# Logic Test

Technical Manual

Alva®

June 2021

# Contents

1 Background 1
1.1 Intelligence . . . 2
1.2 General Mental Ability . . . 4
1.3 Logical ability . . . 5
1.4 GMA and job performance . . . 6

2 Administration 9
2.1 Adaptive testing . . . 9
2.2 Calculation of scores . . . 10
2.3 Interpretation of results . . . 11
2.4 Validation tests . . . 13

3 Psychometric methods 14
3.1 Measurement theory . . . 14
3.1.1 Introduction to the measurement model . . . 15
3.1.2 Measurement model . . . 15
3.1.3 Graphical model . . . 16
3.1.4 Comparing the One- and Three-Parameter Logistic Models . . . 17
3.2 Parameter estimation . . . 18
3.2.1 Standardization . . . 19
3.2.2 Updates and Implementation . . . 20
3.3 Scoring . . . 21
3.4 Item selection . . . 23

4 Test Development 24
4.1 Item design . . . 24

4.2 Standardization . . . . . . . . . . . . . . . . . . . . . . . . . 24
4.3 Calibration . . . . . . . . . . . . . . . . . . . . . . . . . 25
4.4 Continuous development . . . . . . . . . . . . . . . . . . . . . . . . 27
4.4.1 What we do . . . . . . . . . . . . . . . . . . . . . . . . 27
4.4.2 Quality checks . . . . . . . . . . . . . . . . . . . . . . . . 27
4.4.3 Logic Test Update June 11, 2021 . . . . . . . . . . . . . . . . . . . . 28

5 Test quality 29
5.1 Construct validity . . . . . . . . . . . . . . . . . . . . . . . . 29
5.2 Reliability . . . . . . . . . . . . . . . . . . . . . . . . 31
5.2.1 Temporal stability . . . . . . . . . . . . . . . . . . . . . . . . 31
5.2.2 Information . . . . . . . . . . . . . . . . . . . . . . . . 32

6 Adverse Impact 35
6.1 Sample . . . . . . . . . . . . . . . . . . . . . . . . 35
6.2 Level of education . . . . . . . . . . . . . . . . . . . . . . . . 36
6.3 Age . . . . . . . . . . . . . . . . . . . . . . . . 37
6.4 Gender . . . . . . . . . . . . . . . . . . . . . . . . 38

7 Description of data 40
7.1 Raven’s SPM+ norm group . . . . . . . . . . . . . . . . . . . . . . . . 40
7.2 Linear test sample . . . . . . . . . . . . . . . . . . . . . . . . 41
7.3 Standardization sample . . . . . . . . . . . . . . . . . . . . . . . . 42
7.4 Logic Test Update sample (2021-06-11) . . . . . . . . . . . . . . . . . . . . 43

8 References 44

# List of Tables

2.1 Standard Score Interpretation . . . . . . . . . . . . . . . . . . . . . . . . 11

4.1 Percentile score distributions for two norm groups of SPM+ and Alva's logic test . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26

6.1 Score distribution in Alva's database, 2021-06-11 . . . . . . . . . . . . . . 36
6.2 Score distribution by education level . . . . . . . . . . . . . . . . . . . . 37
6.3 Score distribution by age group . . . . . . . . . . . . . . . . . . . . . . . 38
6.4 Score distribution by gender . . . . . . . . . . . . . . . . . . . . . . . . 39

7.1 Raven's SPM+ Norm group (Poland, 2000) . . . . . . . . . . . . . . . . . 40
7.2 Linear test sample . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
7.3 Standardization sample . . . . . . . . . . . . . . . . . . . . . . . . . . 42
7.4 Logic Test Update sample (2021-06-11) . . . . . . . . . . . . . . . . . . . 43

# List of Figures

1.1 Probability Distribution of the IQ Scale . . . . . . . . . . . . . . . . . . . 3
1.2 Catell-Horn theory of intelligence . . . . . . . . . . . . . . . . . . . . . 6
1.3 The mediating role of knowledge on the relationship between GMA and job performance . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.1 Example task . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
2.2 Probability Distribution of the Standard Ten Scale . . . . . . . . . . . . 12

3.1 One-Parameter Logistic Model: Item Response Function (IRF) . . . . . . 16
3.2 One-Parameter Logistic Model: Graphical Model Representation . . . 17

5.1 Test Information Function and Corresponding Reliability Coefficients . 34

# 1 Background

Alva's logic test is a computerized adaptive test aimed for professional assessments in recruitment settings. This technical manual describes the theory and empirical evidence behind it.

Alva's logic test assesses your logical ability, i.e., your ability to process complex information and draw accurate conclusions from it. This is an important part of General Mental Ability (GMA), which has been proven in a vast amount of research to predict job performance in a large variety of roles and industries. GMA has been shown to be the single assessment tool with the strongest predictive power of performance. Logical ability is related to the capacity to solve problems, interpret information, learn new things, and make decisions. The more complex the position, the more impact will mental ability have.

Alva's logic test is a non-verbal, figures-based test; a format that is widely used in both research and practice. This format is very useful, since it minimizes the role of previous experience or domain-specific knowledge. The test requires the test taker to identify patterns and relationships in a material that is rather abstract. Test takers' ability to do so indicates to what extent they are proficient at solving basically any task containing complex or incomplete information.

The fact that the test is non-verbal also makes it less sensitive to differing levels of language proficiency compared to other types of GMA tests.

## 1.1 Intelligence

There are many words to keep track of when discussing logical ability and intelligence. It can be quite confusing to try to weed out the similarities and differences between tests on the market, when the vendors are using different terms to describe them.

Let’s start with the oldest and most widely used term; *intelligence*. The origins of measuring intelligence with standardized tests goes back more than 100 years. Most famous is the endeavour by Alfred Binet and Theodore Simon to develop a standardized measurement for assessing students’ intellectual development. The goal was to identify children who were falling behind developmentally and in need of help. This work resulted in the Binet-Simon intelligence scale, which was later revised to the famous Stanford-Binet scale in 1916 (still in use today).

The term *Intelligence Quotient* (IQ), coined by William Stern, was originally intended to describe the ratio between a student’s mental and chronological age. Students who had a mental ability above their peers would get a high IQ while those who were behind got a low IQ.

Today, IQ is instead defined in relation to a population. An IQ score between 85 and 115 is average, while scores below 85 or above 115 indicates intelligence below and above average respectively. The distribution of intelligence in a broad population is often assumed to be gaussian, or normal, which has a strong foundation both in empirical observations and statistical theory (i.e. the Central Limit Theorem).

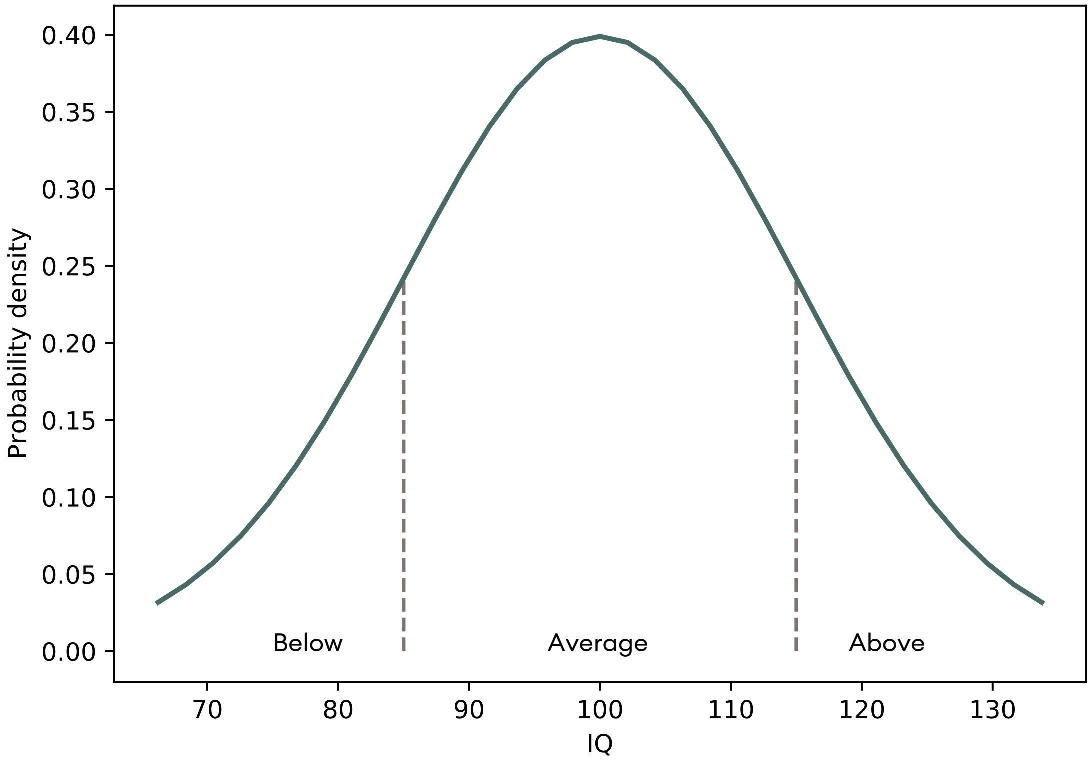

**Figure 1.1:** Probability Distribution of the IQ Scale

It is widely accepted that Intelligence is best measured using multiple types of cognitive tasks. For example, the latest version of the Wechsler Adult Intelligence Scale (WAIS-IV), which is used by psychologists in clinical assessments, consists of 15 separate scales that together form a measurement of IQ. These scales are grouped into verbal comprehension, perceptual reasoning, working memory and processing speed.

There are many theories of General Intelligence today, most of which stems from the work of the psychologist Charles Spearman. In a highly influential paper, he observed that performance on many different cognitive tasks were correlated, and proposed that a general factor (the *g* factor) explained his observations (Spearman, 1904). Later theories include a hierarchy of specific abilities below g, and some even challenge the assumption of one overarching factor. The idea has taken hold however, making the measurement of General Intelligence the goal of most IQ tests today.

Taking all of this into account, the definition of intelligence we adopt at Alva is inspired by David Wechsler, Howard Gardner, Linda Gottfredson and Robert Sternberg to name a few:

> Intelligence is the global ability to process information, think rationally, solve problems, deal with complexity and learn efficiently from experience.

## 1.2 General Mental Ability

While General Intelligence and IQ are commonly used terms in academia, clinical psychology and in day-to-day conversations, it is more common to talk about General Mental Ability (GMA) in organizational and industrial psychology.

Our understanding is that the different terminology is not due to IQ and GMA being different theoretical constructs, nor that the measurements are different in nature. Rather, it's due to the fact that the goal of measurement is different. While IQ tests aim to produce a score that describes individuals' intelligence in relation to the general population, GMA tests are mainly used for ranking candidates in recruitment settings. The reference population is therefore more narrow, focusing only on the working population for example. When a part of the adult population is excluded, in this case non-working adults who can be expected to have a lower level of GMA on average, the ability scale is shifted.

A consequence of this shift is that many individuals get a lower score on a GMA test than on an IQ test. This makes some sense, since it is often of interest to differentiate between individuals that are in the higher ranges of ability in recruitment settings. Therefore, the tasks will be more difficult and the scale won't cover the very lowest ability range. This will of course depend on the test in question, the quality of the tasks, the method for calculating scores, the data collected during test development among other things.

Another consequence is the frequent use of norm groups in tests used for recruitment. Instead of committing to build a single scale that accurately reflects the population of interest, many test publishers resort to providing many different scales and leaving the choice of reference sample (or “norm group”) to the test administrator. In our view, this has caused a lot of confusion regarding the meaning of scores in GMA tests. Our ambition is to provide *one* well designed and properly calibrated scale that can be used across all types of recruitments.

## 1.3 Logical ability

Alva’s logic test measures one central aspect of GMA, which is also referred to as abstract reasoning, figure reasoning, matrix reasoning and fluid intelligence in some contexts. It is inspired by an established test format first introduced by John Raven in 1938, which can be found in modern versions of Raven’s Standard Progressive Matrices (SPM) and Advanced Progressive Matrices (APM) among other contemporary logic ability tests.

In the Cattell-Horn theory of intelligence (Horn & Cattell, 1966), General Intelligence is divided into fluid and crystallized ability. Raven’s matrices and Alva’s logic test are both designed to measure fluid ability (*gf*), which is the ability to reason and solve novel problems without relying on previously acquired knowledge and skills. It can be described as the source of intelligence that an individual uses when he or she doesn’t already know what to do. In contrast, crystallised ability (*gc*) stems from learned experience and is reflected in tests of knowledge, general information, vocabulary and other acquired skills. While *gf* is relatively stable over time, *gc* often increases with age due to the accumulation of knowledge (for a more nuanced discussion, see Hartshorne and Germine, 2015 and Deary, 2012).

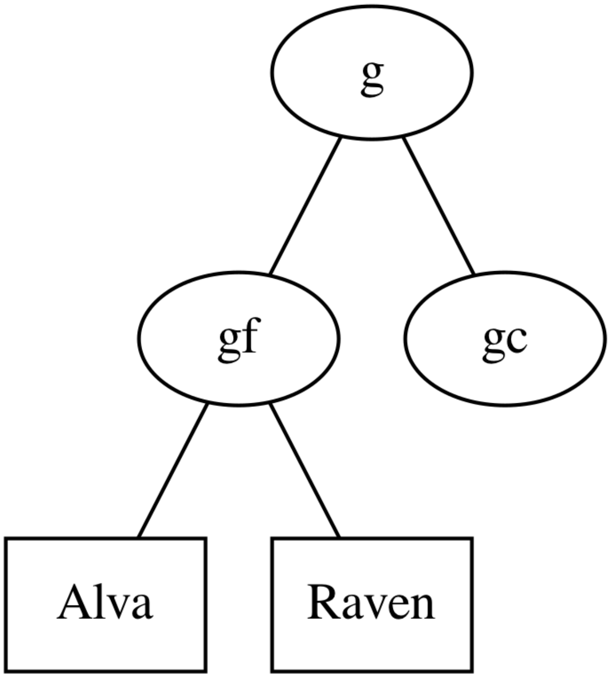

**Figure 1.2:** Catell-Horn theory of intelligence

While logic ability tests do not fully capture GMA or *g*, they have consistently been shown to have a high correlation with - or a high "loading" on - general intelligence. This, together with the practical advantages of being independent of language and other crystallized abilities, makes them ideal for efficiently approximating GMA in many settings.

## 1.4 GMA and job performance

The value of using GMA tests for identifying high potential candidates in recruitment is well known. The utility and predictive validity of GMA tests has been studied since the beginning of the 20th century. Research shows conclusively that people with high GMA are more likely to be top performers than people with low GMA. In the words of Sternberg and Hedlund (2002):

> The so-called general factor (g) successfully predicts performance in virtually all jobs (Schmidt & Hunter, 1998). We do not believe there are any dissenters to this view. [...] The issue is resolved, and it is not clear that further research will do anything more than to replicate what has already been replicated many times over.

It is also clear that the effect of GMA on job performance is stronger as the complexity of the job increases. For unskilled jobs, it has been estimated to *r*=.39 (which is still a strong relationship in the context of psychological science) and *r*=.74 for professional and managerial jobs (Hunter et al., 2006).

These findings are however always discussed on an aggregated level, and the effects are calculated for groups of people. Yet, as managers, recruiters, candidates and employees we are mostly concerned about single individuals. What does it mean for one individual to have a high GMA?

A common theory is that GMA is related to job performance through learning. That is, a person with high GMA is likely to accumulate relevant skills and knowledge, which in turn makes them more likely to perform well.

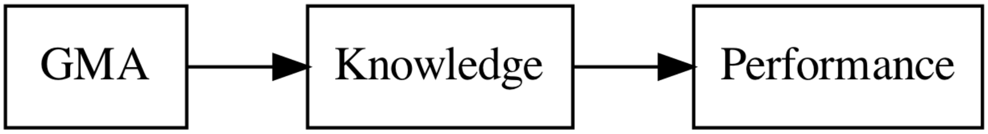

**Figure 1.3:** The mediating role of knowledge on the relationship between GMA and job performance

Looking at the schematic above, it is clear that there are more things at play than intelligence when it comes to job performance. Even the most intelligent person will not acquire any knowledge without putting in time and effort. And without relevant knowledge, no amount of GMA will make a top performer.

High intelligence can be seen as a competitive advantage, much like being tall in a team of basketball players. It certainly helps, but there's a lot more to being a good basketball player than being tall. And there's also a lot more to being a top performer than having a high GMA.

We believe that GMA should be evaluated in relation to the demands of the job and weighted together with other information about the individual. Personality, previous experience, existing knowledge are all relevant data that should be taken into account. What an individual lack in one area can be compensated by strengths in other areas.

# 2 Administration

## 2.1 Adaptive testing

In Alva's adaptive logic test, test takers use their problem solving and abstract reasoning skills to solve logical tasks. They are presented with 20 tasks, to be solved each one in turn.

In each task, the objective is to identify the missing piece in a matrix of figures. The tasks are selected for each individual, based on previous responses, in order to be challenging but not impossible.

There are six options to choose from. There is one, and only one correct answer to each task. The correct answer follows a logical pattern that is applied from top to bottom down the columns, and from left to right across the rows.

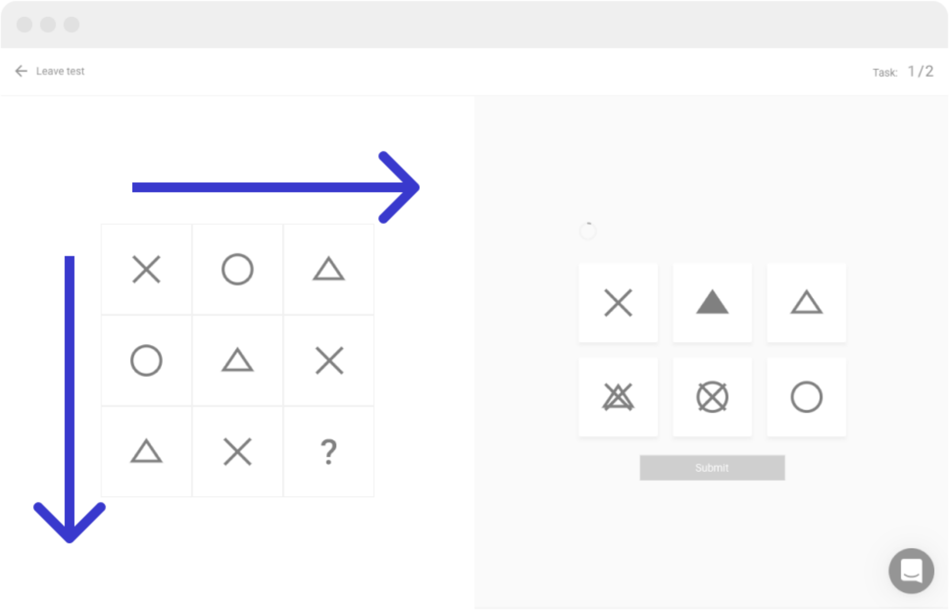

**Figure 2.1:** Example task

There is a time limit of two minutes for each task. Most people complete the test in 15-20 minutes.

Before starting the actual test, test takers get two sample tasks for practice and getting used to the interface.

## 2.2 Calculation of scores

Scores are calculated in real time after each recorded answer as an estimation of their logical ability using an algorithm called EAP (eq. 3.14). Before any answers have been recorded, an average level of logical ability is assumed along with a wide uncertainty range. This assumption is encoded in a *prior distribution*. After recording the first answer, a score is calculated taking both the prior distribution and the likelihood of the observed score (if the answer was correct or not) into account (eq. 3.16). This is repeated throughout the test, and as more answers are recorded, the certainty of the score increases.

The first task is selected randomly from the easiest tasks in the item bank, and the following two are selected to be increasingly challenging. This applies to all test takers, and the purpose is to help them “warm up” to the test before getting more challenging tasks. After the warm-up tasks, subsequent tasks are selected randomly from the 10 most informative tasks in the item bank given test takers’ estimated logical ability (eq. 3.18).

The test is completed when answers to 20 tasks have been recorded. This includes both tasks where the test taker provided an answer and tasks where the time ran out (in which case the answer will be treated as incorrect).

## 2.3 Interpretation of results

The standard score is an estimation of logical ability, relative to the adult working population. The scale is commonly referred to as the Standard Ten (STEN) scale, and it has a mean of 5.5 and a standard deviation of 2.

**Table 2.1:** Standard Score Interpretation

| Standard score | Interpretation | Percentile range |
|---|---|---|
| 1-2 | Below average | 1-6 |
| 3-4 | Slightly below average | 7-30 |
| 5-6 | Average | 31-69 |
| 7-8 | Slightly above average | 70-93 |
| 9-10 | Above average | 94-99 |

The most common standard scores are 5 and 6. The percentile ranges for these scores are wide due to the fact that they cover a large part of the population. Standard scores above 9 or below 2 are much less common, resulting in narrow percentile ranges.

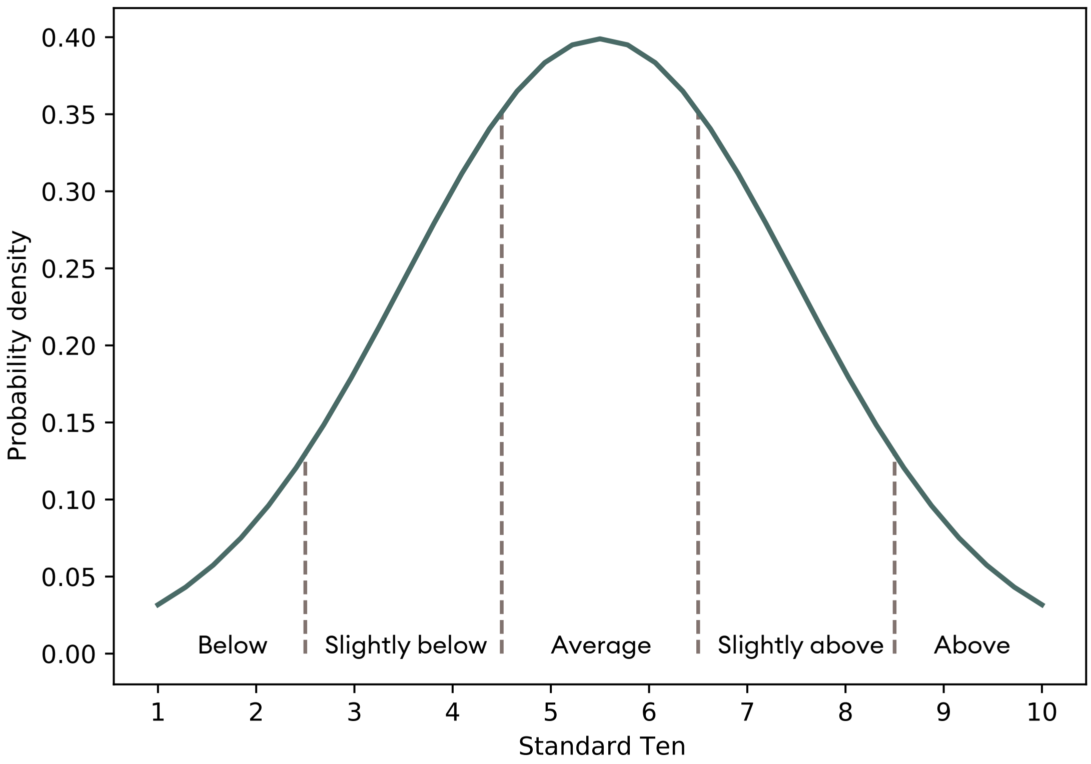

**Figure 2.2:** Probability Distribution of the Standard Ten Scale

The test is adaptive, which means that the tasks are uniquely selected for each individual to match their ability level. A person with a logical ability above average, for example, will get questions that are above average in difficulty. This is why two people with the same number of correct answers can get different standard scores.

## 2.4 Validation tests

Organizations have the possibility to validate candidates' results from the adaptive logic test. This can only be done once per candidate, and Alva strongly recommend administering the validation test on-site under supervision. Best practice is to incorporate the validation test late in the recruitment process, when the candidate visits the office for an interview for example.

The validation test consists of 20 tasks with a time limit of 2 minutes per task, just like the original test. It takes between 15 and 20 minutes to complete. The candidate will get new tasks that they have not been exposed to in the original test.

After the candidate completes the validation test, the administrator gets a side-by-side comparison of the results along with guidance on how to interpret potential differences. Reasons for the differing results include, but are not limited to:

- Getting help from somebody else (cheating)
- Significant differences in mental state due to high-stress levels, illness or other causes.
- Significant differences in testing environment, such as interruptions or distractions.
- Significant differences in the level of effort put in the test

Even with different scores from the original and validation test, it is unlikely that the underlying *construct* of logical ability changes significantly in the short term.

Please note that the candidate will not receive their own results from the validation test. The purpose is for you as an organization to validate candidates' original score and determine their true score.

# 3 Psychometric methods

## 3.1 Measurement theory

Alva's tests are built on second generation psychometrics, also called Item Response Theory (IRT; van der Linden & Hambleton, 1997) or Latent Trait Theory. The measurement theories behind psychometric testing have evolved over the years and while most psychometric tests still rely on first generation psychometrics, or Classical Test Theory (CTT), there are a number of clear advantages in second generation psychometrics:

- IRT supports *adaptive testing*, where the most informative questions are presented based on your previous responses. In this way, candidates are presented with questions most relevant for them, making testing more efficient.
- IRT supports *item banking*, which means that the pool of questions in the test can be continuously developed and increased.
- IRT scoring increases the accuracy of the results, by taking item characteristics into account. This means that fewer questions are needed to get accurate results.
- While CTT deals with a fixed number of questions to form a test scale, IRT deals with each question, or item, separately. This makes the process of continuously improving a test simpler - adding one question doesn't change the entire scale.

### 3.1.1 Introduction to the measurement model

The statistical model used in Alva's logic test is called the One-Parameter Logistic (1PL) model (Birnbaum, 1968). This model describes test tasks using one parameter - difficulty.

The difficulty of a task, which simply refers to how difficult the test is, defines its location on the latent ability scale. An individual who attempts to solve a task that is perfectly matched with his or her logical ability has a 50% chance of finding the correct answer.

In the 1PL model the discrimination parameter, which is defined as the "steepness" of the slope separating individuals with higher logical ability from individuals with lower logical ability, is shared for all tasks.

By taking these characteristics into account, logical ability can be estimated with higher precision than in classical tests.

### 3.1.2 Measurement model

This section describes the mathematics behind Alva's logic test in detail. The 1PL model specifies a statistical relationship between individuals' logical ability, denoted by the greek symbol $\theta$, and the probability of solving a given task correctly. We define $X_{i,j}$ as the outcome for person $i \in \{1, ..., N\}$ for item $j \in \{1, ..., M\}$. The probability of a correct response for person $i$ on a given item $j$ is given by the following Item Response Function (IRF):

$$p(X_{i,j} = 1 | \theta_i, \alpha, \beta_j) = \frac{1}{1 + e^{-\alpha(\theta_i - \beta_j)}} \tag{3.1}$$

The parameters of the model are:

- $\theta_i$: Logical ability of person $i$
- $\alpha$: Discrimination of all items in the test
- $\beta_j$: Difficulty of item $j$

To illustrate, we can plot the IRF for a hypothetical item with known parameters. The form of the function is sigmoidal, with the maximum slope defined by $\alpha$ and the midpoint of the curve defined by $\beta$.

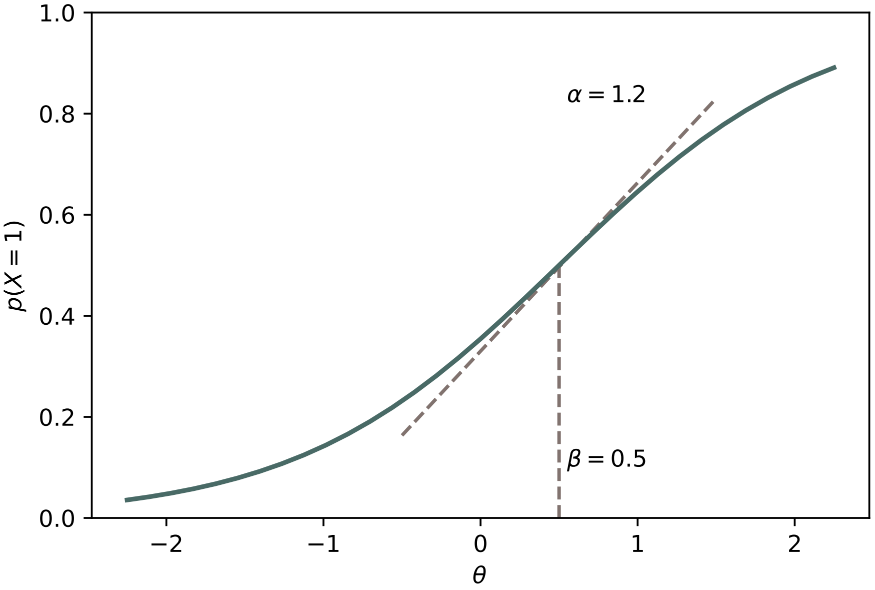

**Figure 3.1:** One-Parameter Logistic Model: Item Response Function (IRF)

### 3.1.3 Graphical model

An alternative way to represent the 1PL model is by the equivalent graphical model. In this notation, a circle represents a latent random variable ($\theta$, $\alpha$, and $\beta$) and a shaded circle represents an observed variable ($X$). Rounded "plates" represent repetition along observations (persons, questions and options) and arrows represent directed relationships between variables.

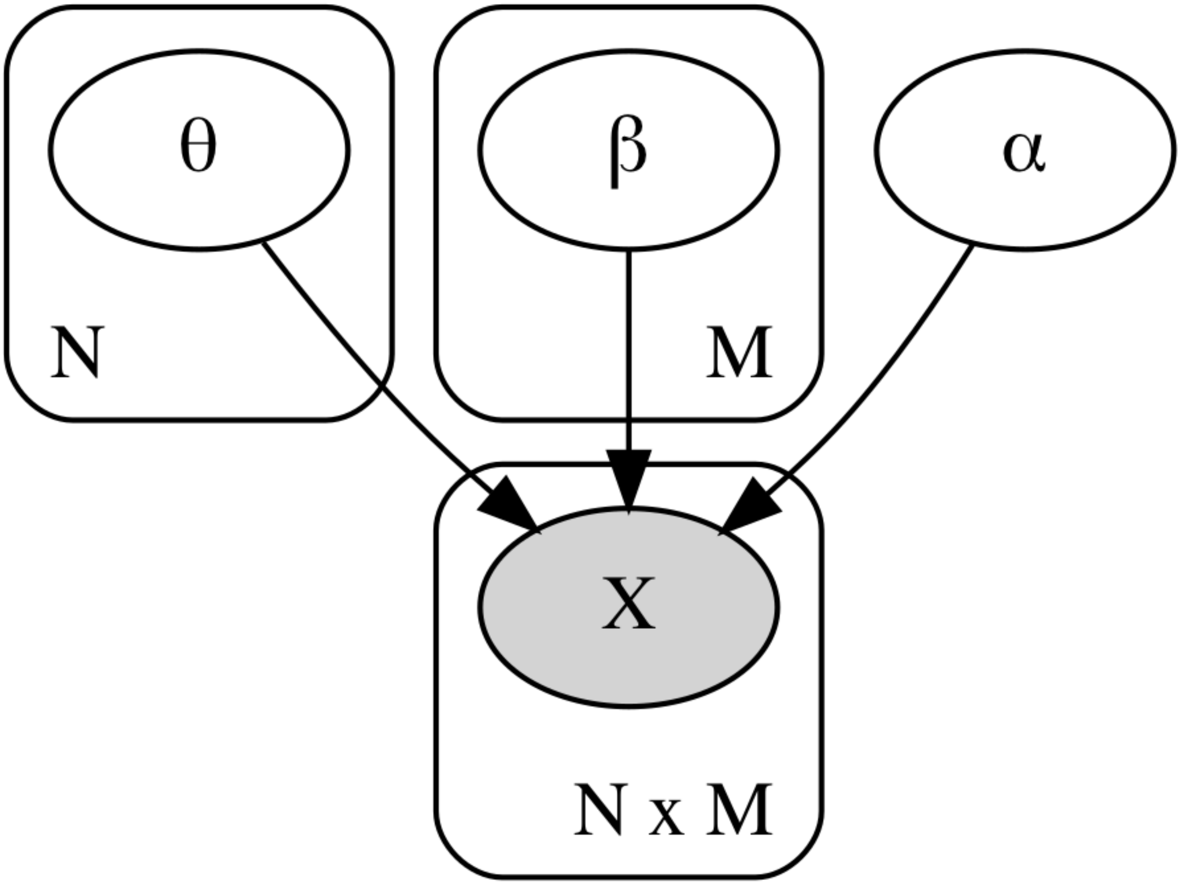

**Figure 3.2:** One-Parameter Logistic Model: Graphical Model Representation

### 3.1.4 Comparing the One- and Three-Parameter Logistic Models

Until our recent update in June 2021, the statistical model used in Alva's logic test was the Three-Parameter (3PL) Logistic Model. This model captures three characteristics in which items can vary - difficulty ($\beta_j$), discrimination ($\alpha_j$) and guessing ($\gamma_j$).

In our latest update of the logic test, we moved away from the 3PL model for two reasons. First, we have discovered that the guessing parameter isn't relevant for our adaptive testing procedure. The reason for including a guessing parameter is to capture the probability of a correct answer by chance alone, which would be when a test taker can't identify the pattern that solves the task. This is more likely to happen when the ability of a test taker is far below the difficulty level of an item. However, due to the adaptive testing procedure, this is extremely rare. Alva's test takers are consistently presented with tasks that match their ability. Second, the discrimination parameter is no longer allowed to vary across items. Using the 3PL model, some great tasks almost never got presented to any test takers. That was due to having a slightly lower discrimination parameter, which made the item selection algorithm skip them. In the 1PL model, however, using a global, or common, discrimination parameter the exposure is more even across all tasks.

## 3.2 Parameter estimation

To be able to use an IRT model for scoring, the model parameters need to be *trained* using observed data. This term is common in statistics and machine learning, and it basically means numerical optimization is used to estimate the parameters in the model. The results from this process are the parameter values that best explain observed data.

A fully observed dataset with $N$ individuals and $M$ items would lead to $N + M$ parameters and $NM$ observations. This model is quite complex, with many parameters relative to the number of observations, which means that the risk of overfit is high. Overfit is an issue when parameter estimates vary a lot depending on the particular dataset used for estimation, making them unreliable for making inferences outside of observed data (generalization). In scenarios like this, it is best practice to apply regularization to reduce the variance in the model and prevent overfit (Hastie, Tibshirani & Friedman, 2009).

At Alva, we use Bayesian Inference for parameter estimation (Fox, 2010). We apply gaussian *priors* (probability distributions assigned to model parameters before observing any data), which is equivalent to L2-regularization (Koller & Friedman, 2009), to control model complexity and generate stable results from the optimization.

Our models are implemented in the Probabilistic Programming Language NumPyro (Phan, Pradhan & Jankowiak 2019), which enables efficient optimization from state-of-the-art sampling methods (Hamiltonian Markov Chain Monte Carlo with the NUTS sampler) and Variational Inference. We were inspired by Luo & Jiao (2017), who have provided code for implementing the 3PL in a Bayesian framwork using similar software (Stan; Carpenter et al., 2017).

New tasks and a constantly growing database of users enables us to continuously analyze and calibrate model parameters. With more data, we learn more about task characteristics and the distribution of logical ability in the population. As a result, the precision of the test improves over time.

### 3.2.1 Standardization

In the previous adaptive version of Alva's logic test we set a standard normal prior probability distribution for both the ability and the difficulty parameters:

$$\theta_0 \sim \mathcal{N}(0, 1) \tag{3.2}$$

$$\beta_0 \sim \mathcal{N}(0, 1) \tag{3.3}$$

In that standardization process, observed results from Raven's Standard Progressive Matrices - Plus version (Raven, Raven & Court, 1998) were used as informative prior, to calibrate the model toward this gold standard measurement of logical ability.

$$\theta_0' \sim \mathcal{N}(\theta_{SPM+}, 0.5^2) \tag{3.4}$$

Then we set an informative normal prior for the discrimination parameter, regularizing the model towards the Rasch model:

$$\alpha_0 \sim \mathcal{N}(1, 0.1^2) \tag{3.5}$$

With the constraint:

$$\alpha > 0 \tag{3.6}$$

The guessing parameter was based on the a priori probability of randomly selecting the correct option, which for the logic test is $1/6$. The beta distribution is apropriate, since it is bounded to the interval $[0, 1]$:

$$\gamma_0 \sim \text{Beta}(1, 5) \tag{3.7}$$

Finally, the responses $X$ were modelled as bernoulli trials with the probability of success from eq. 3.1:

$$X_{i,j} \sim \text{Bernoulli}(p_{ij}) \tag{3.8}$$

### 3.2.2 Updates and Implementation

In the updated version of Alva's logic test the prior probability distributions assigned to the model parameters were updated. For the old items in the test we set a prior probability distribution based on the existing difficulty parameters:

$$\beta_j \sim \mathcal{N}(\mu_{\beta_j}, \sigma^2_{\beta_j}) \tag{3.9}$$

Here $\mu_{\beta_j}$ and $\sigma_{\beta_j}$ refer to the posterior mean and standard deviation from the standardization process described above.

For the new items in the test we set a standard normal prior probability distribution for both the ability and the difficulty parameters:

$$\theta_j \sim \mathcal{N}(0, 1) \tag{3.10}$$

$$\beta_j \sim \mathcal{N}(0, 1) \tag{3.11}$$

Then we set an informative prior for the discrimination parameter, regularizing the model towards the value 1.7:

$$\alpha \sim \mathcal{L}og\mathcal{N}ormal(0.5, 0.2) \tag{3.12}$$

The responses $X$ are modelled as bernoulli trials with the probability of success from eq. 3.1:

$$X_{i,j} \sim \text{Bernoulli}(p_{ij}) \tag{3.13}$$

## 3.3 Scoring

In CTT, scoring consists of two steps - first a raw score is calculated (typically a sum over the responses) and second a transformation from the raw score to a normed score (typically using a norm table). The normed score depend directly on the sample ("norm group") used to generate the norm table. It also requires responses to all questions in the scale, making adaptive testing either impossible or requiring some creative tricks from the test developer.

In IRT, scoring is a statistical estimation of the construct of interest, $\theta$. The resulting score is often on a standard normal scale, often referred to as the *z*-scale (see eq. 3.2). The scoring process only depends on the observed responses and the parameters of the administered questions, not any specific sample. There are no concepts of raw scores or norm tables in IRT.

The question parameters, however, depend on the sample(s) used in the parameter estimation process. One can argue that the parameter estimation process is comparable to the norming process in CTT, but with the advantage of being more flexible in including prior information and aggregating over multiple data sources.

In Alva's tests, the Expected A Posteriori (EAP) algorithm is used for scoring (van der Linden & Pashley, 2010; van der Linden & Glas, 2000). This is an application of Bayes' theorem, where EAP is defined as the expectation of the latent trait over the posterior distribution.

$$\hat{\theta}_{EAP} \equiv \int \theta g(\theta|x_1, \dots, x_j)d\theta \tag{3.14}$$

The Posterior Standard Deviation (PSD) is used to quantify the uncertainty of the score. This is roughly comparable to the Standard Error of Measurement (SEM) in CTT, but with the advantage of being specific to each individual. The SEM, by contrast, is assumed to be constant for all values of $\theta$. PSD is defined as the standard deviation of the posterior distribution.

$$PSD \equiv \int [\theta - \hat{\theta}_{EAP}]^2 g(\theta|x_1, \dots, x_j)d\theta^{1/2} \tag{3.15}$$

First, the unnormalized posterior distribution of the latent trait given the observed responses is calculated as a product of the prior given by eq. 3.2 and the likelihood function given by eq. 3.1.

$$g(\theta|x_1, \dots, x_j) = \theta_0 \prod_{j=1}^{m} p(X_j = x_j \mid \theta, \alpha, \beta_j) \tag{3.16}$$

Second, the posterior expectation and posterior variance are estimated using the quadrature method for appoximating integrals. This is repeated for every administered question, until a satisfactory posterior variance has been reached. This approach is described in detail by de Ayala (2008).

Finally, the scores are transformed from the *z*-scale to the Standard Ten scale using a simple transformation and rounded to the nearest integer.

$$STEN = 2\hat{\theta}_{EAP} + 5.5 \tag{3.17}$$

## 3.4 Item selection

Throughout the test, the score (EAP) and the uncertainty of the score (PSD) are updated after each recorded answer. New tasks are selected adaptively based on how much information they will provide to the next update. In general, the tasks with most information are those where the difficulty level is closely matched to the logical ability of the test taker.

To select the next task throughout the test, the Maximum Posterior Weighted Information criterion (van der Linden & Pashley, 2010) is used.

$$i_k \equiv \underset{j}{\text{argmax}} \int I_j(\theta) g(\theta | x_1, \ldots, x_j) d\theta \tag{3.18}$$

That is, the information function $I(\theta)$ is calculated for all remaining tasks in the item bank, which is then multiplied with the posterior distribution of $\theta$ (eq. 3.16). The posterior weighted information is estimated by integrating out $\theta$. Finally, the tasks with the 10 tasks with the highest posterior weighted information are identified and the next question is chosen randomly from them.

The information for task $I_j$ is given by:

$$I(\theta) = \alpha^2 p(\theta)(1 - p(\theta)) \tag{3.19}$$

# 4 Test Development

Alva's logic test has been developed according to European standards, set by the European Federation of Psychologists' Associations (EFPA). Their test review model is the framework used by most certification agencies in Europe. It contains, among other things, a guide for evaluating the quality of the documentation, norms, reliability, validity and result reports.

## 4.1 Item design

45 tasks were designed to measure logical ability in a fixed, linear testing format. The tasks followed a similar structure as other established tests on the market, such as Raven's Matrices and the Figure Reasoning Test (FRT).

In a small pilot study, results on this test were compared to FRT version A, a test used for admission to Mensa in Sweden. In a sample of 39 participants, the correlation was $r$=0.77, which is considered Excellent by EFPA.

## 4.2 Standardization

In 2019, the adaptive version of Alva's logic test was developed entirely using Item Response Theory (IRT). The data used in the process consisted of responses to the existing 45 tasks from the linear test by 2,295 Alva users and responses to 50 newly designed tasks by an additional standardization sample consisting of 286 participants.

As a first step, we estimated parameters for the 45 tasks in the linear test using data from Alva's platform. The number of observations per task ranged between 1,743 and 2,295, with a mean of 2,133 and a standard deviation of 252. Tasks that were not reached due to participants running out of time were filtered out.

In the second step, tasks were divided into 3 parallel sets, so that there was an overlap of 15 tasks between the sets. This is common practice to be able to link observations across parallel sets. Each set consisted of 39 tasks, both old and new. Using Amazon Mechanical Turk, 286 participants were recruited and assigned to one of the sets randomly.

Parameter estimation for the entire bank of tasks was performed using item parameters from the first step as priors in the Bayesian model. This, together with the fact that some tasks were overlapping, ensures that the parameters for the new tasks are on the same scale as parameters for the old tasks. Normed results from Raven's Standard Progressive Matrices - Plus version (SPM+) for 134 participants was used as priors for the ability parameter, to anchor the scale at appropriate values. This procedure of simultaneous estimation with informative priors means that all available information - from the large sample of Alva users and previous testing with SPM+ - is explicitly taken into account in the final parameter estimates.

## 4.3 Calibration

During standardization, two calibration studies were conducted. First, a sample of 134 participants completed both Alva's logic test and Raven's SPM+. Participants' percentile scores for SPM+ were estimated using two norm tables presented in the SPM+ technical manual (tables D. 11 and D. 13; Raven, Raven & Court, 1998). These norms groups were collected in Germany and in Poland. Percentile scores for Alva's test were estimated by transformation of the standard scores using the gaussian cumulative density function. Results were comparable to normed scores from SPM+.

**Table 4.1:** Percentile score distributions for two norm groups of SPM+ and Alva's logic test

| | SPM+ Germany | SPM+ Poland | Alva |
|---|---|---|---|
| Mean | 51.3 | 65.6 | 60.5 |
| SD | 33.7 | 31.0 | 28.8 |
| 25 %ile | 25 | 50 | 45 |
| 50 %ile | 50 | 75 | 64 |
| 75 %ile | 75 | 95 | 84 |

The second calibration study consisted of 55 members of Mensa Sweden who completed the adaptive version of Alva's logic test. Mensa is an organization for highly intelligent individuals and only those with a measured IQ above 130 are admitted. Since Alva's logic test measures an ability that is closely related to IQ, Mensa members should achieve results above average. Specifically, given some measurement error in the Mensa admission process, the average result for Mensa members should be close to 9.1 with a standard deviation close to 1.2[1]. The observed average score was 8.9 and the observed standard deviation was 1.2 which is comparable to the expected values. This confirms that Alva's logic test is well calibrated for a general adult population.

[1] These figures are based on a simulation, where 100,000 samples were drawn from the IQ score distribution (gaussian with a mean of 100 and a standard deviation of 15) to represent true ability levels and random noise was added (gaussian with a mean of 0 and a standard deviation of 9) to represent measurement error. Results above 130 were labeled "admitted" and the mean and standard deviation of the true ability scores were calculated and transformed to the standard ten scale (gaussian with a mean of 5.5 and a standard deviation of 2).

## 4.4 Continuous development

Most test publishers work according to the waterfall model - they develop content for a new test, collect data, implement and launch. Then they do not touch the content until it's time for the next version. Once the new version is out, it is treated as an entirely new product and the difference to the previous version may be very large. Results from the new version are often not comparable to the old version.

At Alva, we started out with a best-in-class logic test, powered by best practices in machine learning and modern test theory. We follow the agile model in the way we develop and iterate on our tests. Instead of waiting up to five years to launch an entirely new version of our personality test (which is common in the industry), we are continuously collecting data and introducing new questions. This way, we are making sure that our personality test stays ahead of the curve and performs even better than before.

### 4.4.1 What we do

- Introduce new tasks to cover the entire range of logical ability.
- Replace old tasks based on psychometric analysis.
- Enrich the IRT model with the latest data.

### 4.4.2 Quality checks

In each iteration, we use the following metrics to ensure increasing test quality while at the same time avoiding radical changes in results for our users.

- The reliability of the test increases or remains at an excellent level according to European standards
- The construct validity increases or remains at an excellent level according to European standards
- The correlation between old and new scores is higher than 0.9

- The average score isn’t changed by more than 0.5 points on the standard scale
- More than 50% of users keep an identical score
- Less than 5% of users change by more than +/- 1 point

### 4.4.3 Logic Test Update June 11, 2021

#### 4.4.3.1 What we did

- Introduced 16 new tasks to cover the entire range of logical ability.
- Removed 2 old tasks based on psychometric analysis.
- Enriched the IRT model with data from 77,327 tests and 1.54 million responses completed in 2019-2021.

#### 4.4.3.2 Quality checks

In this iteration, we used the following metrics to ensure increasing test quality while at the same time avoiding radical changes in results for our users.

- Increased the average reliability from 0.86 to 0.94 for the logic test
- The reliability and construct validity are at an excellent level according to European standards
- The correlation between the old and new scores is higher than 0.9 (*r*=0.995)
- The average score is changed by less than 0.5 points on the standard scale (average change = -0.1 points)
- 75.8% of all Alva’s users will keep the identical score
- 99.9% of Alva’s users will keep a result within +/- 1 point

# 5 Test quality

The usefulness of any measurement depends on the quality of the measurement process. For psychological tests, this comes down to the *validity* of the test results. The most widely accepted definition of validity is:

> the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests (AERA et al. 2014)

In the case of Alva's logic test, the proposed use is selection in recruitment settings. In earlier chapters, we have presented existing theory and research supporting the use of GMA tests to predict job performance. This chapter presents evidence from our own studies.

## 5.1 Construct validity

When using a psychological test, it is of great interest for both the administrator and the respondent that the test is actually measuring what it claims to measure. This is what construct validity is all about.

A *construct* is a theoretical and statistical concept. It is the "hidden truth" about individuals that we assume causes the responses to the questions in a test. With this assumption, we can use the test results to calculate estimates of the construct.

In the case of Alva's logic test, the construct of interest is the logical ability of individuals. The tasks in the tests make up the measurement.

There are several ways to collect evidence for the construct validity of a psychological test. The most common method is to collect data from two tests that are designed to measure the same or similar constructs and calculate the correlation between the scores of the two tests. This is often referred to as the convergent validity of a test.

The hypothesis is that if one and the same construct causes the results of two separate tests, then the scores of the tests should correlate highly. According to European standards from EFPA, convergent validity coefficients above 0.75 are deemed *Excellent*.

The construct validity of Alva's logic test was estimated using a sample of 134 participants who completed both Alva's test and Raven's Standard Progressive Matrices Plus (SPM+). **The observed validity coefficient was *r*=0.83.**

### Sample

The sample consisted of 58% males and 42% females. The age ranged from 20 to 63 with an average of 34.9 and a standard deviation of 9.3. The education level of the participants was high. 63% had completed a Bachelor's degree or equivalent, 7% a Master's degree or equivalent, 29% secondary education or high school and 1% other educational background.

### Instruments

Alva's logic test is a computerized adaptive test aimed for professional assessments in recruitment settings. Raven's SPM+ is an extended version of the original matrix test, used since 1938 to measure abstract reasoning.

### Study

Participants were recruited using Amazon Mechanical Turk, an online crowdsourcing platform. They were asked to complete 39 tasks from Alva's item bank and 36 tasks from Raven's SPM+ in two separate sessions with one week apart. All participants received compensation.

The standard score for Alva's test was calculated using the Expected A Posteriori (EAP) method, with the same model, item parameters and prior distribution as in the live test. The total score was used to represent results on Raven's SPM+. Tasks from sets A and B from SPM+ were left out, due to being overly simplistic, and the total score was estimated using table 35 from the SPM+ manual (Raven, Raven & Court, 1998).

## 5.2 Reliability

Reliability refers to the overall consistency of a measurement. A test with high reliability produces similar results under similar conditions.

While a high degree of validity implies a high reliability of a measurement, the inverse is not true. A test can have a high reliability but measure something completely irrelevant.

In the framework of Classical Test Theory, reliability is defined as the ratio of *true score* variance to the total variance of test scores. Since the true score is impossible to directly observe, reliability is instead estimated using methods such as test-retest reliability, internal consistency and parallel-test reliability. From the reliability coefficient, a Standard Error of Measurement (SEM) can be estimated, which provides an uncertainty range to test scores.

In the framework of Item Response Theory, reliability has a less prominent role. On the one hand, reliability can reasonably be estimated from the Test Information Function (see section below) which is an extension of the concept of reliability. On the other hand, the uncertainty of a measurement is more commonly estimated for each test session using methods like PSD (eq. 3.15).

In this section we will nontheless focus on the more familiar concept of reliability, to be consistent with technical manuals from classical tests.

### 5.2.1 Temporal stability

A desirable property of a psychological measurement is the stability of scores over time and across different situations. A common way to investigate this quality is to administer the test of interest to a group of people twice and calculate the correlation between the scores. This correlation coefficient is also referred to as the *test-retest reliability* of the measurement.

The time between the sessions should be short enough for the underlying trait to remain unchanged, but long enough for the participants to not remember their exact responses to the questions. It is common practice to administer the test with an interval of two weeks, which was also used in this study.

According to European standards set by EFPA, a test-retest reliability coefficient above 0.8 is deemed *Good* and above 0.9 *Excellent*.

The test-retest reliability of Alva's logic test was estimated using a sample of 117 participants who completed the test twice with 14-15 days in between. **The observed correlation was *r*=0.81.**

**Sample**

The sample consisted of 58% males and 42% females. The age ranged from 20 to 63 with an average of 35.1 and a standard deviation of 9.4. The education level of the participants was high. 63% had completed a Bachelor's degree or equivalent, 6% a Master's degree or equivalent, 29% secondary education or high school and 2% other educational background.

**Study**

Participants were recruited using Amazon Mechanical Turk. They were asked to complete a set of 39 tasks from Alva's item bank twice, with two weeks in between.

### 5.2.2 Information

It is well known, even to classical test theorists, that measurement precision is not uniform across the scale of measurement. Tests tend to distinguish better for test-takers with moderate trait levels and worse in the higher and lower score ranges. In IRT, the concept of reliability is extended from a single number to a function called the *Test Information Function*.

In statistics, information (or more specifically, Fisher Information) refers to how strong the relationship is between data and parameters of a statistical model. High levels of information means that parameters can be estimated efficiently. In Alva's logic test, the Test Information Function tells us how well logical ability of individuals can be estimated using the tasks in the item bank (eq. 3.19).

There is a mathematical relationship between the reliability $r$ and information $I$, which allows us to translate from one to the other in a simple way.

$$r = 1 - \frac{1}{I} \tag{5.1}$$

According to EFPA Standards, an average Information across the scale of measurement above 10 is deemed *Excellent* and above 5 is deemed *Good*. This is equivalent to a reliability coefficient above 0.9 and 0.8 respectively.

The average information across all score ranges for Alva's logic test is 16.11, with the full range of possible ability scores meeting the EFPA standards for *Excellent* average Information (*range*: 12.04 - 18.39). This is transformed to an ***Excellent* average reliability of *r*=0.94**, with a maximum of 0.95 and a minimum of 0.92, using eq. 5.1.

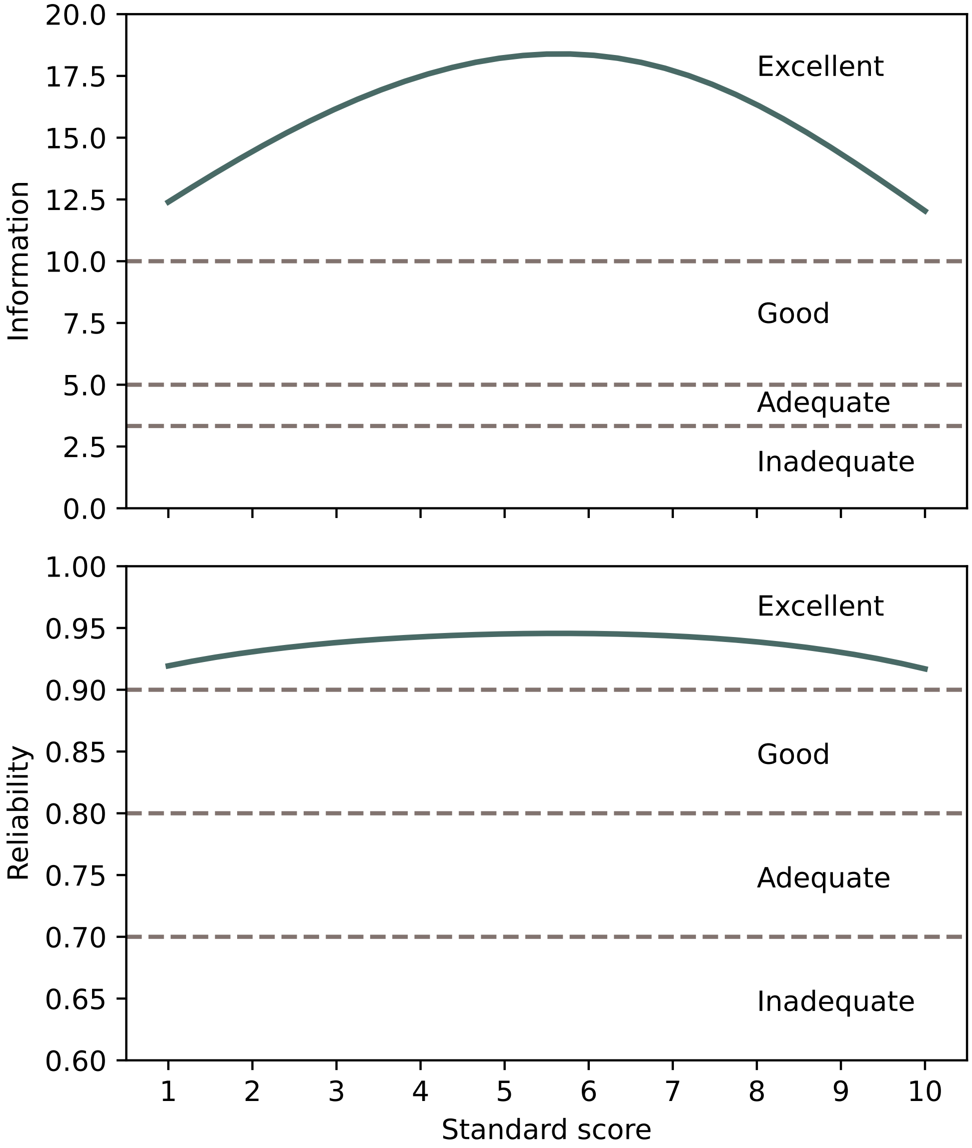

**Figure 5.1:** Test Information Function and Corresponding Reliability Coefficients

# 6 Adverse Impact

To ensure fairness and objectivity in our assessment methods, we collect anonymous demographics data from candidates (voluntarily) after they have completed tests as part of the assessment process. Primarily, we are focusing on three taxonomies: Level of education, Age, and Gender.

## 6.1 Sample

Alva's logic test is developed to represent the "working population" with test results normalized on a STEN-scale, which assumes a normal distribution $X \sim \mathcal{N}(5.5, 2)$. Many of Alva's customers serve in industries with job positions that on a general level is considered "complex", and samples from Alva's database are thus expected to have a characteristic of slightly above 5.5 and 2.

Furthermore, it's important to note that a sample from Alva's database is not a representative sample for the "working population" since it is highly dependent on the current customer portfolio. For example, if Alva's customers primarily were hiring for PhD-positions, the sample would consist of candidates with relatively high age and presumably high logical ability, which in turn would not be representative for the general "working population". The job positions offered by Alva's customers, also in terms of social norms (for example "bus driver" is overrepresented by men) will lay the ground for the sample. What this means is that we look at the statistics and try to explain the variance by looking at various factors.

This analysis is performed on 2021-06-11 on a sample of 77,327 people that have taken Alva's adaptive logic test between 2019-12-09 and 2021-06-11. The characteristics of the sample are shown in the table below.

**Table 6.1:** Score distribution in Alva's database, 2021-06-11

| # Persons | Mean | Std.dev. | Skewness |
|---|---|---|---|
| 77,327 | 6.29 | 2.07 | -0.36 |

The sample mean is 6.29, which is significantly higher than the expected mean for a working population of 5.5. This indicates that the sample on average has higher logical ability than a working population, which most likely coincides with the fact that many of Alva's customers offer more complex and competitive jobs. The sample standard deviation of 2.07 is about what's expected. The skewness of -0.36 indicates that the sample is moderately skewed towards the upper bound on the scale, which also suits well with the underlying assumptions for the sample.

## 6.2 Level of education

Following from our definition of GMA, we expect to see a relationship between highest attained education level and results on the logic test. Due to the connection between GMA and effective learning, individuals with higher GMA should be more likely to both initiate and successfully finish higher education.

By splitting the sample between different levels of education, one observes sample means between 5.48 and 6.78. The different sub-samples all have a sample standard deviation around or below 2 and skewness between -0.42 and -0.13. The sub-sample characteristics are presented in the table below.

**Table 6.2:** Score distribution by education level

| Level of education | # Persons | Mean | Std.dev. | Skewness |
|---|---|---|---|---|
| Secondary Education & High School | 17,952 | 5.48 | 2.16 | -0.13 |
| Post-secondary education, < 3 years | 9,668 | 5.76 | 2.02 | -0.17 |
| Post-secondary education, > 3 years | 38,915 | 6.78 | 1.88 | -0.42 |
| Other/No Response/Rather not say | 10,792 | 6.31 | 2.1 | -0.4 |
| **Total** | 77,327 | | | |

The group with the highest sample mean is the group with "Post-secondary education, 3 or more years", which is expected. Generally, groups with higher education receive higher test results.

## 6.3 Age

We expect a slightly negative correlation between age and logical ability, since this is closely related to the concept Fluid Intelligence (see for example Hartshorne and Germine, 2015 and Deary, 2012).

By splitting the sample into different age categories, we observe that the group between 20-29 are receiving slightly higher test results compared to the total sample. Characteristics of the different sub-samples are presented in the table below.

Overall, one observes a slightly negative correlation between age and results on the logic test, which is according to the expectations. However it's worth noticing that there are big differences in the sizes of the sub-samples. Furthermore, Alva's has a significant amount of "internship programs" which typically attracts high performing young adults. On the other hand, there are also customers with more "blue-collar"-jobs, which is not as prestigious and potentially could attract an older population.

Before drawing any conclusions from this result, one should investigate if some of the variance might be explained by differences in jobs and their different candidate pools.

**Table 6.3:** Score distribution by age group

| Age | # Persons | Mean | Std.dev. | Skewness |
|---|---|---|---|---|
| 16-19 | 2,516 | 5.64 | 2.02 | -0.15 |
| 20-34 | 42,823 | 6.53 | 2.05 | -0.43 |
| 35-44 | 12,814 | 6.2 | 1.99 | -0.35 |
| 45-54 | 6,617 | 5.63 | 1.98 | -0.25 |
| 55-64 | 2,448 | 4.87 | 1.99 | -0.08 |
| Other/No Response/Rather not say | 9,545 | 6.32 | 2.09 | -0.39 |
| **Total** | 77,327 | | | |

## 6.4 Gender

We do not expect any significant relationship between gender and logical ability. We believe that observed differences found in some studies on related tests of fluid intelligence are due to methodological flaws (see for example Blinkhorn, 2005) and/or sample limitations (see for example Dykiert, Gale & Deary, 2008).

In the set of demographic data, the big majority (94%) select “male” (52%) or “female” (42%), which makes it the only two sub-samples relevant to investigate. The characteristics of the sub-samples are presented in the table below.

**Table 6.4:** Score distribution by gender

| Gender | # Persons | Mean | Std.dev. | Skewness |
|---|---|---|---|---|
| Female | 32,339 | 6.02 | 1.97 | -0.26 |
| Male | 40,107 | 6.48 | 2.12 | -0.46 |
| Other/No Response/Rather not say | 4,881 | 6.42 | 2.12 | -0.45 |
| **Total** | 77,327 | | | |

# 7 Description of data

In the development of Alva's logic test, the following samples were used:

- SPM+ norm group: used to calibrate the model to a general adult population.
- Linear test sample: used to train the initial parameters of the model.
- Standardization sample: used to extend the item bank with an additional 50 items.
- Logic Test Update sample (2021-06-11): used to extend the item bank with 16, identify 2 uninformative items, and further improve the parameters of the model

## 7.1 Raven's SPM+ norm group

**Table 7.1:** Raven's SPM+ Norm group (Poland, 2000)

| | n | % |
|---|---|---|
| **Age** | | |
| 19-34 | 83 | 40% |
| 35-44 | 35 | 17% |
| 45-54 | 32 | 16% |
| 55-64 | 56 | 27% |
| **Total** | 206 | |

Other variables are assumed to be population-representative based on the following footnote from the manual for Raven's Standard Progressive Matrices - Plus version, table D.13 (Raven, Raven & Court, 1998):

> [...] The adults were tested individually. A quota sampling procedure was employed. Testers were asked to seek out respondents of appropriate age, sex, place of residence (e.g. large city, small town, village), and education to yeild the correct proportions of people in each of the categories to correspond to the country as a whole. Based on data collected by Aaron, Jackson and Seerden (Greater Fort Bend Economic Development Council, 1998).

## 7.2 Linear test sample

**Table 7.2:** Linear test sample

| | n | % |
|---:|---|---|
| **Gender** | | |
| Male | 1,485 | 65% |
| Female | 780 | 34% |
| Other/Rather not say | 30 | 1% |
| **Age** | | |
| 20-34 | 1,808 | 79% |
| 35-44 | 303 | 13% |
| 45-54 | 119 | 5% |
| 55-64 | 18 | 1% |
| Rather not say | 47 | 2% |
| **Total** | 2,295 | |

## 7.3 Standardization sample

**Table 7.3:** Standardization sample

| | n | % |
|---:|---|---|
| **Gender** | | |
| Male | 190 | 66% |
| Female | 95 | 33% |
| Other/Rather not say | 1 | 0% |
| **Age** | | |
| 20-34 | 182 | 64% |
| 35-44 | 65 | 23% |
| 45-54 | 24 | 8% |
| 55-64 | 13 | 5% |
| Rather not say | 2 | 1% |
| **Education level** | | |
| Upper & Lower secondary | 80 | 28% |
| Post-secondary, less than 3 years[1] | - | - |
| Post-secondary, 3 years or more | 208 | 73% |
| Other/No Response/Rather not say | 0 | 0% |
| **Total** | 286 | |

[1] Not among the options in data collection

## 7.4 Logic Test Update sample (2021-06-11)

**Table 7.4:** Logic Test Update sample (2021-06-11)

| | n | % |
|---|---|---|
| **Gender** | | |
| Female | 32,339 | 42% |
| Male | 40,107 | 52% |
| Other/No Response/Rather not say | 4,881 | 6% |
| **Age** | | |
| 16-19 | 2,516 | 3% |
| 20-34 | 42,823 | 56% |
| 35-44 | 12,814 | 17% |
| 45-54 | 6,617 | 9% |
| 55-64 | 2,448 | 3% |
| Other/No Response/Rather not say | 9,545 | 12% |
| **Education level** | | |
| Upper & Lower secondary | 17,952 | 23% |
| Post-secondary, less than 3 years | 9,668 | 13% |
| Post-secondary, 3 years or more | 38,915 | 50% |
| Other/No Response/Rather not say | 10,792 | 14% |
| **Total** | 77,327 | |

# 8 References

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

Blinkhorn, S. (2005). Intelligence: a gender bender, *Nature, 438*, 31-32. DOI: 10.1038/438031a

Birnbaum A (1968). *Some Latent Trait Models and Their Use in Inferring an Examinee's Ability.* In F Lord, M Novick (eds.), *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, MA.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76:1.*

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology, 63*, 453-482.

de Ayala, R. J. (2008). *The Theory and Practice of Item Response Theory.* Guildford Press, New York, US.

Dykiert, D., Gale, C. R., & Deary, I. J. (2008). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence, 37*, 42-47.

*EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests, version 4.2.6.* (2013). Available online (click here)

Fox, J.-P. (2010). *Bayesian Item Response Modeling.* New York: Springer Hartshorne, J. K, & Germine, L. T. (2015). When Does Cognitive Functioning Peak? The Asynchronous

Rise and Fall of Different Cognitive Abilities Across the Life Span, *Psychological Science, 26:4*, 433-443.

Hartshorne, J. K, & Germine, L. T. (2015). When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span, *Psychological Science, 26:4*, 433-443.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition).* Springer.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57,* 253-270.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594-612.

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge.

Luo, Y. & Jiao, H. (2018). Using the Stan Program for Bayesian Item Response Theory. *Educational and Psychological Measurement, 78:3,* 384-408. DOI:10.1177/0013164417693666

Phan, D., Pradhan, N., & Jankowiak, M. (2019). *Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro*. arXiv:1912.11554

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales (1998 edition).* Oxford: Oxford Psychologists Press.

Spearman, C. (1904). "General Intelligence", Objectively Determined and Measured. *The American Journal of Psychology, 15:2,* 201-292.

Sternberg, R. J. & Hedlund, J. (2002). Practical Intelligence, g, and Work Psychology. *Human Performance, 15,* 143-160.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

van der Linden, W. J., & Glas C. A. W. (Eds.) (2000). *Computerized Adaptive Testing: Theory and Practice.* Netherlands: Springer.

van der Linden, W. J. & Pashley, P. J. (2010) Item Selection and Ability Estimation in Adaptive Testing. In: van der Linden, W. J., & las, C. A. W (Eds), *Elements of Adaptive Testing*. New York: Springer. doi: 10.1007/978-0-387-85461-8_1