Test Construction

Test items: These are the items that make up a test.

Quantitative Item Analysis
Three key aspects to item validity

Item Analysis: This is the general term for all techniques used to assess the characteristics of test items.

Item Analysis can be qualitative or quantitative.

Stages of Quantitative Item Analysis: Generate a pool. Generate administration and scoring instructions. Submit the pool to reviewers for qualitative item analysis. Try out items on pilot sample. Apply methods to results. Delete and modify items. Cross modify. Standardise and score. Collect normative data.

Items differ in the responses they require: **selected-responses or constructed-responses.**

Selected Responses: A limited set of alternatives from which a participant has to choose: multiple choice, true false, ranking, matching etc. They are easily administered and scored and are objectively scored. But allow for guessing, partial knowledge, offer no opportunity for individual skills to shine through, do not measure all constructs (e.g. creativity).

Constructed Responses: Open-ended answers or "fill in the blanks". They provide richer, perhaps more ecologically valid examples of behaviour. However, their validity can be lower, take longer to score, can be subject to experimenter bias.

Hayling Test (Burgess 1996 et al.) A neuropsychological test of executive function. Measures problems with initiations and inhibition. Patients with frontal lobe lesions have problems with one or both. Participants are required to say a relevant or irrelevant word at the end of a sentence.

Item Difficulty: How easy are they, are they equally difficult for all populations? To control for this, we can use **Absolute Scaling (Thurstone, 1925).** This turns sample scores in to z scores and they are plotted on graphs. We see if they are normally distributed (compare samples from different populations). Speed tests, power tests (no time limit) and combinations can cause tests to vary in their difficulty.

Item Discrimination: Do they discriminate between people that have desired attribute and those that don't? Index of Discrimination (e.g. calculate difference in average age of each group) or Correlation Coefficients (e.g. correlations between amnesics and non-amnesics).

Classical Test Theory

Classical Test Theory argues that to combine discrimination and difficulty you can perform an **item-test regression** which calculates the proportion of individuals at each total score level who passed a given item. Longer tests are therefore more reliable than shorter ones.

Item Response Theory (Reise et al. 2000) argues that longer tests should not be more reliable than short ones. Using computer analysis, psychologists can select and eliminate items that are too easy and short, increasing their **construct validity.** This produces overall a more valid test.

Item Fairness: Use panels to weed out stereotypical, offensive items or ethnocentric items (verbal and nonverbal tests).