Definition of Psychometrics:

- It consists of *mathematical* and *statistical models* used to define measures in PSYCHOLOGY
- Psychometrics is a scientific discipline concerned with the question of how psychological constructs (e.g., intelligence, neuroticism, or depression) can be optimally related to observables (e.g., outcomes of psychological tests, genetic profiles, neuroscientific information).
- It consists in the construction of measurement models, in which the construct of interest is represented as a latent variable that acts as a common determinant of a set of observable variables

Aims of Psychometrics:

- developing reliable and valid **measures** (test, interviews, laboratory tests, etc.).
- developing new models, statistical methods, and computational approaches to define measurement proprieties of psychological measures.
- assessing how a psychological measure works across different representative samples of people.

Measure: a standard unit used to express the size, amount, or degree of something.
The measurement is the process of assigning numbers or symbol to the characteristics of the object in exam. Specifically, numbers represent quantities of attributes or the degree to which a quality is present.

For example, consider a scale from 1 to 10 for locating consumer characteristics (preference for the product). Each respondent is assigned a number from 1 to 10 denoting the degree of unfavorableness for the product, with '1' indicating extremely unfavorable and '10' indicating extremely favorable. Here, the measurement is the process of assigning the actual number from 1 to 10 to each respondent.

To sum up, measurement consists of rules for assigning symbols to objects in order to

1. represent quantities of attributes numerically (**scaling**) or
2. define whether the objects fall into the same category or if they fall into different categories with respect to a given attribute (**classification**).

There are different kind of instruments of measurement in Psychology, one of the main one is the **Test.**

Psychological Test:

- can be defined as any *standardized* situation where the individual's behaviour is sampled, observed, and described. This results in an objective and standardized measure of behaviour.
- Tests should satisfy these requirements:
    - *Standardization* of stimuli
    - Objectivity during responses registration (procedure *standardization*)
    - Objectivity during responses evaluation (procedure *standardization*)
    - Objectivity during responses interpretation (lack of arbitrariness)

Standardize: cause (something) to conform to a standard > to bring into conformity with a standard especially in order to assure consistency and regularity

So *standardization* is a key concept and a necessity:

- Consistency and objectivity of how tests are administered and scored. In order to compare one person to another on a test, it is important that they take the test under the same conditions and the same scoring procedure is applied to both.
- **Standardized Situation**: situation where everything is stable.
- **Standardized Measure**: single empirical data linked to a general reference system.
    - A measure is standardized when:
        - → Its rules are clear;
        - → Could be easily applied;
        - → Administrators are required only the skills they acquired in their initial training;
        - → Results do not depend upon the specific administrator.

Advantages of standardized measures:

1. Objectivity
2. Quantification
3. Communication
4. Economy
5. Scientific generalizability

A **measurement scale**, in statistical analysis, is the type of information provided by numbers. Each of the four scales (i.e., nominal, ordinal, interval, and ratio) provides a different type of information. Measurement refers to the assignment of numbers in a meaningful way and understanding measurement scales is important to interpreting the numbers assigned to people, objects, and events.
As I said, there are four type of scale:

TABLE 1-1 STEVENS' LEVELS OF MEASUREMENT, BASIC DEFINING OPERATIONS, PERMISSIBLE TRANSFORMATIONS, EXAMPLES OF PERMISSIBLE STATISTICS, AND EXAMPLES

Scale	Basic operation	Permissible transformations	Permissible statistics	Examples
Nominal	= vs ≠ (equality vs. inequality)	Any one-to-one	Numbers of cases, mode	Telephone numbers
Ordinal	> vs. < (greater than vs. less than)	Monotonically increasing	Median, percentiles, order statistics	Hardness of minerals, class rank
Interval	Equality of intervals or differences	General linear $x' = bx + a$	Arithmetic mean, variance, Pearson correlation	Temperature (Celsius), conventional test scores (?)
Ratio	Equality of ratios	Multiplicative (similarity) $x' = bx$	Geometric mean	Temperature (Kelvin)

Nominal	numbers are used to name or identify people, objects, or events (main example: gender – male or female)
Ordinal:	numbers represent rank order and indicate the order of quality or quantity, but they do not provide an amount of quantity or degree of quality. Usually, the number 1 means that the person (or object or event) is better than the person labeled 2; person 2 is better than person 3, and so forth. Such ordinal scaling does not,

	however, indicate how much more potential the leader has over the person assigned a rating of 2.
Interval:	numbers form a continuum and provide information about the amount of difference, but the scale lacks a true zero. The differences between adjacent numbers are equal or known. If zero is used, it simply serves as a reference point on the scale but does not indicate the complete absence of the characteristic being measured. The Fahrenheit and Celsius temperature scales are examples of interval measurement. In those scales, 0 °F and 0 °C do not indicate an absence of temperature.
Ratio:	Ratio scales have all of the characteristics of interval scales as well as a true zero, which refers to complete absence of the characteristic being measured. Physical characteristics of persons and objects can be measured with ratio scales, and, thus, height and weight are examples of ratio measurement.

IMPORTANT: scales are **conventions** (but this does not mean that they are arbitrary!)
For example:
The Fahrenheit scale for measuring temperature was THE scale for temperature until the absolute zero was discovered. After that, the Kelvin scale was introduced as a new (and more useful) scaling.
Intelligence was initially defined as the ratio of mental age to chronological age (the so-called intelligence quotient; IQ). Now, intelligence is assessed comparing the subject's performance with the performance obtained by same-age

→ These examples show that no "real" scales were discovered. It is better to consider scales as **conventions** that can change. Indeed, more useful conventions may be developed. A convention establishes the scale properties of a measure according to a specifically system of theories. The usefulness (*validity*) of a measure should be assessed considering how it relates to other variables included in the same domain. The "best" scaling procedure results in the simplest net of relationship of the variable of interest with other variables

Models for scaling

Although it is possible to create multidimensional scales, it is better to use **UNIDIMENSIONAL MODELS**: unidimensional models assume that items (i.e., any stimuli used in measurement) measure a single attribute (unidimensional scaling). Multidimensional models often result in ambiguous meaning; thus, they are difficult to measure accurately.

Single Item Measures vs. Multi-item Measures

There are important reasons for relying on **more than one item** to measure a psychological attribute:

1. Individual items usually correlate poorly with the particular attribute in question;
2. Each item tends to relate to attributes other than the one to be measured;
3. Each item has a degree of specificity in the sense of not correlating with any general attribute or factor;
4. Individual items have considerable random measurement error (i.e., they are **unreliable**).

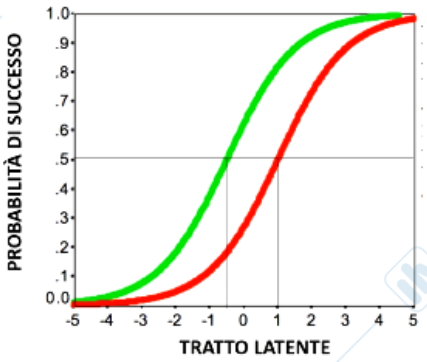
There are two main measurement models: **classical test theory** and **Item response theory**.

Item Response Theory

It is a psychometric theory of measurement based on the concept that the probability that an item will be answered correctly is a function of an underlying trait or ability that is not directly observable (latent trait).
So, it considers the chance of getting particular items right or wrong based on the item characteristic and individual ability.

Item performance is related to the estimated amount of respondent's latent trait. The total score on a test is taken as an estimate of that ability. A person's specified ability (θ) succeeds on an item of specified difficulty.
Item Response Theory defines the intensity of the attribute in terms of true scores.

Item Trace Lines (Item Characteristics Curves): is **the basic building block of item response theory**
Nearly all models for scaling people can be described by different types of curves.



This one is a curve (Rasch Model) that relates the attribute the test measures (in the picture "tratto latente" = latent trait) to the probability of responding in the keyed direction (the so-called response *α* [alpha] – e.g., the presence of a symptom) (in the picture"probabilità di successo" = success probability).
Item characteristic curve shows the expected response probability for people at the same level of the attribute.

Two basic properties of a trace line are its difficulty and its discrimination

- **DIFFICULTY**: how much of the attribute an individual must possess to achieve a given probability of response alpha.
- **DISCRIMINATION**: describes the extent to which the probability of response alpha correlates with the attribute.

Classical test theory

Premises:

1. A test is a sample of a person's behavior, with an underlying idea of the mental functioning.
   A test is not an exact picture of the reality, it's a measure:
   → Based on a theory, specific for each function
   → Depending on an operational model (translation of concepts into operations aimed at evaluating them) and on a measuring model.
2. Each scale assesses one and only one characteristic;

3. It distinguishes among:
   → A single observation (**test item**);
   → Error factor: anything that doesn't make the answer to an item or the final score completely true, valid, representative to the effective reality. we assume that errors are independent among them and from the true score.

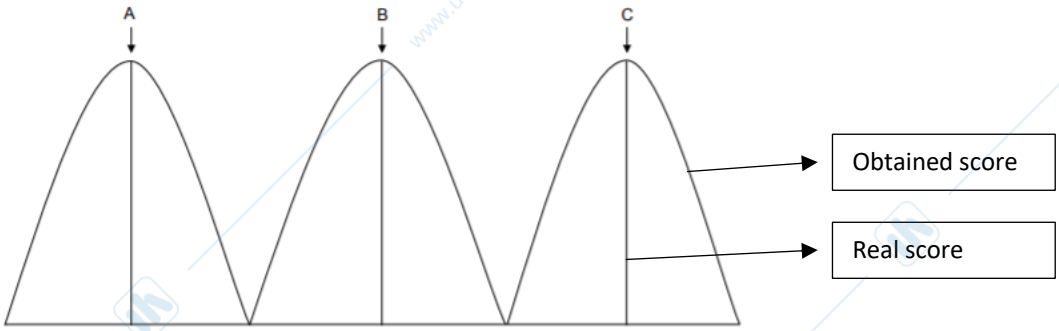
→ A hypothetical, perfectly measured, attribute (i.e., **true score**) hypothetically measured by an infinite number of items (or trials) > it's the score without the error

→ A more general attribute measured by a finite number of items that may be influenced spuriously (i.e., an obtained, observed or **fallible score**)

X	=	T	+	E
observed score		true score		error

Test scores obtained from each person contain an amount of error and an amount of true score

Definition: It is a theory of testing based on the idea that a person's observed or obtained score on a test is the sum of a true score (error-free score) and an error score.
It usually defines the intensity of the attribute in terms of obtained measures (i.e., fallible scores).



The true scores (vertical solid lines) for persons A, B and C fall at progressively higher points along the continuum. Since the obtained score contains random error, it differs from the true score (curved lines). The more the obtained scores in the figure vary about the true score, the more measurement error there is.

If one could give infinite (in realistic terms, many) alternative test forms, the average of the test scores would closely approximate the true score (in case of infinite alternative forms, the average score would correspond exactly to the true score).

The "standard deviation" of each person's error distribution indexes the amount of error. This standard deviation of errors is called the **STANDARD ERROR OF MEASUREMENT** (σmeas).
The standard error of measurement (SEM) estimates how much of the observed score in a test is a precise measure of the attribute: the greater the SEM is, the **SMALLER** the precision of the test is and vice versa.
Reducing the SEM is important both because it increases the precision of the test, both because the SEM is strictly and inversely connected to the reliability of the measure.
During the assessment it is possible to calculate how much we can rely on the test scores and how much these scores are real estimates of what we are evaluating.

Example: imagine if we bought a weight scale: if the SEM of this scale is 5 kg, when the weight scale shows 55 kg, our true weight could be 50 kg or 60 kg. If it was possible to reduce SEM to 2.5 kg, the precision would increase (when the weight scale shows 55 kg, our true weight could be between 52.5 kg and 57.5 kg).

Scores obtained from alternate forms:

- Are distributed symmetrically above and below the true scores
- Follow the normal distribution,
- Add an absolute value to the true score to provide the observed score,
- Are **homoscedastic** (the distribution of these values is constant for A, B and C, since it doesn't change even if the attribute's intensity changes).

Reliability of the Classical Test Theory

Reliability is the extent to which test scores are *not* affected by chance factors.
Reliability Is Consistency: Reliability is the extent to which test scores are consistent
Test scores are reliable to the extent that they are consistent over ...

- different occasions of testing,
- different editions of the test, containing different questions or problems designed to measure the same general skills or types of knowledge, and
- different scorings of the test takers' responses, by different raters.

The correlation of one test (which can be a single item) with another test in the domain is called "**reliability coefficient**" (a measure of the accuracy of a test obtained by measuring the same individuals twice and computing the correlation of the two sets of measures). The correlation between test and true scores is called "**reliability index**."
(it is a statistic that provides a theoretical estimate of the correlation between actual scores of a psychometric test and the assumed true scores).

The symbol for reliability coefficient is letter *r*.
The reliability index is the square root of the reliability coefficient.
The reliability coefficient cannot exceed the reliability index: the correlation between two fallible measures cannot be higher than the correlation between a fallible measure and its true score.
The more tests that have been randomly sampled in a finite domain, the better the reliability estimates.

<u>According to the CTT, there are many ways to define reliability:</u>

1. Considering a test as a behaviour's sample, one definition of reliability is **FREEDOM FROM RANDOM ERROR**, i.e., how repeatable observations are. This definition implies homogeneity of content on multi-item tests and internal consistency or high correlations among components of the overall measure, such as items on a conventionally scored test.

2. reliability as **STABILITY OVER TIME**. Science is concerned with repeatable phenomena, which implies the repeatability of its measurements.

3. reliability as **INTERNAL CONSISTENCY**
Internal consistency is typically a measure based on the correlations between different items on the same test (or the same subscale on a larger test). It measures whether several items that propose to measure the same general construct produce similar scores. For example, if a respondent expressed agreement with the statements "I like to ride bicycles" and "I've enjoyed riding bicycles in the past", and disagreement with the statement "I hate bicycles", this would be indicative of good internal consistency of the test.

Cronbach's alpha is **a measure of internal consistency**. Cronbach's α is a simple and effective reliability estimator. We use it for multiple-choice test (like Scala Likert).

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{j=1}^{K} Var(X_i)}{Var(X_{tot})}\right)$$

In order to calculate Cronbach's α we need to know:

1. The number of items in the test
2. Single item variance
3. Total test variance

If we only have the inter-item correlation, α will be equal to **Spearman and Brown's formula**:

$$\frac{k\bar{r}_{ij}}{1 + (k-1)\bar{r}_{ij}}$$

If items are dichotomous (true/false) α is the computational equivalent of Kuder e Richardson 20 coefficient.

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{k=1}^{K} p_k(1-p_k)}{\sigma_X^2}\right)$$

To summarize:

- α represents the main split-half reliability coefficients, obtainable by splitting the test in all possible halves.
- The reliability coefficient is the expected correlation of one test with another test of the same length.
- α square root estimates the correlation between a test and the true score.
- α is equal to Spearman and Brown's formula, when we use correlations (standardized items) instead of covariances.
- α represents an estimate of the lower limit of the test's reliability (R), i.e., R≥ α.
- From a theoretical point of view, α is an important coefficient. Moreover, it is meaningful and easy to calculate.
- **It should be applied to every new test**.
- Coefficient α considers only random error within session (it requires the independence of the errors).
- When errors are correlated, α implicates an **overestimation** of the measure's reliability and thus a underestimation of measurement error.

4. **TEST-RETEST RELIABILITY**

It is the reliability of a test measured over time. In other words, give the same test twice to the same people at different times to see if the scores are the same.

Usually the scores are not identical, because of random errors. When we need to control the effect of random error among occasions, it is mandatory to use longitudinal designs (a research

design that involves repeated observations of the same variables (e.g., people) over short or long periods of time).
When the time span between the two sessions is ≤ 14 days, we talk about short-term retest reliability.
When the time span between the two sessions is ≥ 6 months, we talk about long-term reliability.
Usually, the longer the time span between test and retest, the smaller the correlation between the two sets of scores.

Retest reliability (i.e., freedom from measurement error) and construct stability (i.e., the correlation between true scores obtained in different time sessions) are two different concepts.

The measure retest reliability can be estimated with **Pearson's correlation coefficient**. It is calculated by administering the same test to the same people in different time sessions.

We consider both sessions to be «parallel» and we assume errors to be independent.

$$r_{X_1X_2} = \frac{Cov_{X_1X_2}}{\sqrt{\sigma_{X_1}\sigma_{X_2}}} = \frac{Cov_{T_1T_2}}{\sqrt{\sigma_{X_1}\sigma_{X_2}}} = \frac{\sigma_T^2}{\sigma_X^2} = r_{kk}$$

Which is an appropriate value of the reliability coefficient?
Cut-off in clinical settings should be ≥ .70 for tests which do not imply absolute judgment.
A value ≥ .90 should be used for tests which imply absolute judgment.
Test with a reliability coefficient < .70 are inappropriate. Clinicians must use tests with strong psychometric properties.
In research, cut-off should not be arbitrary. When the test reliability is .70, the Pearson correlation upper limit with another measure is .84. However, even if the reliability is .50 only, the validity upper limit would be .71 (which is appropriate too).
Considering a measure with desirable characteristics, such as good covering of the domain of interest and unidimensionality, and its scope allows a percentage of errors. Rkk < .70 is not an obstacle to use it in the research field.

This approach has some problems:

- Carryover effect may influence in a spurious way the test-retest correlation (Carryover= temporary or permanent change in a person's behaviour due to previous exposure and one or more experimental procedures, which can modify or impact the subject's performance in a subsequent experimental treatment)
- Considering a specific item, errors in test and in retest tend to be positively correlated. This causes inflation (*aumento*) in the correlation coefficient between the two sets of scores.

5. **INTER-RATER RELIABILITY**
The classical test model could have some limitations, for example when scoring a test involves a rater (in semi-structured interview or rating clinical scales). In this case, scores' replicability is not related to a single rater's internal consistency, but to an agreement among independent raters (**inter-rater reliability**).

Some designs are available to evaluate inter-rater reliability:

1. Video recording of the assessment done by different raters to different people, evaluated by independent group of observers.
2. Pairwise interview design where two or more raters, in turn, evaluate the same people and score the test independently.
   Despite the pairwise interview design has some methodological limits (i.e., assessments are barely independent), it is often used (at least for preliminary evaluation of inter-rater reliability) because of its simplicity.

Generalizability Theory is a statistical framework used to assess inter-rater reliability.

- Reliability is necessary but not sufficient to validity.
- Reliability coefficient's general equation (r_{kk}) is defined as:

$$r_{kk} = r_{xt}^2 = \frac{\sigma_t^2}{\sigma_x^2}$$

- A measure's reliability coefficient is the ratio between true score variance and total variance. The reliability coefficient's square root (r_{xt}) represents a measure's reliability index.

Generalizability Theory

Generalizability theory, or G theory, is a statistical framework for conceptualizing, investigating, and designing reliable observations. It is used to determine the reliability (i.e., reproducibility) of measurements under specific conditions. The purpose of generalizability theory is to quantify the amount of error caused by each facet and interaction of facets (facet = sources of variations, or errors).

Generalizability theory is an extension of classical measurement theory. It is an extension of domain-sampling theory to situations in which sampling proceeds factorially from more than one domain. Logically, it is closely related to issues in experimental design and ANOVA.

The focus of classical test theory (CTT) is on determining error of the measurement. Perhaps the most famous model of CTT is the equation X=T+e, where X is the observed score, T is the true score, and e is the error involved in measurement. Although e could represent many different types of error, such as rater or instrument error, CTT only allows us to estimate one type of error at a time. Essentially it throws all sources of error into one error term. This may be suitable in the context of highly controlled laboratory conditions, but variance is a part of everyday life. In field research, for example, it is unrealistic to expect that the conditions of measurement will remain constant. Generalizability theory acknowledges and allows for variability in assessment conditions that may affect measurements. The advantage of G theory lies in the fact that researchers can estimate what proportion of the total variance in the results is due to the individual factors that often vary in assessment, such as setting, time, items, and raters.

It is used mostly to assess the inter-rater reliability: one of its major applications is when one or more judges (raters) evaluate a series of individuals with respect to multiple attributes.

Two judges may disagree with each other because

- their judgments contain random measurement error,
- they can have a different standard of measurement (one may be consistently more lenient or stringent than the other),
- they can have expectations or prejudices (also related to theoretical model), which may interfere with subject's scores only in some specific instances or for some individuals. This is not random error, but a constant and systematic error and it is called bias (for example, every student heard about the rumors about a teacher giving higher marks to girls rather than boys).

A generalizability coefficient or generic reliability coefficient is a form of intraclass correlation and is symbolized p^2 .It describes how well the average judgments from a sample of one or more judges correlate with the average judgments from a population or universe of potential judges. Symbolically, p^2 may be expressed as:

$$\rho^2 = \frac{\sigma^2_{ind}}{\sigma^2_{ind} + \sigma^2_{error}}$$

A p value is defined as the ratio of true variance among individuals (σ^2_{ind}) and the sum of true variance plus random error variance (σ^2_{error}).
However, what constitute error variance depends on how the ratings are structured (single rating vs composite rating; single rater vs various raters).

To summarize, the generalizability theory allows us to define four possible coefficients whether the goal is:

- to evaluate the reliability of a single rater's evaluation (it is common in clinical setting) or of two or more raters' evaluation;
- To disentangle bias from systematic differences in assessment among raters (lenient or stringent).
- To consider random error within error variance (this approach provides stringent estimates, and it represents the most used approach in inter-rater reliability evaluation in clinical settings).

The intersection between these points leads to 4 basic coefficients for inter-rater reliability estimate. They are usually indicated as p1, p2, p3 and p4.

1. P1 and p3 assess inter-rater reliability of a single rater's score,
2. P2 and p4 assess inter-rater reliability of a sum/average of two or more rater's evaluations.
3. P1 and p2 assess the bias and the systematic error
4. P3 and p4 consider them as a part of error variance (together with random error)

If judgements are qualitative (i.e., presence or absence of a diagnosis) some models for evaluating the agreement among judges are available.
When only two raters are involved and evaluations are expressed by YES/NO (nominal scale or ordinal scale), **Cohen K** represents the index of agreement corrected by random effect (i.e., inter-rater reliability). It can be used in case of dichotomous judgements, represented with a 2x2 contingency table

In this case:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{\sum_{i=1}^{i} p_{ii} - \sum_{i=1}^{i} p_{i+} p_{+i}}{1 - \sum_{i=1}^{i} p_{i+} p_{+i}}$$

K is equivalent to p3 when data are expressed on a nominal scale (po= observed proportion, pe= expected proportion). K can be calculated by statistical software.

The inter-rater reliability coefficient vary between 0 (negative values are possible) and 1: the higher the value, the higher the inter-rater reliability:
Some guidelines are available to evaluate the size of K andp3:
Landis an Kock → < 0 = no agreement,
0-0.20 low agreement,
0.21-0.40= passable agreement,
0.41-0.60= moderate agreement,
0.61-0.80= significant agreement,
0.81-1= almost perfect agreement.

Fleiss → >.75 excellent inter-rater reliability,
values between 0.40 and 0.75 good inter-rater reliability,
<0.4 low agreement.

These guidelines are the most used in psychological field.

QUESTION 1: which coefficient should I use?
ANSWER 1: It depends on the test's goal. If it concerns a semi-structured interview, which implies the future use by different clinicians, I'll use p3 (unless you have some reason to use p1). On the other hand, if It concerns a test like "Strange Stories" (a ToM measure) which request an evaluation and an average score of every subject's responses by different raters, I'll p4 or p2.

QUESTION 2: Should I calculate the coefficient by myself?
ANSWER 2: Even if formulas are not complicated, to derive the variance's components is not so easy. The answer to this question is no. Some statistical software calculates the 4 coefficients. Some of them are free (i.e., R), some of the costs (i.e., SPSS).

The domain-sampling model

Within the scope of classical test theory, a useful model is **DOMAIN-SAMPLING MODEL.**
In psychometrics, it can be defined as the process of choosing test items that are appropriate to the content domain of the test.
With Domain Sampling, tests are constructed by randomly selecting a specified number of measures from a homogeneous, infinitely large pool. A sample of items is reliable to the extent that the score it produces correlates highly with these true scores.

The model assumes an infinite pool of possible items, but it works well as long as the pool of potential items is large.

Many authors use the term "universe" or "population" of items instead of "domain" to refer to the hypothetical collection of items. We do not do so because of the possible confusion with the universe or population of persons or other objects of measurement.