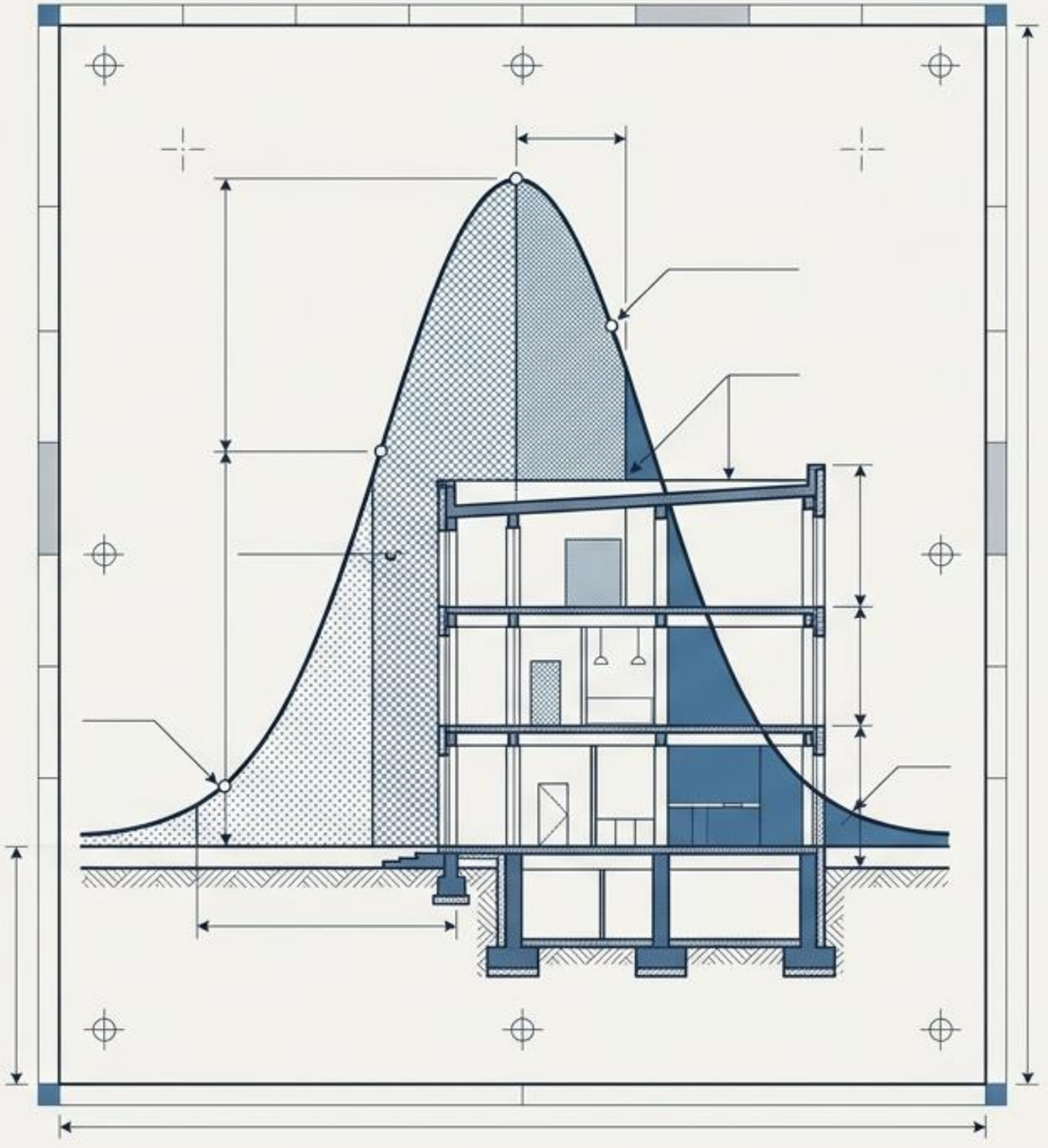
The Blueprint of Precision

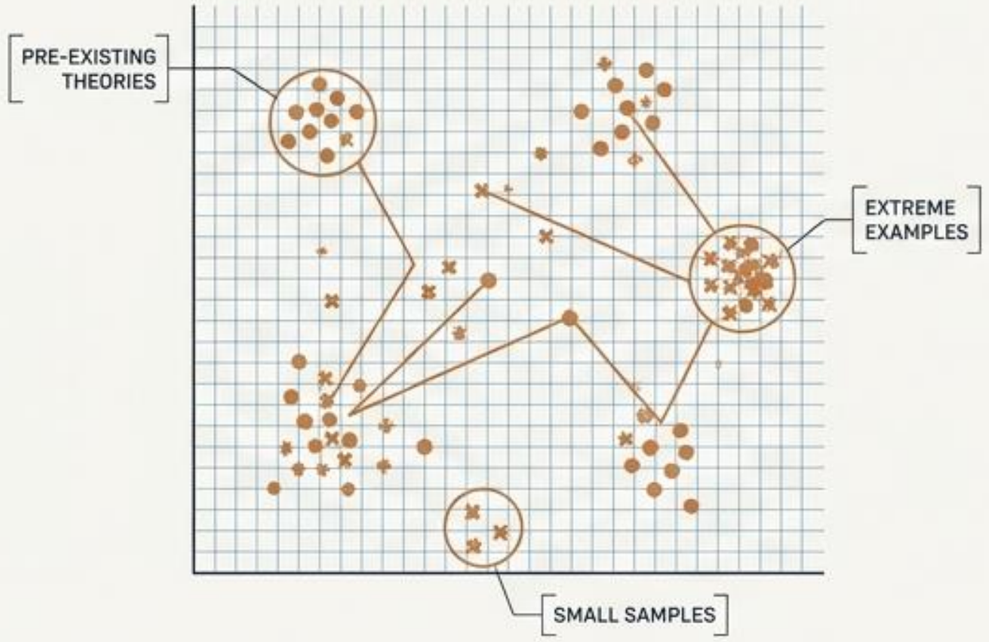
A Practitioner's Guide to Psychometric Assessment and Objective Decision-Making

Aligned with the Code of Ethics for Psychologists Working in Aotearoa New Zealand



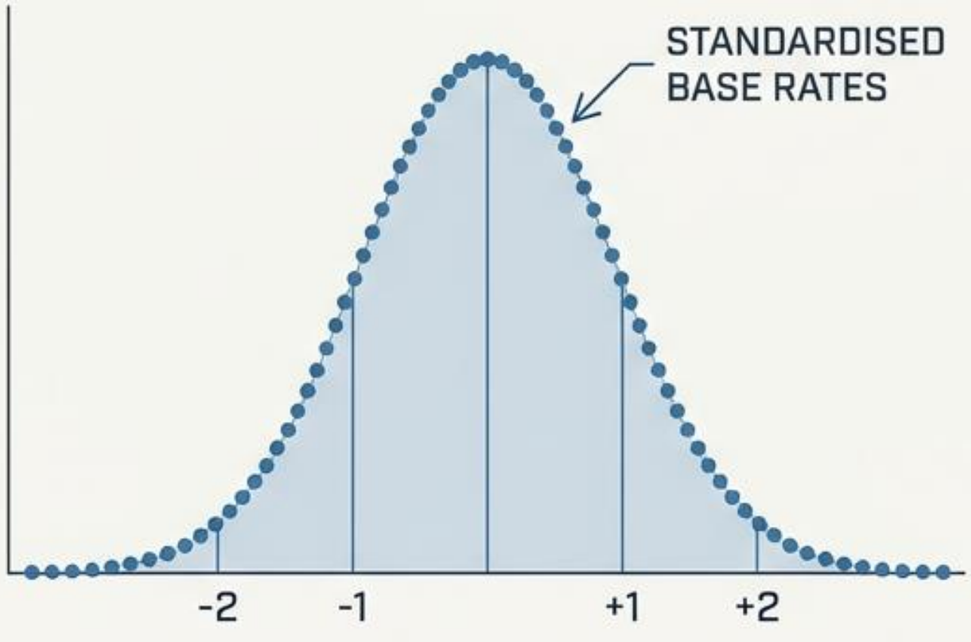
THE BASE RATE IMPERATIVE

FLAWED ATTRIBUTION



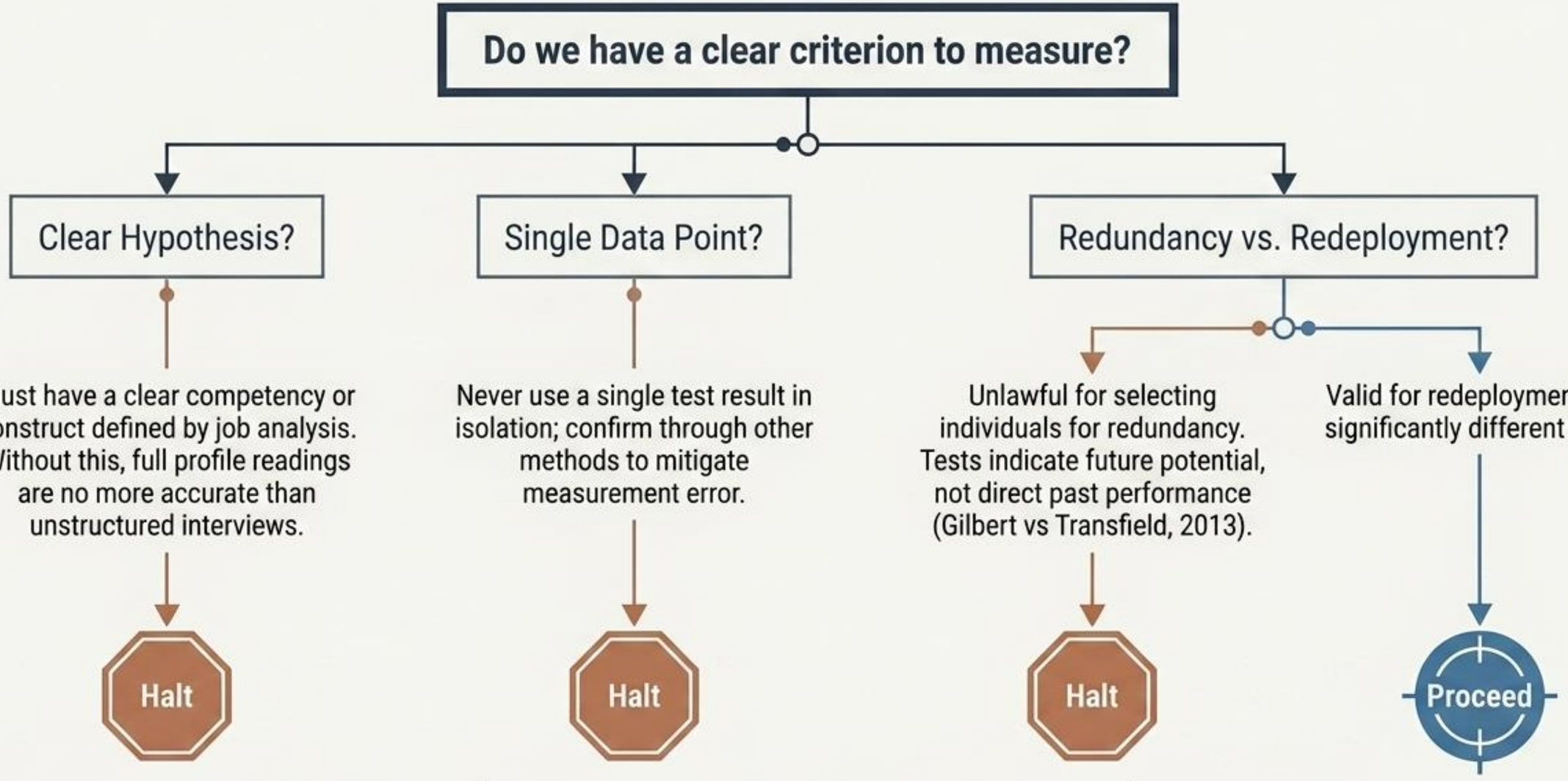
Human judgement naturally seeks to confirm pre-existing theories and is overly influenced by extreme, atypical examples or small sample sizes. Base rate information is systematically underutilised.

OBJECTIVE PROPENSITY



Standardised assessments provide valid base rate data, comparing an individual's behavioural sample against a large population to measure relative propensity with known error margins.

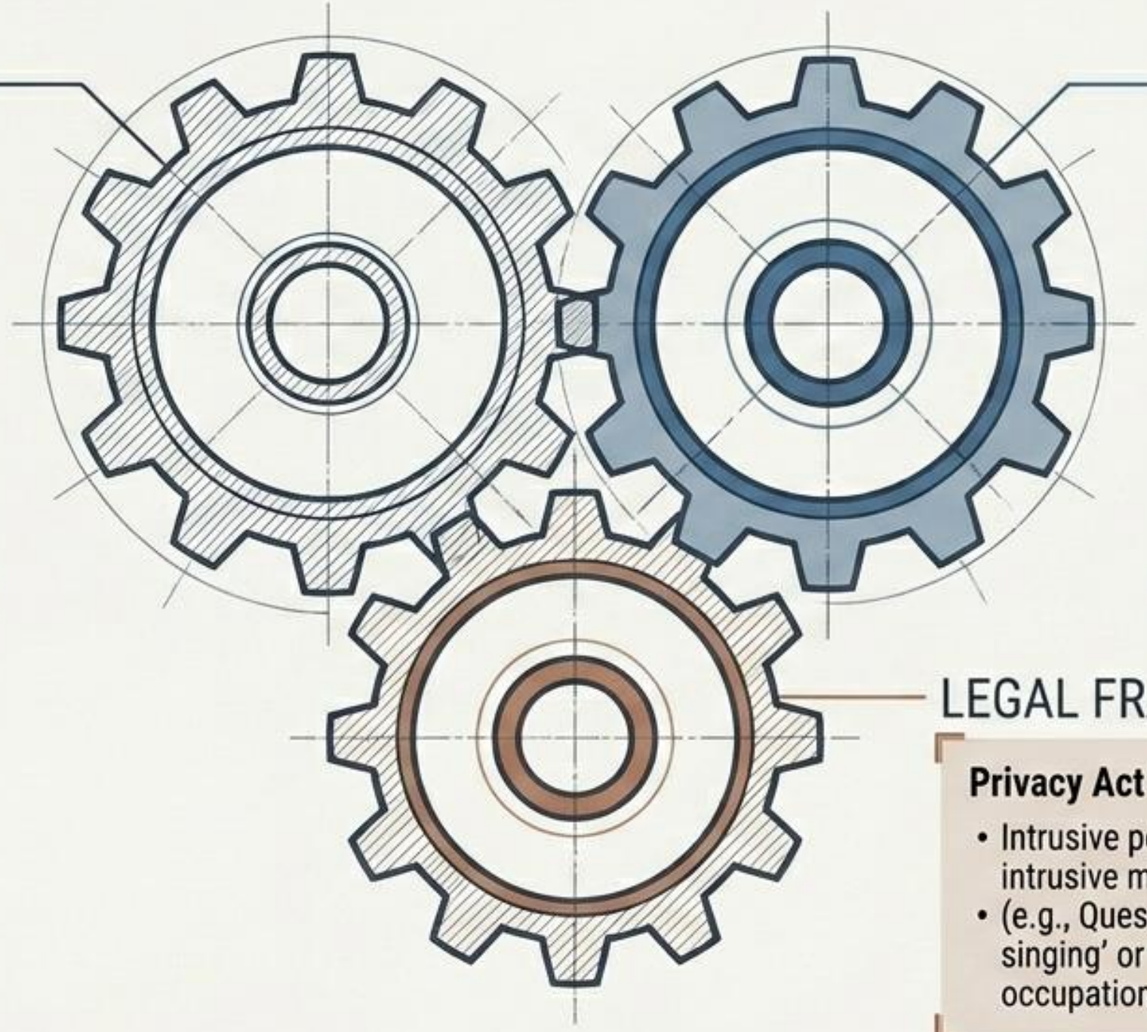
THE ASSESSMENT GATEWAY



ALIGNING CONTENT WITH CONTEXT

JOB DEMAND

TEST CONTENT



Human Rights Act 1993

- Selection decisions must be based on job-relevant criteria.
- Test complexity (e.g., vocabulary, speed) must perfectly mirror actual job requirements.

Face Validity

- Must look relevant to the candidate.
- General content should not require organisation-specific knowledge that disadvantages external applicants.

LEGAL FRAMEWORK

Privacy Act & Code of Ethics

- Intrusive personal questions are forbidden if a less intrusive method exists.
- (e.g., Questioning a candidate's 'hate for opera singing' or 'thoughts about sex' is legally invalid in occupational settings).

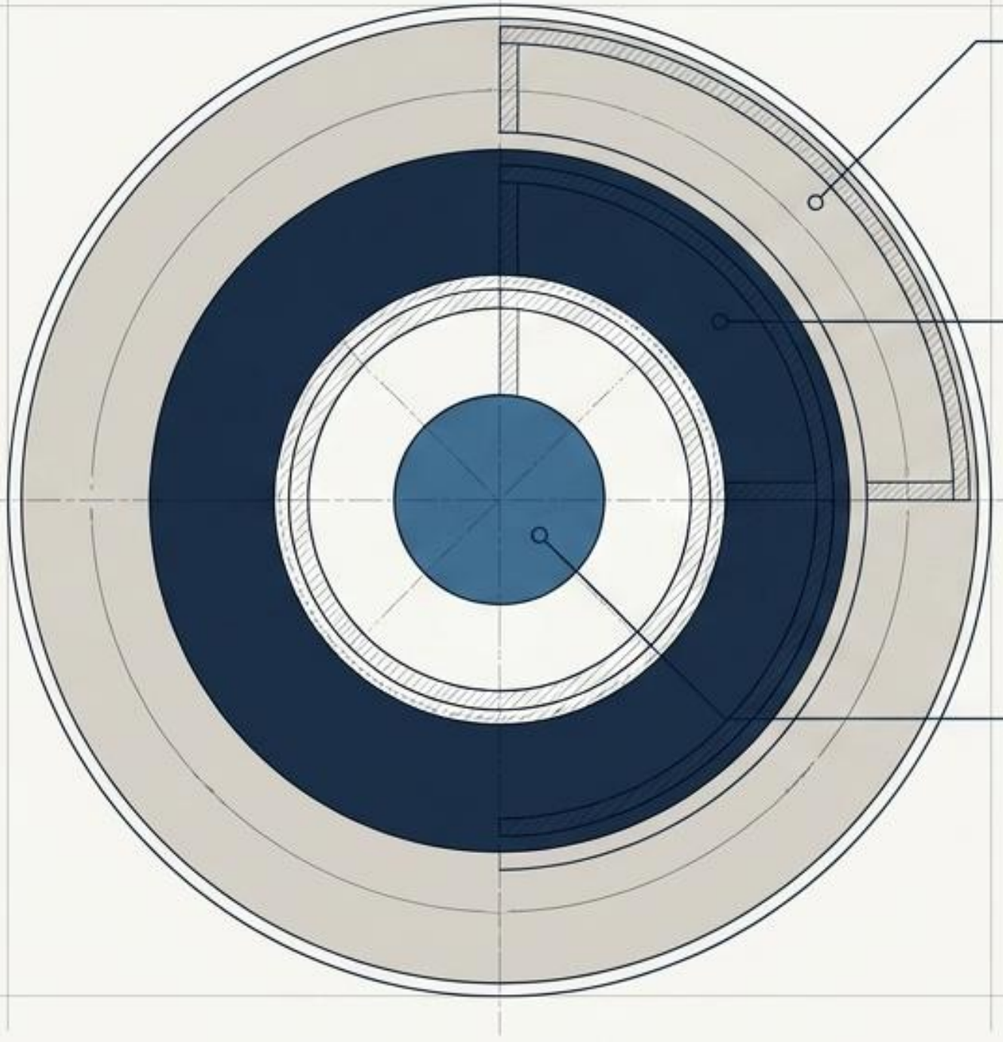
THE RELIABILITY THRESHOLD

The 0.75 Gold Standard.

Reliability sets the upper limit on validity. A 0.75 coefficient provides a manageable margin of error: 0.5 Standard Deviations (equivalent to 1 STEN or 5 T-Scores).

Classical Test Theory	Item Response Theory (IRT)
• **Internal**: Consistency across items (minimum standard).	Focuses on item-level accuracy (Information Function).
• **Test-Retest**: Consistency across time (captures administration/state errors).	Reliability varies by candidate ability (Theta).
• **Parallel**: Consistency across alternate forms.	Uses Marginal Reliability for a single comparative index.

THE THREE TIERS OF VALIDITY



Outer Ring - Face Validity

Qualitative assessment. Does it look relevant and non-discriminatory to the candidate?

Middle Ring - Construct Validity

Theoretical soundness. Must correlate ≥ 0.55-0.60 with tests measuring the same construct (using sample sizes > 100). Factor analysis must prove it measures the claimed number of constructs.

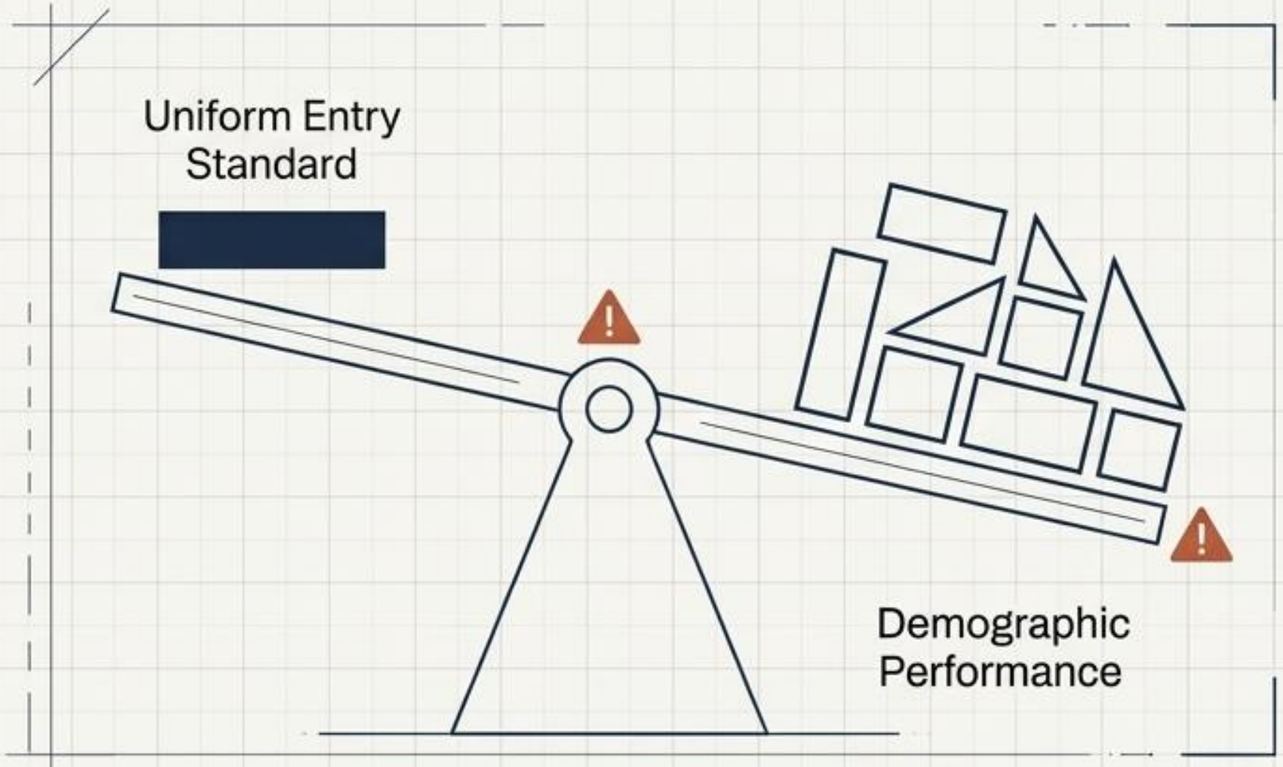
Bullseye – Criterion Validity

Essential for employment. Predicts future behaviour. The coefficient must be ≥ 0.3 (delivering an 80% success rate based on a 1:20 selection ratio). Requires a sample size > 100 (ideally two groups of 50).

Calibrating for Adverse Impact

The Diagnostic Triggers

Group differences exceeding 1 Standard Error of Measurement, or pass rates failing the 'Four-Fifths Rule' (proportion of minority group passing is less than 80% of the majority group).



Differential Prediction

If validation shows a group scores lower on a test but performs equally well on the job, disparate impact is occurring.

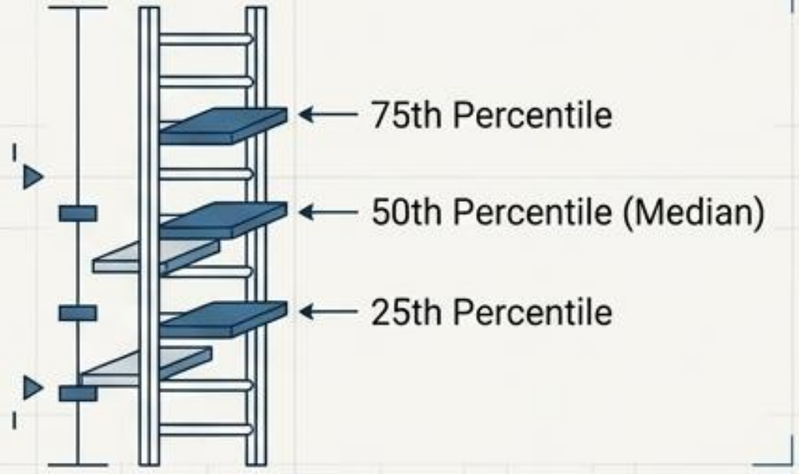
Mitigation Protocols

- Avoid taking only top-down scores; use job-relevant cut-offs.
- Test in the candidate's first language and remove cultural linguistic barriers (e.g., double negatives are absent in Pacific languages).
- Reduce "testing threat" via practice materials and non-threatening terminology.

Establishing the Baseline: Norming Systems

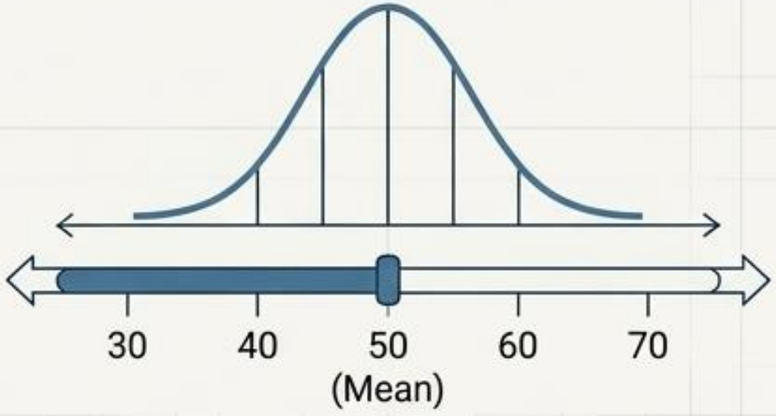
Core Rule: Base rates are only meaningful if the comparison group matches the target population. Norm groups must have N ≥ 150 (preferably 300+) for reliable tools. Use New Zealand-specific norms to ensure cultural validity.

Rank Order (Ordinal)



Format: Percentiles.
Benefit: Easy for candidates and managers to understand. Reflects true score distributions.
Limitation: Unequal units of measurement; cannot be mathematically averaged.

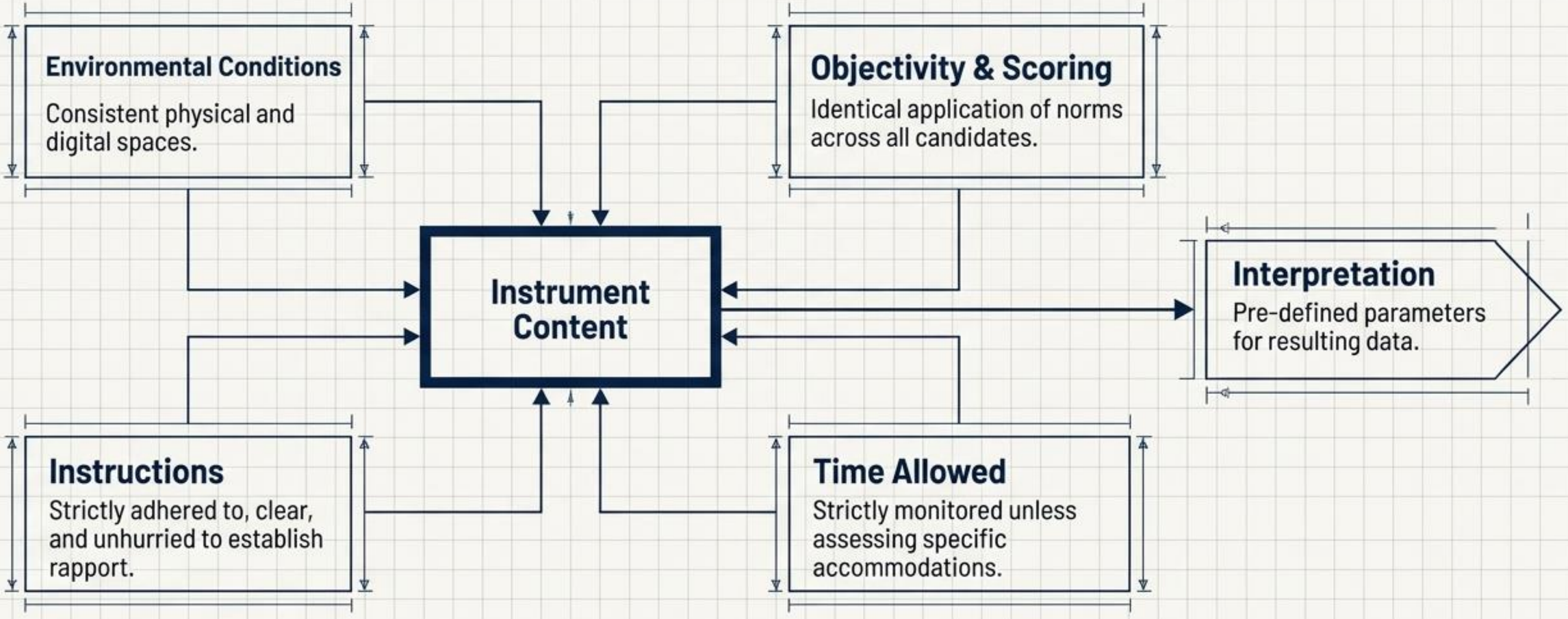
Standard Score (Interval)



Format: T-Scores, STENs, Stanine, IQ.
Benefit: Mathematically manipulable, equal units assuming normal distribution. Essential for complex data weighting and advanced interpretation.

The Engine of Standardisation

The Core Principle: Standardisation is the fundamental mechanism that generates utility. Deviations introduce measurement error and potential discrimination.

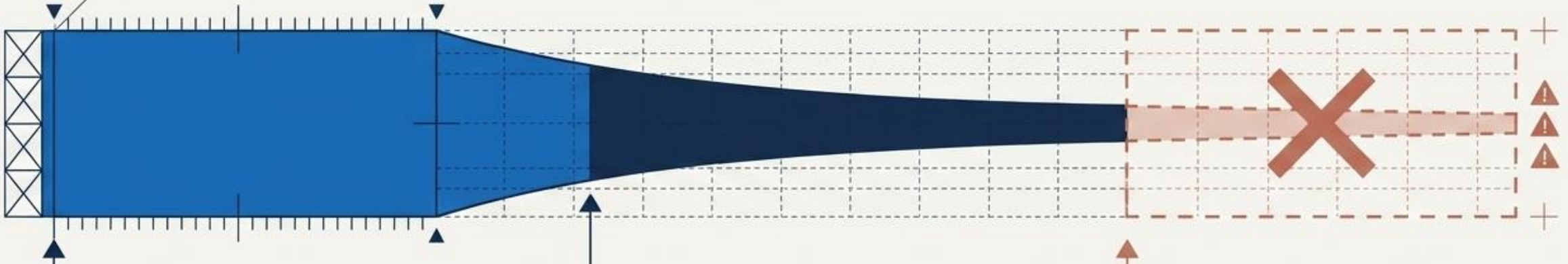


The Digital Frontier: Remote Administration

	Screening	High-Stakes
Unsupervised	Highly effective for volume screening, especially using Computer Adaptive Tests (CATs) that draw from large item banks.	
Supervised		Mandatory for later stages due to identity verification risks. Remote proctoring via webcam or follow-up verification tests under supervised conditions are required.

The Access Risk: Remote delivery risks adverse impact if digital infrastructure is assumed. In NZ, 1 in 5 lack home internet access, rising to 32% for Māori and 35% for Pacific Islanders. Alternative provisions are legally and ethically mandatory.

The Information Lifecycle



Day 0: Immediate (The Feedback Mandate)

The Privacy Act and Code of Ethics require meaningful feedback. Face-to-face is preferred. Uninterpreted raw data or profile charts must never be handed to untrained individuals.

6–12 Months: (Validity Window)

Individuals develop over time. Data remains robust for active selection decisions within this window. Secure, locked storage is paramount.

1–2 Years: (Legal & Practical Expiry)

Decisions can be challenged for up to 90 days (Employment Relations Act) or 1 year (Human Rights Act). Relying on data older than two years is scientifically inaccurate and legally indefensible.

The Psychometric Gold Standard Checklist

Reliability Index

≥ 0.75

Margin of error ≤ 0.5 SD.

Criterion Validity

≥ 0.3 correlation

Sample size N ≥ 100 or 2x50 groups.

Construct Alignment

≥ 0.6 correlation

Must align with proven benchmark tests.

Adverse Impact Limit

80% Pass Rate

Pass rates must clear the Four-Fifths Rule; variations must not exceed 1 Standard Error of Measurement.

Norm Group Viability

N ≥ 150

300+ preferred. Must use localised (New Zealand) context data.

Data Lifespan

Max 2 Years

Raw charts are never to be released uninterpreted.