PSY2041 Psychological testing and assessment

PSYCHOLOGY · MONASH UNIVERSITY · LEVEL 2

Full guide + ask Sia about PSY2041

SIDE 1/2 WHAT MUST BE IN YOUR HEAD · define → compare → calculate → bound the inference CLOSED-BOOK MEMORY SHEET S2 2026 · Examination 40%

0 Exam map

COURSE DATA · BRIEF

LAB 1 · 20% | LAB 2 · 40% | EXAM · 40%

ASSESSMENT	WEIGHT
1 - Written: Lab Report Part 1	20%
2 - Written: Lab Report Part 2	40%
3 - Examination	40%

PERMITTED-MATERIALS TRUTH

"The exam is closed-book (no notes or reading materials permitted without prior arrangement)" — PSY2041 Additional information and resources page, Moodle 2026 (MD_Additional_information_and_resources_md.md).

Final: 100 minutes, 60 marks; multiple-choice, short-answer and extended-answer sections. Build recall around the exact discriminator, then show the evidence boundary.

1 Construct → inference

WK 1

construct — behaviour — score — inference

standardisation + evidence support

A score is evidence about a construct; it is not the decision itself.

FIVE-PART TEST AUDIT

Relevant **behaviour sample** · objective/standardised procedure · quantitative score · interpretable reference · adequate psychometric evidence.

Construct = theoretical attribute inferred from behaviour. Standardisation controls administration and scoring; it improves comparability but does not erase error or guarantee fairness.

REFERENCE	QUESTION ANSWERED	STATE EXPLICITLY
Criterion	met a defined standard/cut point?	standard
Norm	where is score in comparison population?	norm group

EXAM MOVE *Name construct, sampled behaviour, scoring rule, reference and available evidence before trusting the test label.*

1A Inference-boundary cards

WK 1

IF THE STEM GIVES...	YOU MAY SAY...	NOT YET...
identical instructions + scoring	administration/scoring are standardised	scores are reliable or fair
adult reference sample	norm comparison is available	norms fit every population
numerical output	a quantitative score exists	construct meaning is established
one behaviour sample	evidence about sampled behaviour	complete description of person

For culture/fairness prompts use **mechanism → consequence → safeguard**: language or task familiarity may introduce construct-irrelevant variance; interpretation may be distorted; use accessible language, suitable norms and additional evidence.

2 Testing ≠ assessment

WK 1

ASSESSMENT DECISION

referral + several sources + context

TEST SCORE

Testing produces a score; assessment integrates evidence and uncertainty.

TESTING	ASSESSMENT
instrument; standard administration; score	referral question; tool selection; interview/observation/records; integration; professional judgment; limitations

FAIRNESS AUDIT

Check population fit, suitable and current norms, accessible delivery, language/task familiarity, educational opportunity and stereotype threat. The same procedure may be consistent yet support unequal conclusions.

BOUNDARY SENTENCE

"This score contributes to the decision for this population and purpose; it does not determine the person."

3 Reliability forms

WK 2

stable score variance | error

observed = stable + error

Reliability rises as stable variation occupies more of observed variation.

CLAIM/COMPARISON	EVIDENCE	THREAT
occasion ↔ occasion	test–retest	real change, memory, interval
rater ↔ rater	interrater	rules, training, ambiguity
item ↔ item	internal consistency	heterogeneity, coding
form ↔ form	equivalent forms	non-equivalent content

Name what was compared before the coefficient. One reliability form never establishes the others; scope includes population, conditions and interval.

LOW COEFFICIENT → MATCHED REMEDY

WEAK RESULT	INVESTIGATE	POSSIBLE ACTION
rater agreement	ambiguous categories/rules	clarify anchors; train raters
test–retest	change, interval, memory, administration	justify interval; standardise
item coherence	coding, multidimensionality, broad construct	check blueprint before removal

4 Validity argument

WK 2

reliable, not valid | reliable + valid evidence

neither | valid use needs consistency

Consistency is necessary for a useful interpretation, never sufficient.

VALIDITY IS NOT A BADGE

Ask whether evidence supports the proposed **meaning and use** of scores in this population and context.

EVIDENCE SOURCE	QUESTION
Content	Does the domain appear in the tasks/items?
Response process	Are people responding as intended?
Internal structure	Do score patterns fit proposed dimensions?
Criterion	Relation to relevant external outcome?
Construct relations	Do patterns behave as theory predicts?

Reliably repeating biased or irrelevant results remains invalid for the claim. A new population, criterion, timing or decision can require fresh evidence.

EXTENDED ANSWER *Define consistency → define validity support → name the relationship → apply evidence → state the bounded inference.*

5 Claim → design

WK 2 WORKED

CLAIM	COMPARISON THAT TESTS IT
coherent confidence items	items with items
stable over days	occasion with occasion
predicts later help-seeking	earlier score with independently measured later criterion

Trap: short-term stability establishes test–retest reliability only. It does not establish predictive criterion validity.

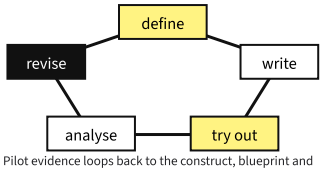
VALIDITY WORDING LADDER

TOO BROAD	REPAIR
"the test is valid"	evidence supports this score interpretation/use
"content experts approved it"	content evidence only; other sources remain
"correlates with outcome"	criterion evidence; check criterion quality and timing
"high reliability proves construct"	consistency limits error; it does not identify meaning

End with population, timing and purpose: "supports stability across this interval in this sample," or "supports prediction of this independently measured criterion in the intended population."

6 Test development cycle

WK 3



Pilot evidence loops back to the construct, blueprint and items.

1. **Conceptualise:** purpose, population, construct, intended use.
2. **Blueprint:** map content facets and required behaviours.
3. **Construct:** instructions, response format, items, scoring rules.
4. **Try out:** realistic sample and conditions; ask how respondents understood items.
5. **Analyse + revise:** difficulty, discrimination, distractors, coverage, burden and fairness; pilot changes again.

DIFFICULTY INDEX

p = number correct ÷ tryout N. Higher p means easier for that sample.

ITEM EVIDENCE	READ IT AS	DO NOT
difficulty	proportion correct	call .70 "70% difficult"
discrimination	differentiates relevant performance	delete on one number
distractors	misconceptions/plausibility	ignore never-chosen options

7 Item-writing checks

WK 3

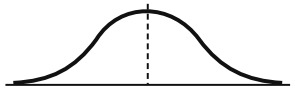
- One clear behaviour; language matched to population; no irrelevant reading load.
- Parallel, plausible options; no cueing; open response has an explicit scoring rule.
- Double-barrelled? Split only if both behaviours belong in the blueprint; otherwise remove off-construct content.

DECISION RULE *Evidence → diagnosis → targeted edit → new pilot check. Psychometric numbers inform judgment; they do not replace the blueprint.*

DEFECT	WHY IT MATTERS	REPAIR
double-barrelled	one response cannot identify which claim	split only if both are on blueprint
implausible distractor	item functions with fewer real options	replace using misconception evidence
irrelevant reading load	samples reading, not target skill	simplify without changing demand
unclear reference	respondents answer different questions	specify time/context

8 Describe before infer

WK 4



mean = median

Symmetry centres mean and median; skew pulls the mean toward the tail.

SCALE	INFORMATION PRESERVED	USEFUL SUMMARY CUE
Nominal	categories	mode/frequency
Ordinal	categories + order	median/ranks
Interval	equal intervals	mean/SD if shape permits
Ratio	equal intervals + meaningful zero	full arithmetic

Positive skew: long right tail, commonly mean > median. **Negative skew:** long left tail, commonly mean < median. Extreme observations move mean and spread more than median and IQR.

STANDARD SCORE

z = (X - M) ÷ SD. Sign = above/below mean; magnitude = distance in SD units. Not a percentage or percentile.

SPREAD	MEANING
Range	maximum − minimum; extreme-sensitive
IQR	spread of ordered middle half
SD	typical dispersion around mean in original unit

9 Correlation boundary

WK 4

Pearson correlation standardises covariance and describes direction and strength of a linear association from −1 to 1. Inspect outliers, restricted range and nonlinearity.

REPORT

variables + direction + approximate strength + linear form + sample/reference frame + causal boundary.

Correlation does not establish direction or cause: reverse direction, third variables, selection and measurement problems remain possible.

SUMMARY CHOICE DRILL

DATA CUE	BEST FIRST MOVE	REASON
unordered categories	frequencies/mode	codes do not carry distance
ordered responses	median/ranks	order survives; gaps need not
strong skew/outlier	median + IQR; show shape	robust centre/spread
roughly symmetric quantitative scores	mean + SD	distance information is useful
two quantitative variables	plot before coefficient	see form, range and influential points

10 Hypothesis logic

WK 5

test statistic under H_0

$p \geq \alpha$ fail to reject | $p < \alpha$ reject H_0

neither branch proves a population claim

The decision is conditional on the null model, assumptions and stated α.

P VALUE

Probability, if H_0 and analysis conditions hold, of a result at least as extreme as observed. It is not P(H_0 true), replication probability or effect size.

REALITY	REJECT H_0	FAIL TO REJECT
H_0 true	Type I error	correct
H_0 false	correct detection	Type II error

Power = probability of detecting an effect of specified size when it exists. A nonsignificant result may reflect a small effect, noise, limited power or a negligible difference; report estimate and confidence interval.

11 Choose the t design

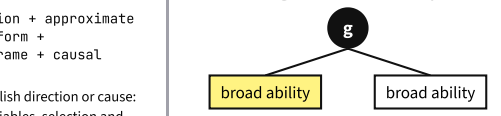
WK 5

SCORE STRUCTURE	PROCEDURE	TARGET
one sample vs reference	one-sample t	mean − reference
two unrelated groups	independent-samples t	group mean difference
same/linked people twice	paired-samples t	mean linked difference

FAST RULE *Identify who contributes each score before reading the scenario topic. Design chooses the error term.*

12 Intelligence hierarchy

WK 6



narrow skills below broad abilities

Global, index and subtest scores answer questions at different levels.

Begin with referral → check precision and pattern coherence → identify relative and normative strengths/weaknesses → consider task demands, language, education, health, fatigue and motivation → integrate behaviour/history → recommend cautiously.

COURSE REFERENCE

Intelligence scores commonly use mean 100, SD 15. A percentile is rank in a norm group, not percentage intelligence.

PSY2041 Psychological testing and assessment

SELF-TEST DRILLS · JUDGEMENT CARDS · EXAM-MORNING RETRIEVAL

Full guide + ask Sia about PSY2041

SIDE 2/2 DRILLS · cover the answers → work the path → name the trap → retest tomorrow

CLOSED-BOOK MEMORY SHEET

60 marks · 100 minutes

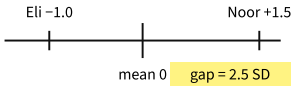
13 Worked drill 1 · z scores

WK 4 WORKED

Prompt. Mean 50, SD 8. Noor = 62; Eli = 42. Compute and interpret.

PATH

```
Noor: z = (62 - 50) ÷ 8 = 12 ÷ 8 = 1.50.
Eli: z = (42 - 50) ÷ 8 = -8 ÷ 8 = -1.00.
```



Direction comes from the sign; magnitude is distance from the reference mean.

Answer. Noor is 1.5 SD above the mean; Eli is 1 SD below. Their standardised positions differ by 2.5 SD in this distribution.

TRAP *Do not translate z into percentage correct or percentile rank.*

14 Worked drill 2 · item index

WK 3 MODEL

Prompt. 84 of 120 students answer correctly.

```
p = 84 ÷ 120 = 0.70 → 70% answered correctly; the item is relatively easy for this tryout sample.
```

Next: compare relevant high/low performers, inspect distractors and check blueprint coverage. Weak discrimination may reflect keying, wording, an implausible distractor, sample mismatch or essential mastery content.

IF YOU SEE	DO
high p	say easier, not "more difficult"
weak discrimination	diagnose before delete
revision	pilot again

15 Distribution traps

WK 4

WHICH IS TRUE?	CORRECT CONTRAST
8, 9, 9, 10, 34	median 9; mean 14; right outlier pulls mean
negative z	below reference mean, not negative construct
r near 0	little linear association, not necessarily no relation
strong r	association, not causation

15A Exact-noun retrieval

SHORT-ANSWER TOPICS

CUE	ANSWER	NEAREST RIVAL
same people, suitable interval	test–retest	interrater
score against defined cut point	criterion reference	norm reference
right-tail distribution	positive skew	negative skew

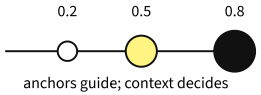
16 Worked drill 3 · Cohen's d

WK 8 WORKED

Prompt. Group A mean 28; Group B mean 24; pooled SD 8. Define direction A − B.

PATH

```
d = (M_A - M_B) ÷ SD_pooled
= (28 - 24) ÷ 8 = 4 ÷ 8 = 0.50.
```



The course benchmark sequence is 0.2, 0.5 and 0.8; it orients, not settles, practical meaning.

Answer. Group A scored 0.50 pooled SD higher than Group B: conventionally medium, with practical importance requiring context.

Reverse the subtraction and the magnitude stays 0.50 while the sign changes. Print the group order first.

17 Evidence · magnitude · precision

WKS 5, 8

LAYER	QUESTION	REPORT
Evidence	how incompatible with H_0?	test statistic, p, decision
Magnitude	how much?	difference, d or r
Precision	which values remain plausible?	confidence interval

A large sample can make a tiny effect statistically significant. A small study can leave a meaningful estimate imprecise. Significance never converts a small effect into a large one.

ASSUMPTION EXPLANATION

Define the condition → name the data object it concerns → explain the standard error/reference-distribution link → state a plot or design check → say how violation could distort inference.

CONDITION	WHERE IT BITES
independence	study design/observation unit
shape + outliers	scores or paired differences
variance comparability	independent-group SE
linearity	Pearson r's target

T-PROCEDURE CARDS

STEM	CHOICE	WHY
class mean vs published reference	one-sample t	one sample, fixed comparison
different participants in two groups	independent t	unlinked scores
same participants before/after	paired t	analyse linked differences

18 Worked drill 4 · signed ranks

WK 9 WORKED

Prompt. After − before differences: +2, −1, +4, −3.

DIFFERENCE	|D|	RANK	SIGNED RANK
+2	2	2	+2
−1	1	1	−1
+4	4	4	+4
−3	3	3	−3

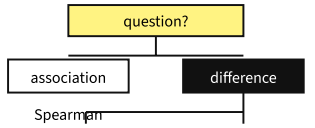
ANSWER PATH

```
Absolute differences → rank smallest to largest → restore original signs → W⁺ = 2 + 4 = 6; |W⁻| = 1 + 3 = 4.
Tiny example is descriptive only.
```

TRAP *Do not rank signed values directly. Remove zeros; average tied ranks; keep subtraction direction visible.*

19 Rank-method chooser

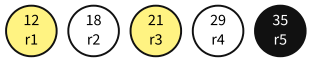
WK 9



Association versus difference first; then linked versus independent observations.

METHOD	TARGET	REPORT AS
Spearman	monotonic association in ranks	ordered association
signed-rank	paired/reference differences	direction/ranks of linked differences
rank-sum	two independent groups	relative group ranks

Ranks preserve order but give up raw spacing. They reduce dependence on distribution form; they do not repair dependence, biased sampling or the wrong question.



Rank transformation protects relative position, not the raw distances.

20 Method traps

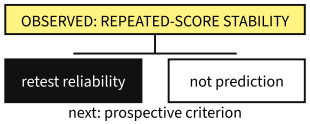
WKS 5, 9

IF YOU SEE...	DO THIS
"not normal" only	name scale, shape, outlier/form issue
same people twice	analyse linked differences
monotonic curve	consider Spearman, not Pearson by habit
rank result	do not silently report a mean claim

21 Worked drill 5 · score claim

COURSE WORKED

Prompt. A screening scale is stable across one week; developer claims it predicts later clinical need.



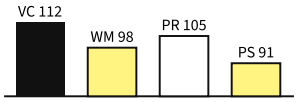
The comparison determines the evidence type and the next design.

Path. Stable construct expected → repeated occasions support test–retest reliability over that interval → later clinical need is a separate criterion → follow initial score prospectively → compare with an independently measured later outcome in intended population → check precision and alternatives.

Answer. Current evidence concerns short-term consistency only. A prospective criterion study is needed before the advertised prediction.

22 Intelligence profile drill

WK 6 WORKED



Illustrative profile: verbal strongest; processing speed lowest.

Referral: timed written work. Describe relative pattern → connect slower processing cautiously → check confidence intervals, discrepancy frequency and broad-score coherence → consider motor, visual, fatigue, anxiety, language and history → corroborate → recommend without diagnosis.

DISTINGUISH *Relative weakness within this profile versus low standing relative to the norm group.*

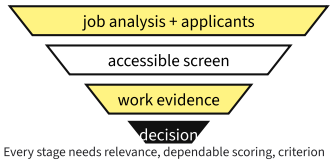
23 Educational decision cards

WK 7

IF EVIDENCE IS USED TO...	NAME
change learning while developing	formative
judge achievement after instruction	summative
flag need for fuller assessment	screening
repeat brief measures and adjust support	progress monitoring
sample learned knowledge/skills	achievement
estimate capacity for future learning/performance	aptitude

24 Organisational testing

WK 10



Every stage needs relevance, dependable scoring, criterion evidence and fairness review.

START HERE

```
Job analysis defines critical tasks and attributes; without it, relevance cannot be judged.
```

METHOD	SAMPLES	TRADE-OFF
job-knowledge	knowledge needed on entry	opportunity/exposure
skill test	defined competence	narrow task
situational judgment	responses to realistic situations	scoring/relevance
work sample	representative task performance	cost/admin burden
assessment centre	several exercises/assessors	resources/training

DISPATCHER APPLICATION

Job analysis → accessible screen of job-relevant processing/prioritisation → finalist simulation of cases, entry and communication → multiple trained assessors with anchored scoring → validate against later performance and added information → monitor subgroup outcomes, access and privacy; provide review route.

TECHNOLOGY *Scale can accelerate bias. Check transparency, privacy, security, accessibility, job relevance, unequal error and human review.*

25 Clinical layers

WK 11



Interview, current observations and standardised tools contribute different evidence.

LAYER	PURPOSE	BOUNDARY
MSE	describe current appearance/behaviour, speech, mood/affect, thought, perception, cognition, insight/judgment	observation ≠ diagnosis
Diagnosis	classify pattern/course against criteria	check impairment and alternatives
Formulation	explain individual pattern	revisable hypothesis

26 Clinical judgement cards

WK 11

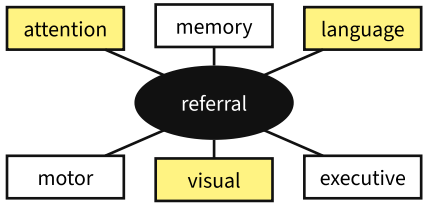
Observation drill. Quiet speech, frequent pauses, downward gaze. Do not write "depressed and uncooperative." Record the behaviours; ask about mood, anxiety, fatigue, pain, language comfort, culture and setting; integrate history, MSE domains and measures.

IF YOU SEE...	DO THIS
high symptom score	say symptoms justify fuller assessment; do not diagnose from score alone
screening threshold	consider false-positive/false-negative consequences and follow-up
diagnosis requested	criteria + course + impairment + differential explanations + context
sources disagree	ask what each sampled; gather resolving evidence
interpretive adjective	replace with observable noun/verb first

INTERVIEW ORDER *Purpose/consent/confidentiality → open account → focused onset/frequency/severity/triggers/impairment/coping/risk → context → summary check → plan.*

27 Neuropsychological battery

WK 12



One low score cannot carry a domain, diagnosis or localisation claim.

Battery selection starts from the referral, relevant domains and suitable norms. Interpret quantitative pattern plus strategy, error type, persistence and behaviour; integrate history and collateral information.

TASK ANCHORS

```
Wisconsin Card Sorting: shifting/rule discovery · Tower of London: planning · clock drawing: visuospatial, constructional and executive demands · grooved pegboard: fine motor speed/dexterity. Each also depends on other processes.
```

Sixty-minute sequence: concise interview → manageable attention → difficult planning task early, ideally within first 30 minutes → verbal learning while attention is adequate → break if needed → interpret later motor speed with fatigue, pain and sensory/motor constraints. In the worked scenario, fatigue rises after 40 minutes.

LOW-SCORE TRAP → FIX *Never map one low score directly to a damaged area. Check task demands, premorbid ability, language, education, mood, fatigue, pain, medication and sensory or motor constraints. Seek convergence across other tasks, error patterns, strategy, history and collateral evidence. Then give a bounded functional recommendation or name the next hypothesis to test.*