PS416
Personality & Individual Differences

Intelligence A
Tools and Methods for Intelligence Testing

Maxwell J Roberts
Department of Psychology
University of Essex
www.tubemapcentral.com

version date: 11/11/2020

Plan for my Three Topics

1) Topic A: Tools and Methods for Intelligence Testing
    - Reliability, Validity, Standard Error of Measurement
    - Intelligence test formats and items
2) Topic B: The Structure of Intelligence
    - Conceptualisations of intelligence
    - Single Factor, Hierarchical, Multiple Ability and Contextual theories
3) Topic C: Uses of Intelligence, Critiques of Intelligence
    - Criticisms of the construct of intelligence
    - Using intelligence tests to predict people's futures
    - Criticisms of the validity of intelligence tests

Tools and Methods for Intelligence Testing

- Origins of intelligence testing
  - Why test intelligence
    How to predict intellectual future
- Basic requirements for psychometric tests
  - Reliability, Standard Error of Measurement, Validity
- Intelligence test formats
  - Individual versus group tests
- Intelligence test items
- Individual tests: examples
- Group tests: examples
- Testing versus theory

Origins of Intelligence Testing

- Why test intelligence?
  - Useful to be able to predict people's intellectual futures
    - Anticipating outcomes, e.g., education
    - Identifying interventions needed, e.g., education
    - Diagnosis and prognosis, e.g., clinical psychology
    - Selection, e.g., job candidates
    - Theorising: structure of abilities might affect the above

Origins of Intelligence Testing

- How might intellectual future be predicted?

1) Evaluate achievements of parents?
    - Successful parents -> successful children?
    - Might have use irrespective of stance on the effects of nature/nurture on differences in intelligence
    - *A crude measure, siblings can differ markedly*

Origins of Intelligence Testing

- How might intellectual future be predicted?

2) Evaluate progress after training has been given
    - Successful completion of training predicts future success?
    - Approach taken by Binet in the 1900s
        - Some children absorb general/cultural knowledge faster than others
    - Approach taken by British Universities
        - A Level results diagnostic of likelihood to succeed at continuing education?
    - *In many cases time invested might be wasted*

Origins of Intelligence Testing

- How might intellectual future be predicted?

3) Identify prerequisites necessary for future success
    - Certain specific skills are foundations for success?
    - Approach taken by driving tests
        - Training time irrelevant,
          can key skills be demonstrated?
    - *Prerequisites can be hard to identify*

Origins of Intelligence Testing

- How might intellectual future be predicted?

4) Evaluate general cognitive abilities?
    - General ability is the foundation for all future achievement?
    - Tasks must be abstract, requiring basic thinking ability without drawing on any *domain specific* knowledge/skill
    - Approach taken in many occupational settings, a general intelligence test is part of the selection process

Origins of Intelligence Testing

- How might intellectual future be predicted?

4) Evaluate general cognitive abilities?

- *What are the most appropriate foundations?*
  - What are the fundamentals of cognitive ability
    - Galton: perception?
    - Raven: logical reasoning?
  - What tasks best measure their efficacy?
- *Prediction errors inevitable whatever the system*
  - Disconnection between test content and real life leads to allegations of systematic error, i.e. unfair discrimination

Basic Requirements for Psychometric Tests

- Content question for intelligence tests has been settled
  - Test dominated by abstract logical reasoning tasks/puzzles
  - Intelligence = general logical reasoning skill
- Intelligence tests are used to affect people's futures
- Like all such tests, basic criteria must be satisfied
  - High reliability
  - Low Standard Error of Measurement
  - Validity

Reliability

- A **reliable** test measures a construct *accurately*
  - Accuracy is identified in terms of ***consistency***
  - Accurate measures give similar scores on repeated occasions
    - Ruler example, an unreliable ruler is useless!
- All types of reliability similar, measured by *correlation coefficient* between at least two scores
- Reliability is technically easy to compute and determine, other than a very few pitfalls

Reliability

- **Test-Retest reliability**
  - Accurate scores will be stable over time
  - Test same people twice on two separate occasions
  - Allow reasonable gap, e.g. three months
  - Look at Test1-Test2 correlation between paired scores
  - $r = .7 \sim .8$ is a good range to aim for
- Difficult for children, mood questionnaires, etc.

Reliability

- **Inter-Scorer reliability**
  - Accurate scores will be judged consistently
  - For tasks without clear objective marking scheme
  - Same responses scored by two independent judges
  - Look for high correlation between paired rating scores
- Needs caution, scorers might be acquaintances with similar expectations/prejudices

Reliability

- **Parallel-Form reliability**
  - Accurate scores will be consistent amongst different versions of the same test
  - Test same people twice with two different test versions
  - Look for high correlation between paired test scores
- Very useful but rare, each individual test requires considerable effort to create

Standard Error of Measurement

- Test scores are *dependent measures*
  - Test scores are *estimates* of true scores
  - Test score may be a *good* or *bad* estimate of a true score
- Intelligence test score is a *dependent measure* of intelligence
  - Intelligence test score may be a *good* or *bad* estimate of true intelligence level
- Standard Error of Measurement measures how likely it is that the test scores are good estimates

Standard Error of Measurement

- Measurement errors are normally distributed
  - Smaller errors common, larger errors rare

- Standard Error of Measurement calculated using reliability
  - Consistent test scores more likely to be accurate (close to true scores) than inconsistent scores
  - Used to give a 95% confidence interval, how likely it is that a test score is close to the true score

Standard Error of Measurement

- Test with high SEM
  - Individual scores are not necessarily close to true scores
  - Scores discriminate poorly between candidates

- Test with low SEM
  - Individual scores more likely to be close to true scores
  - Scores discriminate well between candidates

Standard Error of Measurement

- The more measurements are taken, the more likely that errors of measurement will cancel each other out
  - Important practical implications
    - Aggregate scores from many tests more likely to be accurate than a single score from just one test
    - Tests with many tasks are more likely to be accurate than tests with fewer tasks
    - Longer tests with many items are more likely to be accurate than shorter tests with fewer items

Validity

- A **valid** test is a *genuine* measure of the construct claimed to be measured
- Unreliable tests cannot be valid by definition
- Reliable tests *might be valid*
- Validity is difficult to determine conclusively, sometimes impossibly so, with many pitfalls for the various methods

Validity

- **Predictive validity**
  - Intelligence tests should predict success for tasks that require intelligence
  - Test score and task performance should be correlated
  - Even low correlations are useful (Topic C)
- Difficult to measure success in some contexts, are these measures valid and reliable?
- No conclusive guarantee of validity, multiple factors may determine success at complex tasks

Validity

- **Differential validity**
  - Intelligence tests should ***NOT*** predict success for tasks that *do* ***NOT*** require intelligence
- This validity is useful to demonstrate, but tasks may have differential correlations for many trivial reasons
  - E.g. measures of job success are valid/reliable for a maths teacher, and invalid/unreliable for a PE teacher

Validity

- **Concurrent validity**
  - Intelligence tests should be highly correlated with established benchmark measures of intelligence
- How is it known that the benchmarks are valid?
- Conformity/internal consistency within a discipline is not evidence that the discipline is non-dysfunctional
- Can hinder development of new, improved measures

Validity

- **Construct validity**
  - An intelligence test is valid if it demonstrates multiple converging validities, for example, as described above
- Unfortunately, adding together many inconclusive findings does not result in conclusive findings overall

Intelligence Test Formats

- Generally comprise abstract logical reasoning tasks
  [Abstract: not linked to a particular specific context]
- Intelligence = something to do with
  the ability to solve these types of task

- Although all tests include logical reasoning tasks there are still important differences between them:
  - What tasks should be used, does their content matter?
    - Shapes, patterns, words, numbers, etc.
  - What types of intelligence/ability can be measured?
    - General, verbal, non-verbal, numerical, spatial etc.

Intelligence Test Formats

- Tests therefore differ in:
  - The types of tasks
  - The number of different tasks
  - The extent to which they attempt to be culture-neutral
    - The extent that general/practical knowledge is included as part of the test
    - The extent that items draw upon knowledge of word meanings, or generally require language understanding
  - Whether administered individually or as a group

Intelligence Test Formats

- **Individual versus Group Tests**
  - Individual tests
    - Taken face to face, with a trained administrator assessing one single individual
    - Essential where supervision needed
    - ✓ Use a wide variety of tasks [reduces SEM]
    - ✓ Motivation can be maintained (e.g., targeted difficulty)
    - ✓ Performance can be interpreted (e.g., near misses)
    - ✗ Administrator must have appropriate training
    - ✗ Time-consuming to administer
  - Best for diagnosis, often used in educational settings

Intelligence Test Formats

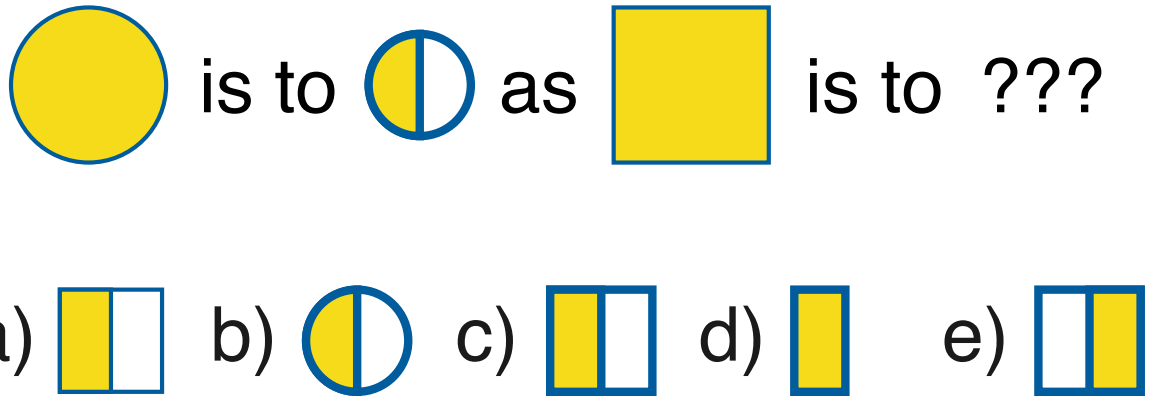
- **Individual versus Group Tests**
  - Group tests
    - Several people take test simultaneously under exam conditions supervised by a (less-trained) individual
    - ✓ Relatively fast and efficient to administer
    - ✓ Advanced training not necessary for administrator
    - ✓ Objective scoring, usually multiple choice
    - ✗ Restricted range of tasks/items
    - ✗ Restricted answer formats, restricted interpretation
  - Best for screening, often used in occupational settings

Intelligence Test Items

- Many different formats of logical reasoning puzzles
- All of the following examples have been used in intelligence tests
- All tests are *age matched*: item difficulty pitched at a level appropriate for people being tested

Intelligence Test Items

- Analogy (non-verbal)



Intelligence Test Items

- Analogy (verbal)

???	Tree
Egg	Chicken

a) Coop b) Nest c) Chick

d) Seed e) Sapling

Intelligence Test Items

- Sequence (verbal/general knowledge)

Which letters are next?

J, F, M, A, M, J, ???, ???

a) F, M b) F, A c) J, M

d) J, A e) J, F

Intelligence Test Items

- Sequence (numerical)

Which numbers are next?

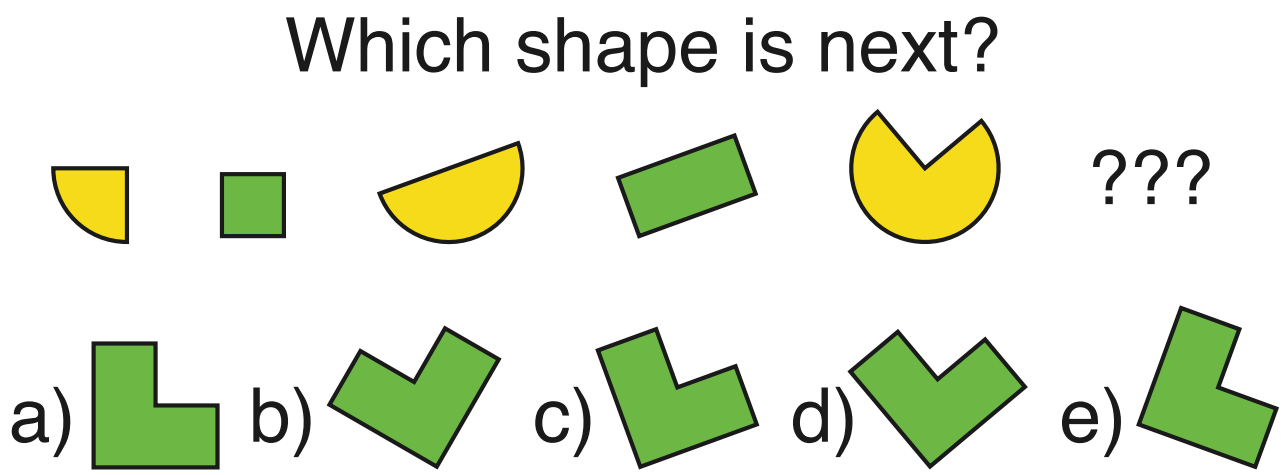
2, 3, 4, 6, 6, 9, ???, ???

a) 8, 10 b) 12, 16 c) 8, 12

d) 12, 12 e) 8, 16

Intelligence Test Items

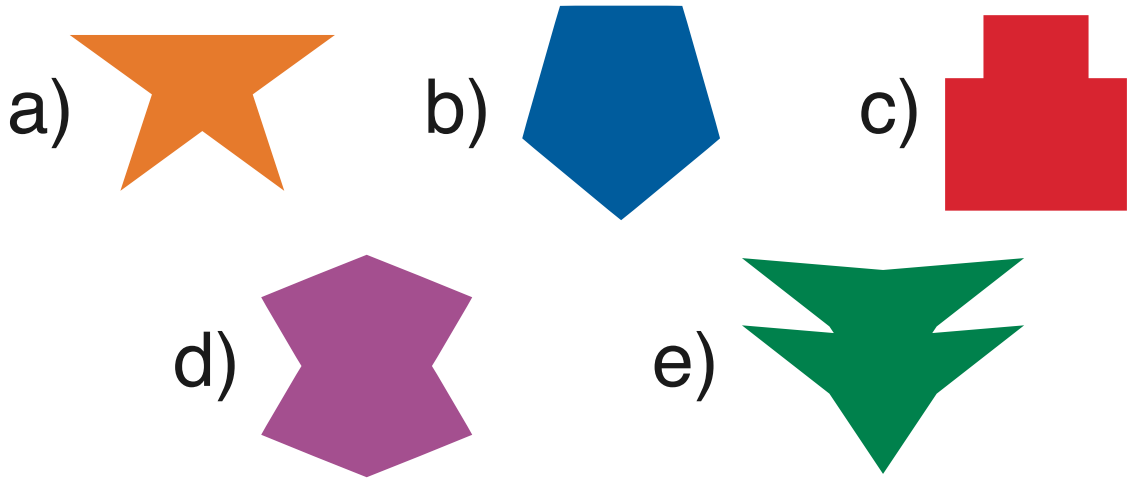
- Sequence (non-verbal)



Intelligence Test Items

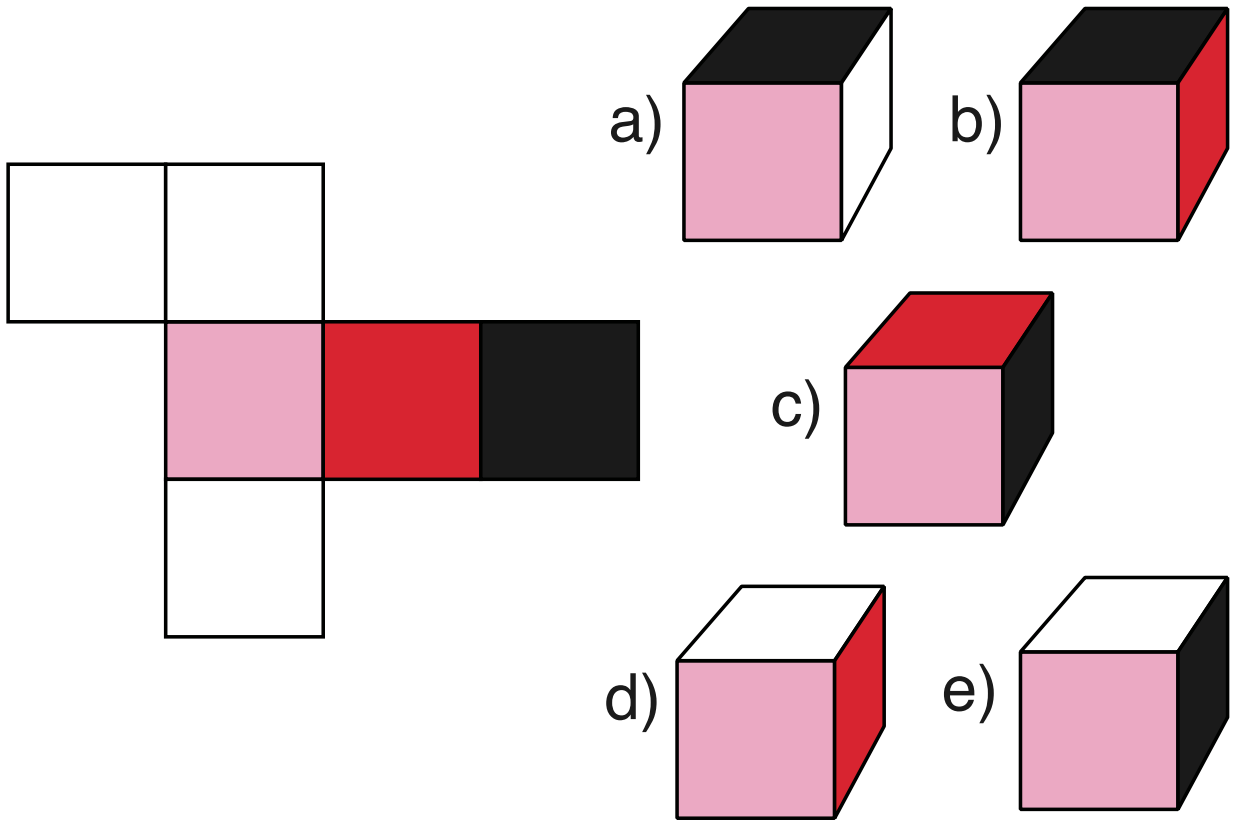
- Odd one out (non-verbal)

Which shape is the odd one out?



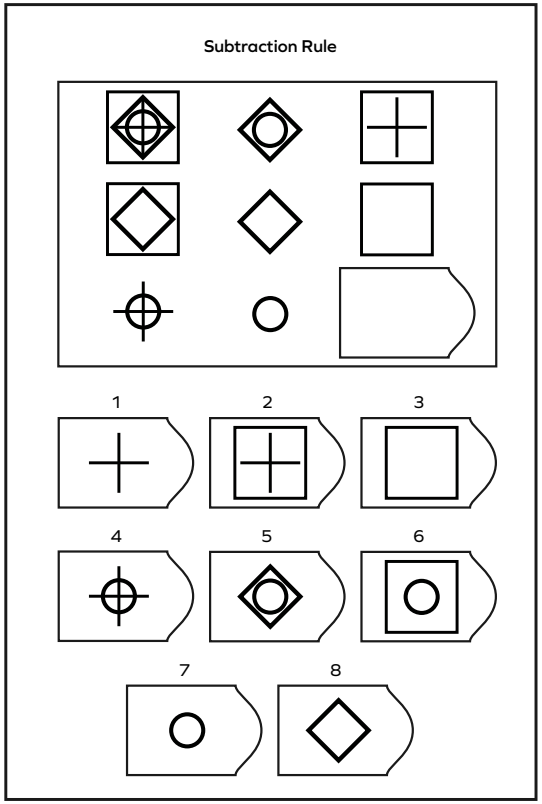
Intelligence Test Items

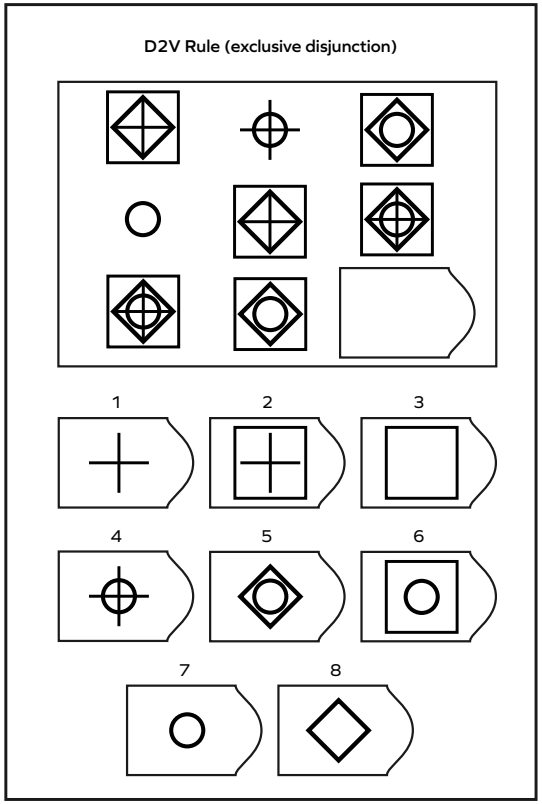
- Spatial manipulation (non-verbal)



Intelligence Test Items

- Many different logical relationships can be encoded in a non-verbal matrix
- Often more than one logical relationship in the same item





Individual Tests: Examples

- **Stanford-Binet** test (1904 onwards)
  - First useful intelligence test
  - Most recent version 2003
  - Can calculate scores for general intelligence or subscales:

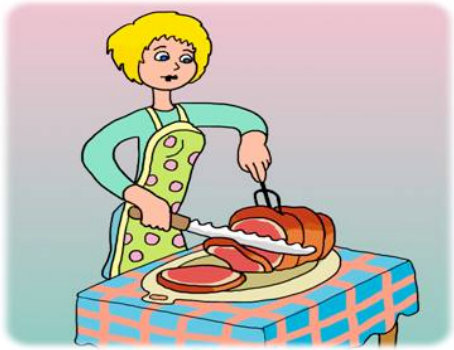
    Fluid reasoning (see Topic B), Knowledge, Quantitative reasoning, Visual-Spatial reasoning, Working Memory

- Of all the tests, this one has the greatest weight put on verbal and general knowledge

Individual Tests: Examples

- Some recent Stanford-Binet test practice items

Look at the picture below. There is something very silly or impossible about this picture. What is wrong with this picture?



How many petals are on this flower?

1) What happens to water when it freezes?

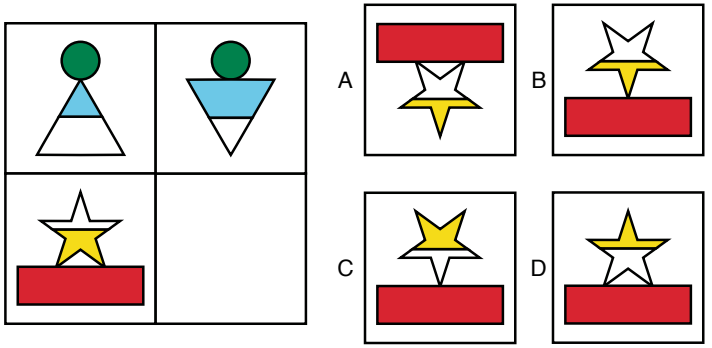
2) Can you name three vegetables that are green?

3) Can you name something you must turn on to use?

4) How is sand different from soil?

5) Which invention is older, the wheel or the computer?

Do you see the shirt in the first box? Point to a shirt in the row on the side that is identical.



Do you see these four boxes? In the top row the pictures go together in a certain way. Now look at the bottom row. Do you see the empty box? Which of the four pictures on the side goes with the picture in the bottom box the same way as the two pictures in the top row go together?

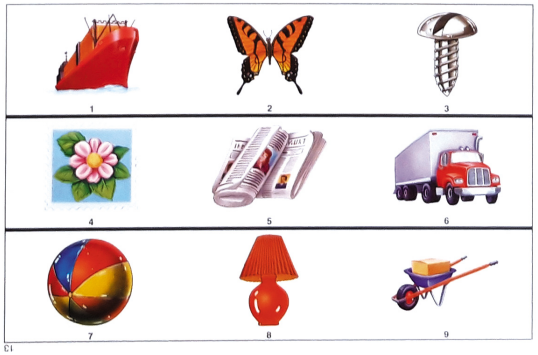
Individual Tests: Examples

- **Wescher Scales** (1930s onwards)
  - WAIS for adults, WISC for children
  - Most recent version 2008
  - Variety of items: logical, verbal, numerical, spatial and also memory, general knowledge, etc.
  - Balance is more towards basic cognition and less towards general knowledge
  - Can calculate scores for general intelligence or subscales (see slide 41)
  - WISC tasks include *vocabulary*, *arithmetic*, *block design, picture completion, picture concepts,* and *matrix reasoning*

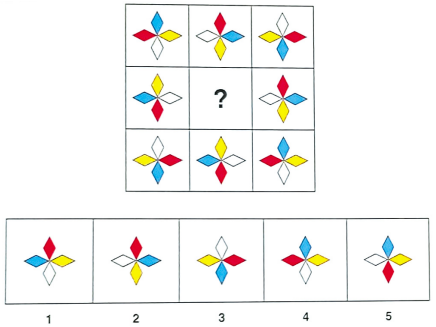
Individual Tests: Examples

- Examples of WISC items

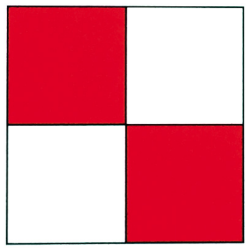
Picture concepts

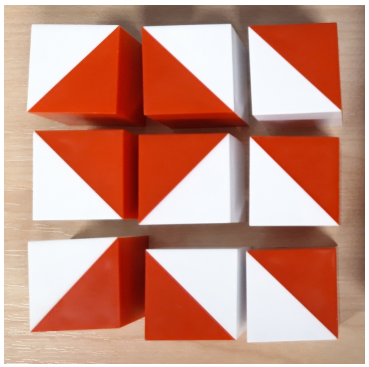


Matrix reasoning



Block design

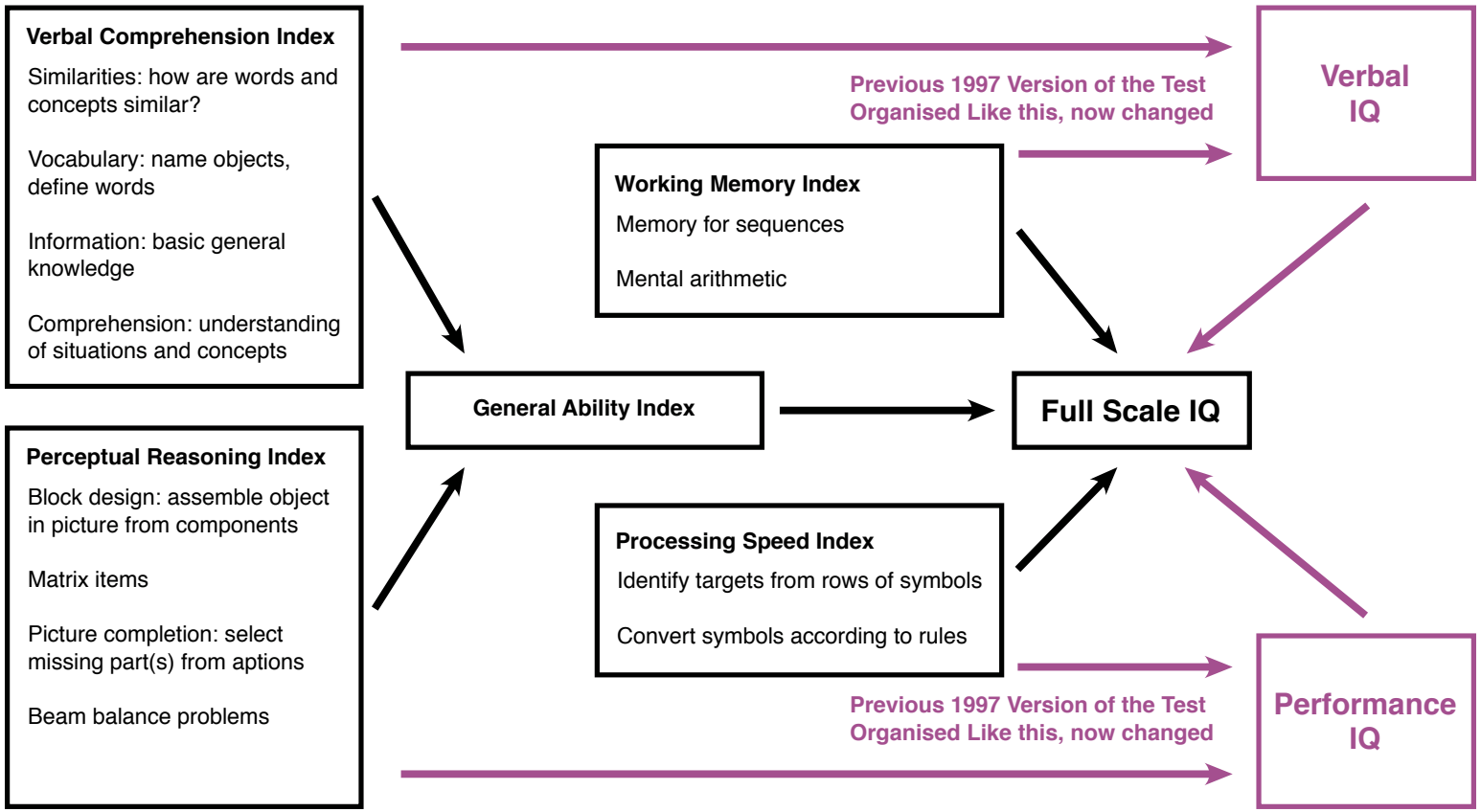




Picture completion

Individual Tests: Examples

- Subscales from the WAIS, and changes over time



Group Tests: Examples

- Items focus more towards logical reasoning (as opposed to memory and general knowledge)
  1) Identify a rule or relationship from item elements
  2) Apply the rule to generate the correct answer
- Items focus more towards relationships between shapes and patterns (as opposed to words/numbers)
- Often only one task
- Almost always multiple choice
- High score = finding the one single correct answer for each item (limited creativity)

Group Tests: Examples

- **Raven Progressive Matrices** (1940s onwards)
  - Completely non-verbal, matrix format (slide 36)
  - Progressive: items steadily increase in difficulty from very easy to very difficult
  - Draws people in, easy to understand the requirements minimal instructions necessary
  - Provides one single general intelligence score; specifically designed to measure ***g*** (see Topic B)
  - Statistical testing indicates Ravens matrices provideee one of the purest measures of general intelligence
- Has become the gold-standard of intelligence tests
- Very widely used

Group Tests: Examples

- **Culture Fair Tests** (Cattell, 1940s onwards)
  - Non-verbal, four different tasks including small matrices (slide 36) and sequences (slide 33)
  - Four different tasks less convenient to administer, continual stop-go
  - Provides one single general intelligence score, intended to measure ***fluid intelligence*** (see Topic B)
- Despite additional tasks, this test does not offer enough improvement in measurement to justify its inconvenience

Group Tests: Examples

- **AH tests** (Heim, 1960s onwards)
  - Two separate forms, one has tasks based on shapes and patterns, the other numbers and words
  - Includes just about every possible item type
  - Like a self-administered Stanford-Binet test
  - Each item needs its own instructions, less convenient
  - Provides general intelligence score, or separate scores:
    verbal: word meanings and numerical reasoning
    non-verbal: abstract logic and transforming shapes
- Fallen out of use, inconvenient format
- Separate Verbal/Non-verbal scores are probably not statistically justified (see Topic B)

Testing versus Theory

- All the tests outlined have good reliability
- Validity is a more complicated issue
- Test development preceded intelligence theory (Topic B)
  - Intelligence testing commenced *before* the creation of necessary statistical techniques and also computers
  - Formats originated from the hypotheses of researchers, rather than empirical research
  - Tests are good for measuring general intelligence, but validity of subscales/more specific abilities less clear
- Disagreements about the existence/measurability of general intelligence (Topic C)

Copyright Notice

- The text of and organisation of this presentation is copyright ©Maxwell J Roberts, 2020. These slides may be distributed in unaltered form, but must not be reused or reformatted for presentations or teaching, whatever the purpose, and they must not be rehosted for downloading at any other web site, without the express permission of the copyright holder.
- The following images are copyright ©Maxwell J Roberts and may not be reused for any purpose except for fair-use educational/illustrative purposes. In particular, they may not be used for any commercial purpose (e.g., textbook, non-open-access academic journal) without the express permission of the copyright holder.
- Slides 29-36, 41
- All subjective evaluations expressed are the personal opinions of the author
- *This slide must not be deleted from this presentation*