

MEDHA AI

APPSC PAPER I: GENERAL STUDIES & MENTAL ABILITY · PSYCHOLOGY

Quantitative analysis and psychological measurement

Free study notes · 1 note · 25 Oct 2026

Free study material from Medha AI. For notes, practice, mock tests and previous papers for your exam, start free at **app.medhaai.in**.

medhaai.in

Notes · Practice · Mock tests · Previous papers

Start free at
app.medhaai.in

1 Measurement in psychology: levels of measurement, statistical concepts, reliability, validity and norms

Psychological measurement assigns numbers by rules on four levels (nominal, ordinal, interval, ratio); statistics then tests hypotheses while risking Type I and Type II errors; and a good psychological test is built step by step (items, try-out, item analysis) until it is reliable (consistent), valid (measures what it claims) and has norms so that a raw score can be read as a percentile or a standard score (z, T, stanine).

Core facts

- **Measurement** = assigning numbers to objects or events according to rules. The four scales were set out by **S.S. Stevens (1946)**.
- **Scales of measurement** are also called **levels of measurement**. Memory aid: **NOIR**.
- **Nominal**: names or categories only (gender, blood group, city of birth, roll number). Average: **mode**.
- **Ordinal**: categories in **rank order**, gaps unequal (merit ranks, grades A/B/C, beginner / intermediate / fluent). Average: **median**.
- **Interval**: ranked and **evenly spaced**, but **no true zero** (temperature in °C or °F, IQ and test scores). Average: **mean**.
- **Ratio**: equal intervals **and a true zero** (height, weight, age, reaction time, temperature in **Kelvin**). Ratios such as "twice as heavy" make sense.
- **Null hypothesis (H_0)**: there is **no** difference or relationship. **Alternative hypothesis (H_1)**: there is one.
- **Level of significance (α)**: usually **0.05 (5%)** or **0.01 (1%)**.
- **Type I error**: rejecting a **true** null hypothesis (a **false positive**); its probability = α .
- **Type II error**: failing to reject a **false** null hypothesis (a **false negative**); its probability = β .
- **Correlation (r)** runs from **-1 to +1**; the sign is the **direction**, the size is the **strength**. **-0.5 to -1** = a **strong negative** correlation (Cohen: 0.5 and above is large; Guilford's older scale calls 0.4-0.7 moderate).
- **Psychological test**: an **objective and standardised measure of a sample of behaviour** (Anastasi's idea).
- **Reliability** = **consistency** of scores. **Validity** = the test measures **what it claims** to measure.
- **Norms** = the **average performance of a representative (standardisation) group**, used to interpret a raw score.
- **Grade norm** = the **mean raw score** of pupils in a given grade. **Age norm** = the mean score of children of a given age.
- **Standard score** = the distance of a person's score from the **mean**, in **standard deviation** units.
- **z** = $(X - \text{Mean}) \div \text{SD}$ (mean 0, SD 1). **T** = $50 + 10z$ (mean 50, SD 10). **Stanine** = "standard nine", 1 to 9, mean 5, SD about 2.

- **Deviation IQ** = $100 + 15z$ (mean 100, SD 15; Wechsler).

Explained

How to compute the mean, median, SD and r is in **STA-CT-01**, **STA-DK-01**, **STA-DK-02** (normal curve) and **STA-CR-01**; a classroom summary is in **EDU-ST-01**. This note covers only what a psychology paper asks beyond the arithmetic.

1. The four levels of measurement

Scale	Properties	Examples	Statistics allowed
Nominal	identity (categories)	gender, religion, car brand, marital status	mode, frequencies, chi-square
Ordinal	+ order	ranks, Olympic top five, Likert ratings	+ median, percentiles, Spearman rank r
Interval	+ equal intervals	°C, °F, IQ, exam and personality scores	+ mean, SD, Pearson r , t-test
Ratio	+ true zero	height, age, weight, Kelvin, reaction time	+ ratios, geometric mean, coefficient of variation

- Each scale has **all the properties of the one above it**, plus one more.
- **True zero** means zero is the complete absence of the thing: 0 kg is no weight, but 0 °C is not "no temperature" and an IQ of zero is not "no intelligence". So most psychological scores are treated as **interval**, not ratio.
- **Likert-type ratings** (strongly agree ... strongly disagree) are strictly **ordinal**; in practice their totals are often treated as interval.
- Variables may also be **discrete** (counted in whole numbers: number of children) or **continuous** (any value: time, height).

2. Statistical ideas in words

- **Descriptive statistics** summarise data (mean, SD, frequency table). **Inferential statistics** go from a **sample** to the **population** (tests of significance, estimation).
- A **parameter** describes a population; a **statistic** describes a sample.
- **Parametric tests** (t-test, ANOVA / F-test, Pearson r) assume interval or ratio data and a roughly **normal** distribution. **Non-parametric (distribution-free) tests** (chi-square, Mann-Whitney U, Spearman's rho) suit nominal or ordinal data or small, skewed samples.
- A **distribution** is the set of all the n scores in a dataset, arranged by how often each value occurs.

3. Hypotheses and errors

- Variables, kinds of hypotheses and sampling are introduced in **PSY-ME-01**; here is how a hypothesis is tested.
- A **hypothesis** is a testable statement about the relationship between two or more variables. Example: "Students who sleep 8 hours recall more words than those who sleep 4 hours."

- H_0 : sleep makes no difference to recall. H_1 : it does.
- **Directional (one-tailed)** H_1 states the direction ("8-hour sleepers recall **more**"); **non-directional (two-tailed)** says only that there is **a difference**.
- The researcher **tests the null hypothesis**. If the result would be very unlikely by chance (p less than α), H_0 is **rejected** and the result is called **statistically significant**. Otherwise H_0 is **retained** ("not rejected"; we never "prove" it).
- **p-value**: the probability of getting a result this large if H_0 were true.

Decision	H_0 actually true	H_0 actually false
Reject H_0	Type I error (α)	correct decision (power)
Retain H_0	correct decision	Type II error (β)

- Court analogy: H_0 = "the accused is innocent". Convicting an innocent person = **Type I**; acquitting a guilty person = **Type II**.
- **Power** = $1 - \beta$: the chance of detecting a real effect. Larger samples raise power.
- Making α stricter ($0.05 \rightarrow 0.01$) **reduces Type I errors** but **increases Type II errors**.

4. Reading the size of r

r (either sign)	Guilford's description	Cohen's convention
below 0.20	slight, almost negligible	below 0.10: trivial
0.20–0.40	low	about 0.10: small
0.40–0.70	moderate, substantial	about 0.30: medium
0.70–0.90	high, marked	0.50 and above: large (strong)
0.90–1.00	very high	

- -0.8 is **stronger** than $+0.6$: the sign only gives direction.
- $r = 0$ means no **linear** relation; a curved relation can still exist.
- r^2 (coefficient of determination) = the share of variance the two variables have in common: $r = 0.6 \rightarrow 36\%$.

5. Steps in constructing a standardised test

1. **Planning**: state the purpose, the group it is for, the content and the item format; draw a **blueprint (table of specifications)**.
2. **Writing items**: prepare more items than needed (an item pool); objective types (multiple choice, true–false, matching, completion) or essay; get experts to review them.
3. **Preliminary try-out (pilot)**: give the draft to a sample like the final group; fix unclear items and set the time limit.
4. **Item analysis**: keep items of suitable **difficulty** that **discriminate** between high and low scorers; check distractors.
5. **Final form**: arrange the selected items (usually easy to hard) with standard instructions and a scoring key.

6. **Reliability and validity:** check them on a fresh sample.

7. **Norms:** give the test to a large **representative** standardisation sample and prepare norm tables.

8. **Manual:** describe administration, scoring, norms, reliability and validity.

6. Item analysis

- **Difficulty index (p)** = the proportion of examinees who answer the item **correctly**. $p = 0.9$ is an **easy** item; $p = 0.2$ a **hard** one. Items around **0.5** spread scores best; many test-makers keep items between about **0.2 and 0.8**.
- **Discrimination index (D)** = (correct answers in the **upper** group – correct answers in the **lower** group) ÷ number in one group. The upper and lower groups are usually the top and bottom **27%** (T.L. Kelley). D runs from **-1 to +1**. Common rule: 0.40 or more very good; 0.30–0.39 good; 0.20–0.29 revise; below 0.20 poor; **negative D** → **discard** (weak students do better on it than strong ones).
- Example: 12 pupils in each group; 11 of the upper and 5 of the lower answer correctly. $p = (11 + 5) \div 24 \approx \mathbf{0.67}$; $D = (11 - 5) \div 12 = \mathbf{0.5}$: a good item.
- **Distractor analysis:** a wrong option that nobody chooses, or that strong pupils choose more than weak pupils, needs rewriting.

7. Reliability: kinds

Kind	How it is found	Coefficient of	Main problem
Test-retest	same test, same people, twice	stability	memory and practice; change over the interval
Parallel (alternate, equivalent) forms	two equivalent forms	equivalence	hard to build truly equal forms
Split-half	one sitting; odd vs even items, corrected by Spearman–Brown	internal consistency	depends on how the test is split
Kuder–Richardson / Cronbach's alpha	from all the items at once	internal consistency	assumes items measure one trait

- **Kuder–Richardson** formulas **KR-20 and KR-21 (1937)** are for items scored **right / wrong**; **Cronbach's alpha (1951)** extends this to items with several points (rating scales).
- **Inter-rater (scorer) reliability:** agreement between two or more scorers; essential for essays, observation and projective tests.
- Reliability coefficients run from **0 to 1**; about **0.70+** is accepted for research and **0.90+** for decisions about individuals.
- **Longer tests** are more reliable (more items average out chance); a **more varied (heterogeneous) group** gives a higher coefficient; guessing, unclear items and poor conditions lower it.

8. Validity: kinds

Kind	Question it answers	How it is checked
Face	does it look as if it measures the trait?	judgement of test-takers; not a technical proof

Kind	Question it answers	How it is checked
Content	do the items cover the whole field fairly?	experts compare items with the blueprint
Criterion-related	do scores agree with an outside criterion ?	correlation with the criterion (validity coefficient)
Construct	does it measure the theoretical trait (e.g. anxiety)?	many lines of evidence (convergent, discriminant, factor analysis)

- **Concurrent validity**: criterion measured **at the same time** (a new depression scale against a psychiatrist's present diagnosis).
- **Predictive validity**: criterion measured **later** (an entrance test against marks at the end of the first year; an aptitude test against later job performance).
- **Construct validity** (Cronbach and Meehl, 1955): **convergent** = high correlation with tests of the same trait; **discriminant (divergent)** = low correlation with tests of different traits.
- **Reliability and validity**: a test **must be reliable to be valid**, but a reliable test **need not be valid**. A clock that is always 10 minutes fast is reliable (consistent) but not valid (accurate). Reliability is necessary, not sufficient.

9. Norms

- A raw score (say 42) means nothing alone; norms tell how it compares with the reference group.
- **Norm-referenced** tests compare a person with others; **criterion-referenced** tests compare a person with a fixed standard (e.g. "can type 40 words a minute").

Norm	Built from	Example of reading
Age norms	mean score of each age group	a score equal to the 10-year average = "age equivalent 10" (mental age)
Grade norms	mean score of pupils in each grade	reading score = class 6 average
Percentile norms	% of the group scoring below each raw score	percentile rank 80 = better than 80%
Standard-score norms	distance from the mean in SD units	z, T, stanine, deviation IQ

- **Percentile** is not **percentage**: 70% correct is a percentage; PR 70 means a higher score than 70% of the norm group.
- Norms may be **national**, **regional** or **local**, and must come from a sample like the people being tested.

10. Standard scores

Score	Formula	Mean	SD
z	$(X - M) \div SD$	0	1
T	$50 + 10z$	50	10

Score	Formula	Mean	SD
Deviation IQ	$100 + 15z$	100	15
Stanine	9 bands, each half an SD wide	5	about 2

- Example: a test with mean 60 and SD 10. A raw score of **75** gives $z = +1.5$, $T = 65$, deviation IQ = **122.5**, stanine **8**. A raw score of **45** gives $z = -1.5$, $T = 35$, deviation IQ = **77.5**, stanine **2**.
- **Stanine** shares of a normal group, from 1 to 9: **4, 7, 12, 17, 20, 17, 12, 7, 4 per cent** (total 100). Stanine 5 is the middle 20%. Stanines were used widely by the US Army Air Forces in the Second World War.
- **Ratio IQ** (older) = Mental Age $\div$ Chronological Age $\times$ 100 (Stern's quotient, multiplied by 100 in Terman's Stanford-Binet, 1916). **Deviation IQ** (Wechsler) replaced it because ratio IQ does not work for adults.
- Standard scores let us **compare** results on different tests (a T of 60 in English and a T of 55 in Mathematics: the English result is relatively better).

Common traps

✗ "Temperature in Celsius is a ratio scale." ✓ It is **interval** (0 °C is not "no heat"); **Kelvin** is ratio.

✗ "IQ is a ratio scale, so IQ 140 is twice IQ 70." ✓ IQ is treated as **interval**; such ratios are meaningless.

✗ "Ranks are interval data." ✓ Ranks are **ordinal**: the gaps between them are not equal.

✗ "Type I error is accepting a false null hypothesis." ✓ Type I = **rejecting a true** H_0 ; accepting a false H_0 is **Type II**.

✗ "-0.8 is weaker than +0.6." ✓ Strength is the **size**; -0.8 is the **stronger** correlation.

✗ "A reliable test is always valid." ✓ Reliability is **necessary but not sufficient** for validity.

✗ "Test-retest gives the coefficient of equivalence." ✓ Test-retest = **stability**; parallel forms = **equivalence**.

✗ "Face validity is the strongest evidence." ✓ It is only appearance; **construct** validity is the most complete.

✗ "A norm is the score every student should reach." ✓ A norm is the **average performance** of a representative group, not a target.

✗ "Percentile rank 70 means 70% marks." ✓ It means scoring **higher than 70% of the group**.

✗ "T-score has mean 0." ✓ **z** has mean 0, SD 1; **T** has mean **50**, SD **10**.

Quick revision

- NOIR: nominal (names, mode), ordinal (ranks, median), interval (equal gaps, no true zero, mean), ratio (true zero).
- °C and IQ = interval; Kelvin, height, age, weight = ratio.
- H_0 = no difference; Type I = rejecting true H_0 (α); Type II = retaining false H_0 (β).
- r from -1 to $+1$; ± 0.5 to ± 1 strong; sign = direction only.
- Test construction: plan → items → try-out → item analysis → final form → reliability, validity → norms → manual.
- Difficulty p = proportion correct; discrimination $D = (U - L) \div n$, upper/lower 27%.
- Reliability: test-retest (stability), parallel forms (equivalence), split-half + Spearman-Brown, KR-20, alpha, inter-rater.
- Validity: face, content, criterion (concurrent now, predictive later), construct.
- Norms: age, grade (mean of a grade), percentile, standard score.
- $z = (X - M) \div SD$; $T = 50 + 10z$; stanine 1-9, mean 5; deviation IQ $100 + 15z$.