Saville Assessment

Willis Towers Watson

swift

Flexible Online Learning: Swift Occupational Ability

training

Introduction

This guide covers the content included in the Saville Assessment Swift Occupational Ability Flexible Online Learning course and can be used as reference material during and after the training.

Contents

Key Figures 3

Theorists in Intelligence Testing 3

Singular vs. Multi-faceted Intelligence 4

Cognitive Ability Tests 5

Introduction to Testing 7

Talent Trends and Challenges 7

Projective Tests 8

What is a Psychometric? 8

Can-do and Will-do Assessments 9

Types of Assessment 10

Externally-referenced Tests 10

Norm-referenced Tests 10

Testing Theories 10

Test Security 11

Gradient Step Testing (Fixed Length) 12

Job Analysis and Assessment Choice 13

Common Methods of Job Analysis 14

Considerations for Choosing Assessments 15

Screen Out, Select In 16

Test Feedback 17

Why do we give Test Feedback? 17

Feedback of Saville Aptitude Tests 17

Features of Different Feedback Forms 18

Feedback Process 19

Feedback Tips 20

Test Scores and Test Norms 21

Frequency Distribution 21

The Normal Distribution 21

The Normal Distribution Performance Bandings 22

Rank-ordered Norm Scores 22

Describing Percentiles 23

Standardized Scores 24

Calculating Standard Deviation 26

SDs and the Normal Distribution 27

Other Ways of Describing Scores 27

Norms 29

Choosing Norms 29

Norm Sampling Methods 30

Correlations and Reliability 37

Interpreting Correlations. What is Correlation? 37

Strength and Direction of Correlations 31

Using Correlations 33

Reliability 34

Types of Reliability 35

Margin of Error 35

Validity and Test Utility 37

What is Test Validity? 37

Types of Validity 37

Meta-Analysis and Validity Generalisation 39

Test Utility 40

Best Practice and Ethics 42

Proper Data Management – GDPR 42

Discrimination 43

Fair Assessment 43

Using Tests Responsibly 45

Key Figures

Theorists in Intelligence Testing

Charles Spearman

- Observed that those who perform well in one ability tend to do better on others
- Proposed a higher factor of general intelligence - which he coined 'g'

Louis L. Thurstone

Multi-faceted view that intelligence comprises Seven Primary Mental Abilities

- Verbal Comprehension
- Word Fluency
- Number Facility
- Spatial Visualization
- Associative Memory
- Perceptual Speed
- Reasoning

People can have high mental ability in one area, while being lower in others.

Philip Vernon

Hierarchical structure of intelligence

- Broke down 'g' into Academic and Practical ability
- Academic factor refers to abilities including reading comprehension and arithmetic reasoning
- Practical was more focused on mechanical and spatial abilities

Raymond Cattell

Fluid and Crystallized Intelligence

- Fluid intelligence - the ability to deal with novel and abstract problems. It was thought to be *genetic and therefore immune* to culture and environment; for example, it could not be taught and thought to decline with age
- Crystallized intelligence - grounded in knowledge, expertise and wisdom *learnt over time, and thus thought to increase with age*

Howard Gardner

Theory of Multiple Intelligences

Some intelligences are best measured using specific skill assessments

- Linguistic
- Logical
- Spatial
- Musical
- Kinesthetic
- Interpersonal
- Intrapersonal

Singular vs. Multi-faceted Intelligence

To this day, it is still debated whether intelligence is a single construct or formed of multiple constructs.

Singular

Some academics favor a singular form of intelligence. As discussed, this would mean that someone who is good at one subject is likely to be good at another.

An example of a test that could be used to assess general, or singular, intelligence would be an IQ Test; the Wechsler Adult Intelligence Scale, for example. This test measures intelligence over four broad areas: Verbal Comprehension, Perceptual Reasoning, Working Memory and Perceptual Speed, and combines these to give a Full Scale IQ score. The fourth edition of this test, which was developed in 2008, is often used by neuropsychologists to assess overall cognitive function of the brain.

Multi-faceted

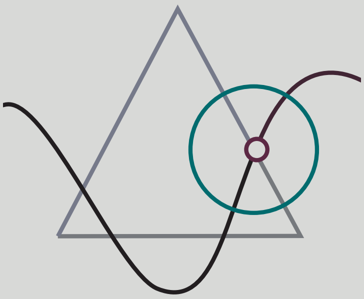
Some applied psychologists and practitioners lean towards a multi-faceted view of intelligence. An example of tests used to measure specific ability areas would be aptitude tests; a verbal reasoning assessment or an error checking test, for example. Practitioners tend to have a concern for the relevance of a test to the job, rather than being concerned with theoretical considerations.

Whilst evidence shows that tests of general intelligence predict job performance just as well as tests of specific abilities, in an occupational setting, it makes practical sense that you are only measuring abilities which are found to be relevant to the role. It is important to understand the level of general ability and the kind of specific abilities that are required for a role; assessing the related attributes of an individual enables sound judgments about job-fit to be made, and measuring specific abilities that have clear job relevance results in greater candidate, manager and legal acceptability.

On this basis, Vernon's hierarchical model integrating 'g' with more specific abilities - under the umbrellas of practical and academic intelligence - carries great merit.

Cognitive Ability Tests

- Cognitive ability/aptitude tests are generally found to be the strongest single predictors of work performance
- Applications: workplace selection, development and career guidance
- Predicts academic performance and broad life outcomes such as occupational attainment and stable employment, and even mortality, in addition to job performance



Saville Assessment Aptitude Tests

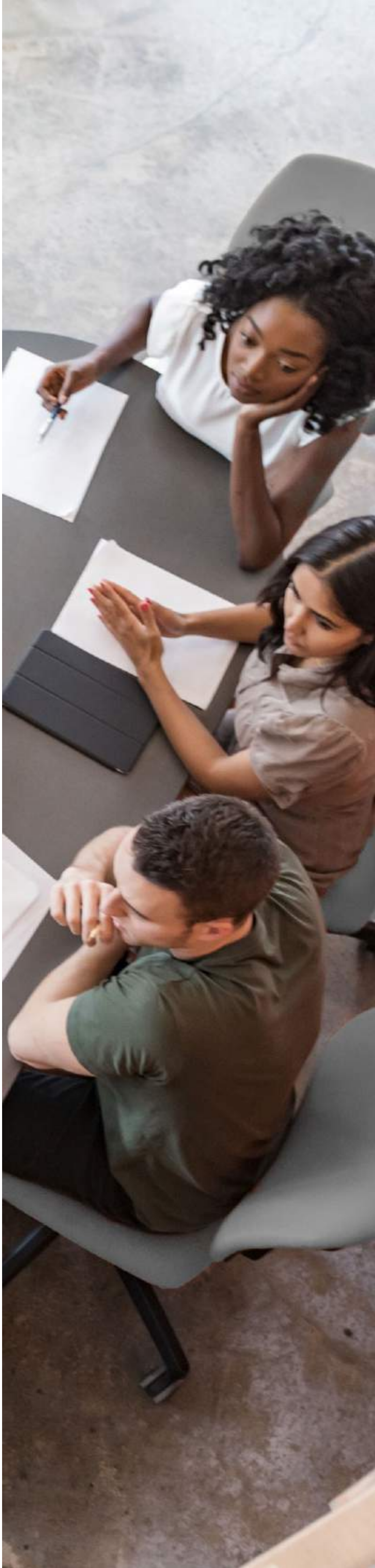
- Saville Assessment offers both measures of single aptitude and combined aptitudes (Swift tests)
- Single aptitude tests allow you to target an in-depth cognitive aptitude that is central to success in a role
- Combined tests such as Swift target the relevant areas more broadly and efficiently
- Single tests give a score for the single aptitude area which can be used for decision making. The decision-making score from a combined Swift test is the overall score

Introduction to Testing

Talent Trends and Challenges

These are some common trends and challenges in selecting Talent. Think about ones that have affected your organization and consider how aptitude assessments could be used to help.

- Talent is global
- Applicant numbers per role are increasing
- Organizations / behaviors, cultural fit, values
- Candidate experience is critical
- Diversity and inclusion, fairness in assessment
- Recruitment processes are speeding up
- Online, mobile and remote assessment is the norm
- Security of assessment materials is still a risk
- Everyone's talking about Big Data



Projective Tests – Thematic Apperception Test

Projective tests such as the Thematic Apperception Test and the Inkblot give candidates stimuli (*see figures 1 and 2*) that are open to interpretation. Candidates provide their interpretation of the picture, which is evaluated by the interviewer/assessor; it is claimed that this method uncovers individuals' unconscious needs or drives.

These kinds of assessments may be used in more clinical or therapeutic settings, however, they can be very subjective and are not suitable for selection. Subjectivity means they lack reliability and validity which are key attributes of psychometric assessments.

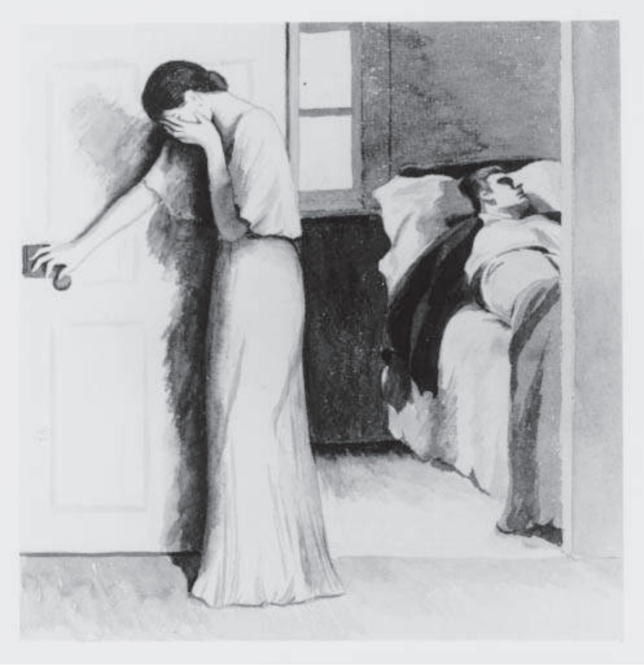


Figure 1 Thematic Apperception Test

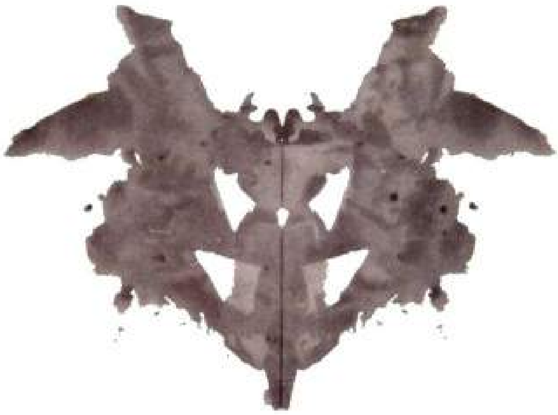


Figure 2 Rorschach/ Inkblot Test

What is a Psychometric Test?

Rather than focusing on tests that are subjective to candidates and assessors, we want to focus on psychometric assessments.

These can be defined as an assessment of a clearly defined psychological attribute, typically scored using a numerical scale or category system, to describe individual differences.

Psychometric Breakdown

Psycho – The mind

Metric – Measurement

We take Psychometric to mean the measure of the mind.

Can-do and Will-do Assessments

Will-do Tests

These measure typical performance, examples of which are listed below:

- Interest inventories/questionnaires
- Personality questionnaires
- Motivation questionnaires
- Job performance
- Attitude surveys
- 360 degree assessments

Interest inventories/questionnaires measure the things an individual is interested in. This type of information may be useful in career guidance. Personality questionnaires look at styles of behavior, for example the Occupational Personality Questionnaire (Saville et al, 1984) and the Professional Styles and Focus Styles versions of Saville Assessment's Wave.

Motivation questionnaires measure what people want to do. Note: this can also be measured by the Wave questionnaire detailed above. Rating scales look at measures of job performance.

Attitude surveys are often of great interest in market research. 360 degree assessments ask for ratings from bosses, colleagues and subordinates. Saville Assessment has developed the Wave Performance 360 questionnaire to gather self and other ratings online.

Can-do Tests

These assess maximum candidate performance, examples of which are listed below:

- Aptitude
- Achievement/attainment
- Intelligence tests (IQ)
- In-tray
- Work sample
- Trainability tests

Aptitude tests measure abilities that underpin future potential – examples include Saville Assessment's verbal, numerical and diagrammatic analysis tests.

Achievement/attainment tests look at an individual's level of current knowledge – examples include school exams or a driving theory test. Intelligence tests (IQ) are a mixture of aptitude and attainment, one common measure of IQ is the Wechsler Adult Intelligence Scale. In-tray exercises/business simulation exercises are tests which assess skills at particular tasks and are often very useful in assessment centres. Work sample tests present applicants for a job with a sample of the work they will be expected to undertake in the job. Trainability tests assess how well individuals respond to training.

Key Benefits: Aptitude Tests

This **Swift Occupational Ability** course focuses on the Can-do – Maximum Performance assessments. These are some of the key benefits of using ability tests and we will expand on them throughout the course:

- Benchmarks against external group
- Single most valid predictor of work performance
- Measures lots of different types of ability
- Efficient online assessment
- Fair and consistent treatment of candidates
- Supplements other sources of information
- Sophisticated question banking to protect the security of the content

Types of Assessment

Maximum performance “Can-do” tests can be split further into Externally-referenced and Norm-referenced tests.

Externally-referenced Tests

Externally-referenced tests ask “how well does the candidate compare against an expected standard?” Candidates have to reach a required standard on the test, irrespective of how anyone else performs.

A good example of this is a driving test: to pass, one must meet a set required standard on the test to be given a driving licence. This type of testing is also common in education – students generally are expected to reach a set standard to achieve a certain qualification or grade.

- A set standard needs to be achieved to pass these kinds of assessment
- Examples of externally referenced tests are driving tests and exams in education

Norm-referenced Tests

Aptitude tests are one type of ability test which are typically norm-referenced and indicate an individual’s level of aptitude compared to other people.

Aptitude tests that use norm referencing focus on predicting the future performance of individuals and typically do not require the test-taker to have specific knowledge or experience to do well. This type of referencing is popular in recruitment where selection is competitive and must only involve role-relevant tasks.

- Norm-referenced aptitude tests aim to predict future performance
- They do not require individuals to have specific prior knowledge
- Norm-referencing is popular in recruitment as you can benchmark individuals

Testing Theories

Questions, or items, in a test can either be fixed-content or item-banked. Fixed content means that all candidates are shown the same questions; item-banked means that questions are drawn from a large bank of content and candidates are therefore unlikely to see exactly the same content as each other. **Classical Test Theory** and **Item Response Theory** use these different types of content.

Classical Test Theory

- Classical Test theory (CTT) is more generally used for fixed-content testing rather than item-banked testing
- In CTT the number of questions answered correctly = ability. However, this assumes all items are of equal difficulty

Item Response Theory

- Item Response Theory (IRT) scores candidates in a consistent way, despite them seeing different test content
- IRT takes account of question discrimination (the information provided by the test item), difficulty and pseudo-guessing within a test drawn from a bank of content to estimate a candidate’s ability
- IRT provides a value called the Test Information Function which describes how well a test estimates a candidate’s ability. A single question is likely to have a lower Test Information Function than a whole test so generally a longer test is a better indicator of candidate ability than a shorter test

Test Security

With all testing there are potential issues relating to the security of test content. This is particularly true for remote online testing. One method for reducing security issues is to present different content to different candidates by holding content in a bank. Item-banked tests can use either Adaptive Testing or Gradient Step Testing.

Adaptive Testing (Uncontrolled Length)

With Adaptive Testing, items are randomly drawn from the bank as the test narrows in on the ability level of the candidate. The number of items in a test will be different for different candidates.

Candidate 1		Candidate 2	
Easy Level 1	○ ✓	Easy Level 1	○ ✓
Easy Level 2	○ ✓	Easy Level 2	○ ✗
Medium Level 1	○ ✓	Easy Level 1	○ ✓
Medium Level 2	○ ✓	Easy Level 2	○ ✗
Difficult Level 1	○ ✗	Difficult Level 1	
Medium Level 2	○ ✗	Medium Level 2	
Medium Level 1	○ ✓	Medium Level 1	
Medium Level 2	○ ✓	Medium Level 2	
Difficult Level 1	○ ✗	Difficult Level 1	
Medium Level 2	○	Medium Level 2	

Item Bank

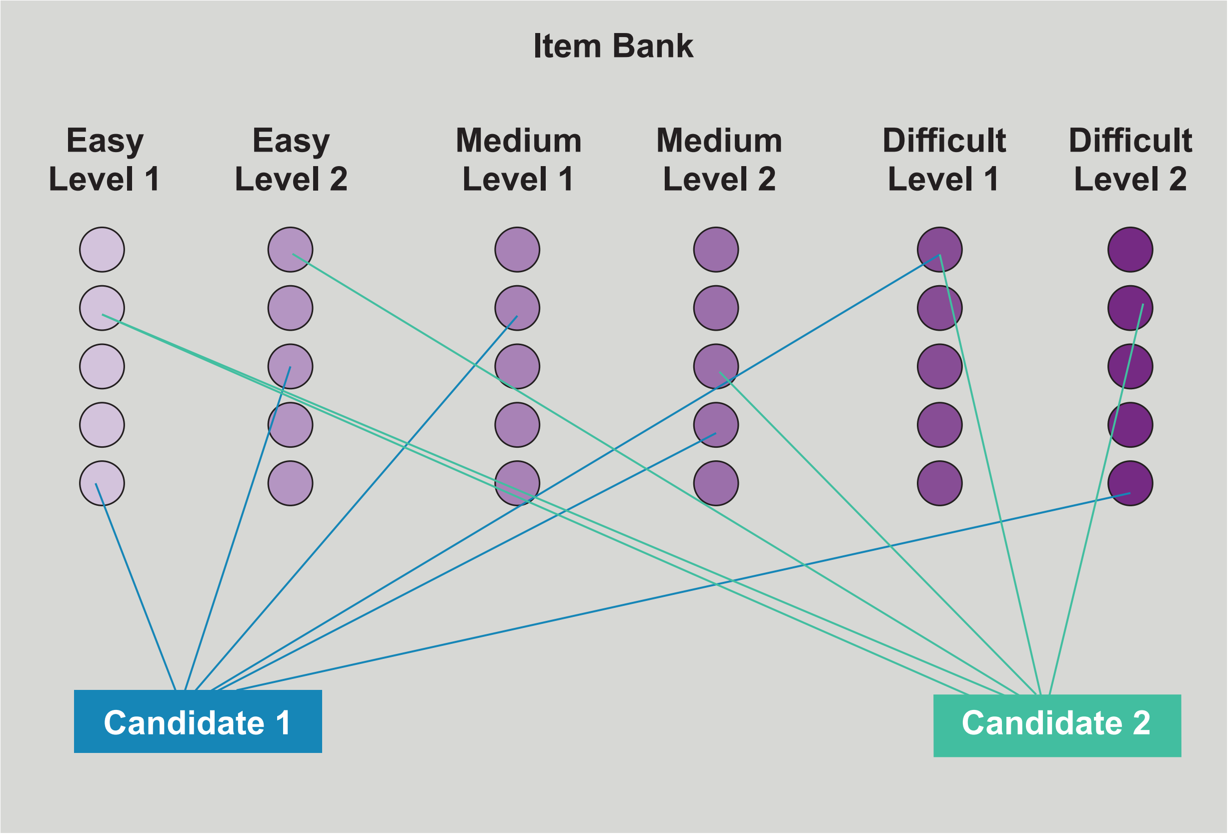
Easy Level 1	Easy Level 2	Medium Level 1	Medium Level 2	Difficult Level 1	Difficult Level 2
○	○	○	○	●	●
○	○	○	○	●	●
○	○	○	○	●	●
○	○	○	○	●	●
○	○	○	○	●	●

Gradient Step Testing (Fixed Length)

Items are randomly drawn from the bank and the number of items and difficulty level remains consistent. Each candidate would do a six-item test of equal difficulty.

Gradient Step Testing is often preferred to Adaptive Testing as:

- Candidates are given tests of the same difficultly which gives the perception of consistency
- Candidates are given tests of the same length which gives them the feeling that they have an equal opportunity to show what they can do



Job Analysis and Assessment Choice

An important concept of Job Analysis is that the analysis is conducted on the job, not the person. While data may be collected from incumbents through interviews or questionnaires, the product of the analysis is a description or specification of the job, not a description of the person to be hired.

Job Analysis is an essential pre-requisite to choosing which psychometric tests and questionnaires to use. In assessment, good job analysis focuses on things that can be defined clearly and measured well.

What is Job Analysis?

Job Analysis is a detailed process to identify and determine the particular job duties and requirements, and the relative importance of these duties for a given job.

Why do we do job analysis?

- Defining role profiles/job descriptions/person specifications
- Job sizing; job analysis can help determine the overall size of a role and therefore the appropriate pay grading required for it
- Developing a framework of criteria for assessment e.g. behavioral competencies

Good Job Analysis leads to:

- Things that can be defined clearly
- Measurable concepts

Less effective Job Analysis leads to:

- Loosely defined behaviors/skills which cannot be measured easily
- Behaviors/skills which cannot be measured easily

Common Methods of Job Analysis

Traditionally, job analysis was very time consuming and involved methods to collect information from multiple sources.

Structured interviews:

- Job holders can be interviewed about important behaviors required to be effective in their role, e.g. Critical Incident Technique prompts an individual to explain the positive or negative impact of an action on a specified outcome
- Line managers can also be interviewed to establish the requirements to perform well in a given role, e.g. Repertory Grid Comparisons can be used to compare competencies in terms of their importance for a job
- Visionary interviews can be conducted in a structured way with a mixture of stakeholders to establish the key requirements for a role going forwards

Job content reviews:

Another method of job analysis is job content review. Reviewers analyze what is important for a given role by studying the job via different methods that can include

- Diaries
- Observing the job
- Doing the job
- Task/job analysis questionnaires
- Validation research

Validation research

Another way to conduct job analysis is through validation research. This can take time and be costly.

- Large samples of job holders or applicants
- Establishing statistical links between test scores and job performance

Methods like these, including structured interviews, focus groups and visionary interviews can now also be supplemented with much faster, online methods such as the Saville Assessment Job Profiler, a multi-rater assessment or in-person or online card sort exercises. Using these methods in combination can be much more resource friendly as they are less time-consuming.

Job Profiler and Card Sort

The Saville Assessment Job Profiler tool and the Wave Performance Culture Frameowork Card Sort can be used to supplement different job analysis methods

Saville Assessment Job Profiler

- Job Profiler is an online tool that takes 15 minutes to complete
- It can be used to survey different stakeholders within an organization on the importance of different behaviors and aptitude areas to a given role
- Stakeholders are asked to rate 36 behaviors and 6 aptitude areas on a 1 – 7 scale from Not Important to Critically Important, giving an overview of which areas are most relevant to the job in question. The resulted job profile aggregates the views of all stakeholders together to provide key guidance on which behaviors to assess and which aptitude areas should be evaluated using suitable aptitude assessments
- Stakeholders can also leave free-text comments on what they think is crucial to performing well in a given role

Saville Assessment Card Sort

The Hire Talent Card Deck includes: Behavior cards showing the section and dimension levels of the Wave Performance Culture framework, Ability cards showing the dimension and facet levels of the Wave Performance Culture framework, Scale cards providing structure to rank each indicator's level of important and a Question card providing direction for card sort exercises.

Using a card sort activity, stakeholders are encouraged to discuss and identify all performance indicators using 12 Behavior section cards and six Ability dimension cards. Subsequently, the Question and Scale cards can be used to facilitate further discussions of the level of important of each indicator, and to confirm the selection of relevant aptitude assessments from the Saville Assessment portfolio.

Considerations for Choosing Assessments

Early Considerations for Choosing Assessments

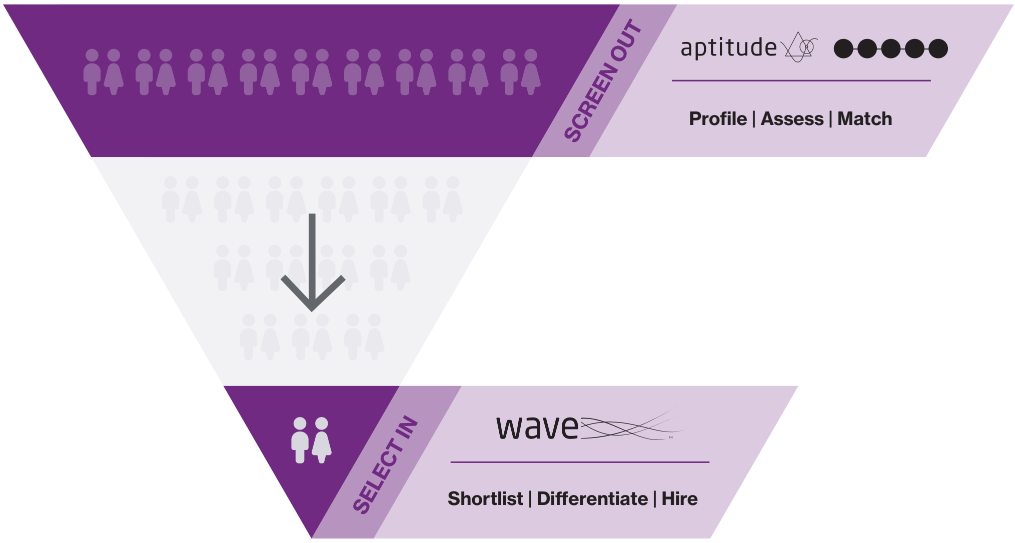
Below are some things you need to consider early on when deciding the assessments that you will use. You should do the test yourself and ask:

- Does it look good?
- Does it make sense?
- Is the content relevant to the role?
- Does the content appear fair and inoffensive?

Other things that needed to be considered are:

- Is it psychometrically sound?
- How much does it cost in total?
- What are the administration practicalities (screening online, supervised final stage, number of candidates, etc.)?

Screen Out, Select In



You should also think about the point at which these assessments will be used, are they early-on screening assessments used to remove large numbers of applicants or are they later in the process, selection tools used to shortlist and differentiate a smaller number of candidates.

Our Methods of Screening

Aptitude tests

- Longer assessments that look at one ability in depth or shorter tests that assess several abilities more broadly

Behavioral screening questionnaires

- Short behavioral assessments that can provide one fit score for rapid decision making in screening

Language tests, e.g. Workplace English

- Workplace English tests assess an individual's ability to understand workplace-relevant sentences in English

Situational Judgment Tests

- Situational Judgment Tests or SJTs provide engaging, realistic, work-related previews of the role by presenting candidates with scenarios they are likely to come across in the job

Why Use Aptitude Tests?

Hire

- Aptitude tests are mostly used for recruitment, either in screening or selection

Build

- Tests are used less frequently for individual development, although career guidance and planning tools often contain an aptitude component
- They do predict training performance!

Lead

- Despite many leaders' avoidance of testing, cognitive ability has been shown to be especially predictive of performance at senior levels

Test Feedback

Why do we give Test Feedback?

Even if candidates are not successful, research shows that they are more positive about an organization if they have received feedback on their performance. In a review of a large communication company's selection process, it was found that 6% of rejected applicants disconnected from the company as customers. Over the course of a year it is estimated that £4.4 million revenue was lost because of the poor candidate experience and subsequent disconnection.

Feedback also gives candidates a chance to understand why tests are used by the organization and the rationale behind using them. Candidates and feedback providers can discuss the pattern of results displayed, and discussion of examples from their working life can aid understanding. Where there are differences between ability test results and findings gained from other assessment center exercises involving numerical calculations, for example an in-tray forecasting exercise, these can be explored. Candidates can gain valuable insight into their relative strengths and development areas, which can help guide personal development in specific areas. Candidates have the right to see any information held on them, including assessment results. It is best practice to provide this in an appropriate and accessible form, such as verbal or written feedback. Candidate feedback reports are designed for this purpose.

Public Relations

- Candidates are often clients too
- When candidates have a poor experience they often take their custom elsewhere which can lead to huge losses for companies

Candidate Experience

- To assist in explaining why a job offer was not made
- To increase the recruited candidate's self-awareness
- To develop individuals in key ability areas

Assessor Experience

- To understand results by seeking examples/explanation
- To understand conflicts with other assessment data

Applicable Legislation

- To comply with applicable legislation

Feedback of Saville Aptitude Tests

Saville Assessment ability tests give additional, unique information regarding test performance on different categories of questions on a test. These measures break down the overall Total Score into Item Type or Aptitude Area Sub-Scores which help pinpoint very specific strengths and development needs, providing recruiters and candidates with an in-depth understanding of result patterns. The test-taking style measures provide added insight into how the candidate completed the test. Depending on whether the test administered is a supervised or unsupervised version, different forms of test taking style information are provided.

Unsupervised, item-banked Tests:

Pace – how quickly the individual has responded to the questions

Aptitude – how well the individual has performed on the test

Supervised, fixed-content Tests:

Speed - the proportion of test questions answered in the allocated time

Accuracy - the proportion of correct answers

Caution - the difference between Speed and Accuracy

Saville Assessment online profile charts and feedback reports are designed to be used by a wide audience, including candidates, trained users and line managers, giving more flexibility. Graphic displays are used to ensure quick and straightforward interpretation and feedback.

Features of Different Feedback Forms

You can choose to deliver feedback in different ways.

Report

- In high volume hiring processes you can share the report with candidates without a feedback session
- These reports are intended to be used by hiring managers, trained users and candidates
- You may choose to only give spoken feedback on this report in smaller recruitment pools or later into a selection process alongside feedback from other parts of the process (e.g. other assessments or interview)

Telephone

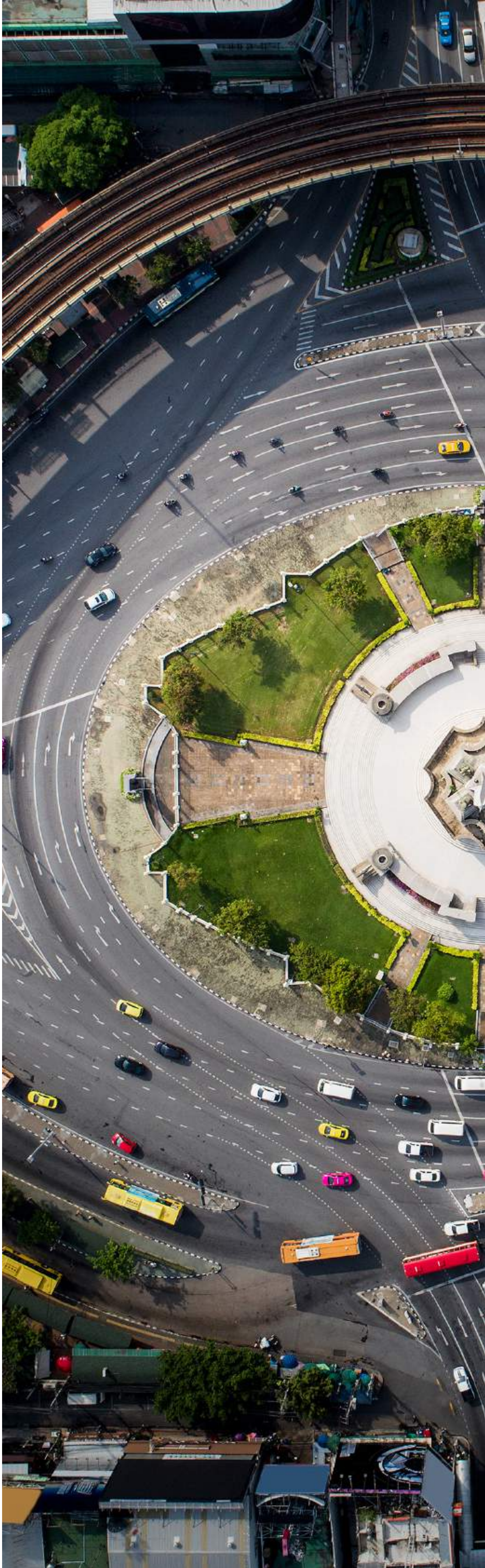
- With a smaller applicant pool or later in a selection process you may want to talk through a report with a delegate over the phone
- You will need to send their report to them ahead of the call
- This is a quick and convenient way to deliver feedback but you may miss out on visual cues from the candidate

Video Call

- A video call can be used in smaller applicant pools or later into selection processes
- Video calls can be quick and convenient and you can also pick up on visual cues such as eye contact which can help to build rapport
- You can “screen share” the report with the candidate

Face to Face

- Giving feedback face to face can be more difficult to organize
- You might only cover aptitude feedback in person alongside feedback on other parts of the process
- Consider how you set up your room and where you position the report to support conversation



Feedback Process

Introduction and purpose

- Timing, purpose, confidentiality, two-way process

Summary of tests used and comparison group

- Recap of the aptitude areas that have been assessed, the rationale for why the tests have been used; their link to the role requirements. Explain the benchmarking of the assessment and its 'shelf life'

Background

- Has the candidate completed assessments like this before?

Encourage self assessment

- Candidate experience when completing
- Reflect on areas of strength and more challenging areas. Gauging how well the candidate thinks they have done can be helpful when delivering below average scores

Discuss overall performance

- Feedback total score, ask candidate for their thoughts and then feedback all sub areas

Review pace and sub-scores

- Break down any sub-scores of a combined assessment
- Pace refers to how quickly the candidate has completed the assessment and it's helpful for the candidate to reflect on for future assessmentss

Summary and review

- Wrap up your session by summarizing everything you have covered and answer any questions the candidate might have

Feedback Tips

Effective Feedback

- Do build rapport (e.g. give eye contact, ask the delegate how they found the assessments)
- Do ensure two-way dialogue (e.g. invite the candidate to ask any questions throughout and be sure to check their understanding)
- Do gauge candidate reactions and impressions (e.g. ask the candidate whether their scores are in line with their expectations, ask what they think about their performance)
- Do discuss development areas if raised by candidate (e.g. if a candidate has referred to a particular area as a challenge for them, you may want to spend a little more time discussing relevant development tips)

Less Effective Feedback

- Giving too much information or too many scores at once can confuse the candidate (i.e. give your candidate time to process the information and space to ask questions)
- Be mindful of using technical jargon (i.e. talking about specific report scores without explaining what they are or what they mean)
- Try not to make value judgments (i.e., “That’s really good...Oh that’s not so great.”)

Test Scores and Test Norms

Ranked Scores

Scoring

An applicant scores 19 correct (their raw score) on a numerical test (28 items).

How well has the applicant performed?

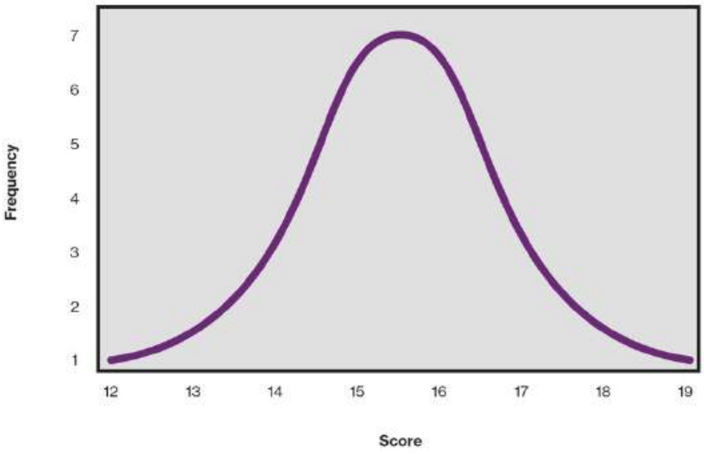
We can't tell, for a score to be meaningful to us, we need some way of comparing the score achieved by an individual on a test against the scores achieved by a representative/relevant sample of people, i.e. a benchmark or norm group.

Frequency Distribution

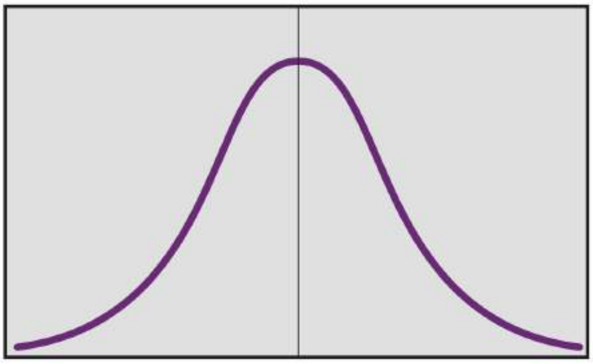
One way to look at a group's scores is to produce a frequency distribution. On the horizontal (x) axis, the scores on the test or assessment are presented and on the vertical (y) axis, the frequency (or count) is presented. The frequency count for each score is plotted on the graph to give us our frequency distribution.

- A frequency distribution can help us to make sense of a group's scores
- This graph shows the frequency of each score achieved on an example assessment
- We can see the highest frequency around 15 and 16 meaning that this is how well most people performed on the assessment

Score	12	13	14	15	16	17	18	19
Tally	|	||	|||	𝍸 ||	𝍸 |	|||	||	|
Count/ Frequency	1	2	3	7	6	3	2	1



The Normal Distribution



Scores which, when plotted, form a smooth curve like the one depicted are said to be 'normally distributed'. This curve is sometimes called a 'bell-shaped curve'. You can see that most scores fall around the average (bulge in the middle), with fewer occurrences towards the far left (low scores) and towards the right (high scores).

Most natural phenomena are normally distributed. If you were to plot the shoe sizes of a large number of women in the UK, you would discover that the distribution normal. You would find the same with height and weight. Using these normal distributions, you are able to get a sense of where you stand compared to others. You can start to answer questions such as: Are my feet big, small or average? Am I tall, short or around average? How well have I done in a numerical test? The normal distribution and its unique properties are the basis for all test norm systems.

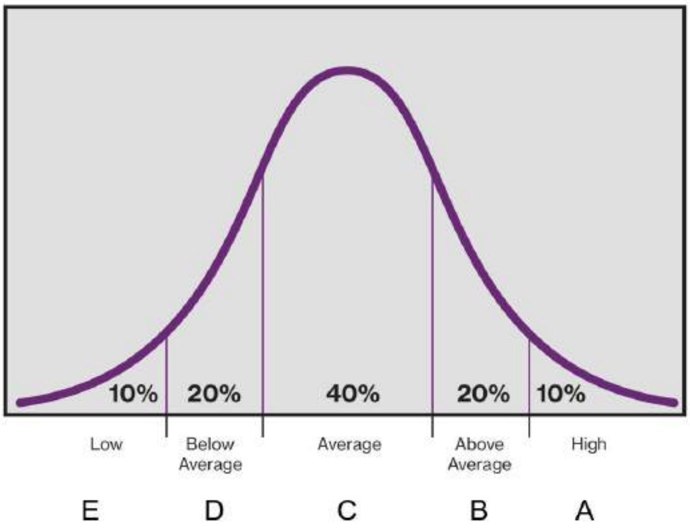
- A smooth 'bell curve' frequency can be described as a normal distribution
- Most scores fall around the middle, or average
- Fewer individuals have scored at the tail ends of the curve showing very few have particularly low or particularly high scores
- We find that many natural phenomena are normally distributed in this way; e.g. shoe size, height, weight

The Normal Distribution Performance Bandings

Grades or bands are a type of rank order norm scoring system.

The area under the curve represents the total percentage of people who have taken a particular test. We are able to chop up the curve into bands of average, above average, below average, high and low. Or indeed grades of A, B, C, D and E. Grades and bands are one of the simplest norm systems we can use.

- Performance bands are one way of describing scores
- We can see the percentage of individuals at each percentage band
- We can label these divisions as Low, Below Average, Average, Above Average and High
- Another example of performance bands are lettered grades like those we see in school and college



Rank-ordered Norm Scores

Percentiles

Percentiles are essentially an extension of the Grade System – they are ‘graded grades’ whereby instead of having just five bands, you have many bands, giving you a more sophisticated grading system. In fact, the percentile system splits the normal distribution into 100 bands, each representing 1% of the comparison group or population under the curve. Percentile is defined as ‘per cent’, or ‘of a hundred’.

The normal distribution can, therefore, be thought of as being divided up into percentiles. A percentile rank indicates the percentage of the norm group a person’s score comes above. A score at the 60th percentile means that the individual performed better than 60% of the group (while 40% of the group have performed better than them). The way to describe a score at the 60th percentile is to say “you have performed better than 60 percent of the comparison group”. This phrasing is useful when feeding back test scores to candidates or line managers.

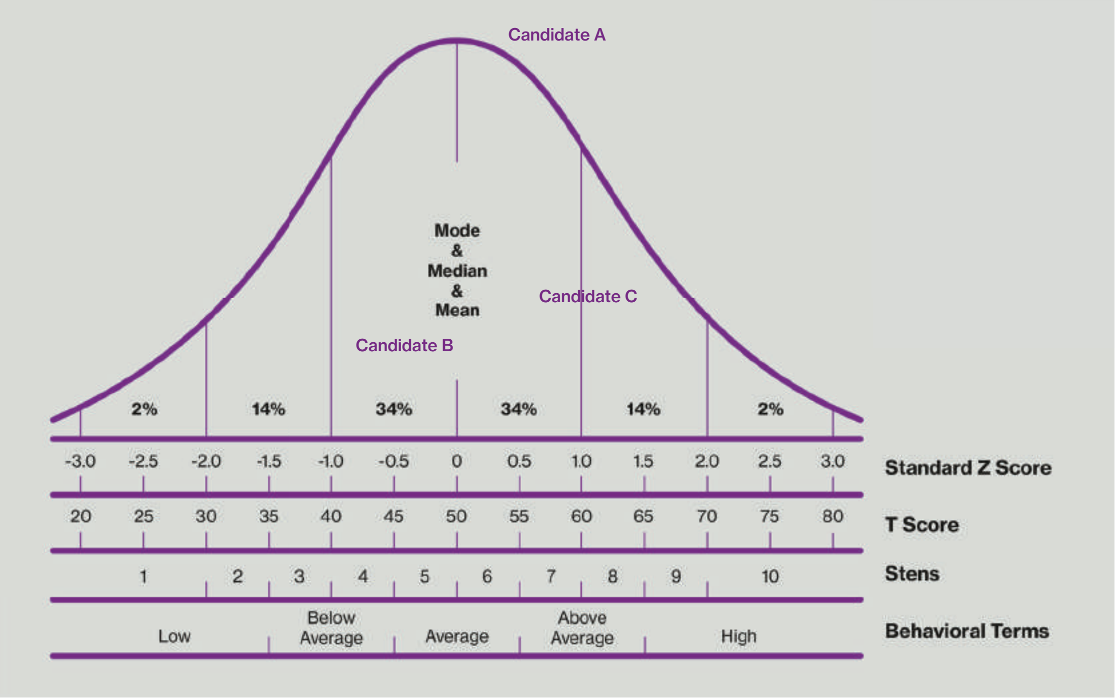
- Percentiles are an extension of performance bands
- Instead of five grade bands, we split a normal distribution into 100 bands, or percentiles, which each represent 1% of the comparison group
- Percentiles are rank-ordered meaning that a person scoring at the 60th percentile is essentially “in front of” 60% of the comparison

Describing Percentiles

- Candidate A: 56th %ile
- Candidate B: 33rd %ile
- Candidate C: 68th %ile

All of these candidates are actually average, even though their percentile scores look quite different. With a normal distribution, the majority of scores fall within the middle of the distribution, around the mean. This is the widest part of the curve.

It is important to remember that percentiles are not equal units of measurement. An increase from the 87th to the 99th percentile is a greater performance improvement than an increase from the 56th to the 68th percentile. Percentile scores can therefore be said to reduce the difference at the extremes and exaggerate scores around the middle of the distribution. When using percentiles, it is, therefore, key that you do not over-read small differences between applicants; one or two raw scores difference near the average can result in a large gap in percentile terms. This leads to a major practical problem - you cannot take percentiles from different tests to produce overall composite scores. In order to get around this problem, we use scores called standard scores, should an occasion arise when you wish to add or take an average of a set of scores.



Standardized Scores

Additional Scoring Systems

Where is the middle, how spread out is the group?

Measures of Central Tendency

There are two key measures of a normal distribution; where the middle is and how spread out the data is. By understanding these measures and how they impact on the shape of the curve, we can start to understand things like how easy or difficult a test is.

Measures of Central Tendency, the Average:

The Mode - The mode is the most frequent score in the set of scores. Occasionally you may have more than one mode in a set of scores.

The Median - The median is identified by lining up scores in order and finding the middle number. It is the number which has 50% of the scores above it and 50% of the scores below. If there are even numbers of scores, then calculate the number which would be half way between the two middle numbers.

The Mean ($\bar{x}$) - The arithmetic mean is often referred to as the average of the set of scores. The mean is calculated by adding all of the scores (X) in the group to find the total and then dividing the total by the number of people in the group (N).

Measures of Spread:

Range - One measure of the spread of the group is the range. The difference in scores between the candidate with the highest score and the candidate with the lowest score is called the range. Consider the scores shown in ***Table 1***. Both Test A and Test B show a mean score of 50, however Test B shows much greater variation in the scores of the group (the lowest is 10 and the highest is 90) compared to Test A (where the lowest score is 46 and the highest 54). The range can be misleading if the set of scores you are looking at contains any outliers, where candidates have extremely low or high scores. This can give you a large range but does not give you a good general idea of the actual spread of scores for the whole group.

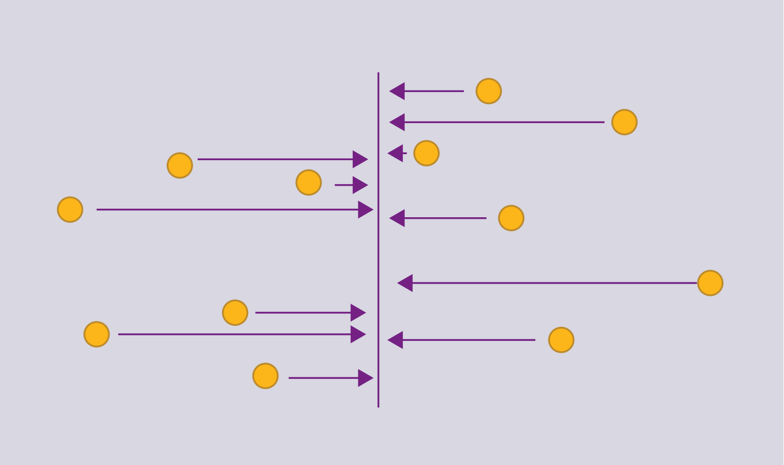
Table 1

	Test A	Test B
Candidate 1	54	90
Candidate 2	53	80
Candidate 3	52	70
Candidate 4	51	60
Candidate 5	49	40
Candidte 6	48	30
Candidte 7	47	20
Candidte 8	46	10
Mean Average	$\bar{x}$=50	$\bar{x}$=50
Range	8	80

- Range is the lowest score subtracted from the highest score
- Range is easily influenced by outliers (extreme scores) so we can't rely on range alone when looking at spread

.

Standard Deviation - The Standard Deviation (SD) is a statistic which tells us about the spread of scores around the average or mean and gives us a more robust measure of what is termed 'dispersion' of scores. This is essentially the average of the spread of scores around the mean of those scores. The SD tells you on average how far away scores are from the average score – it is a measure of the variability of the scores in that group. Thus, a group of scores can exhibit a high degree of variability (high SD) or a low degree of variability (low SD), indicating a more homogenous group.



- SD is a statistic which robustly measures how spread out scores are based on the mean of the data
- The SD tells you on average how far away scores are from the average score

Calculating Standard Deviation

SD = the standard deviation

$\sum$ = the sum of

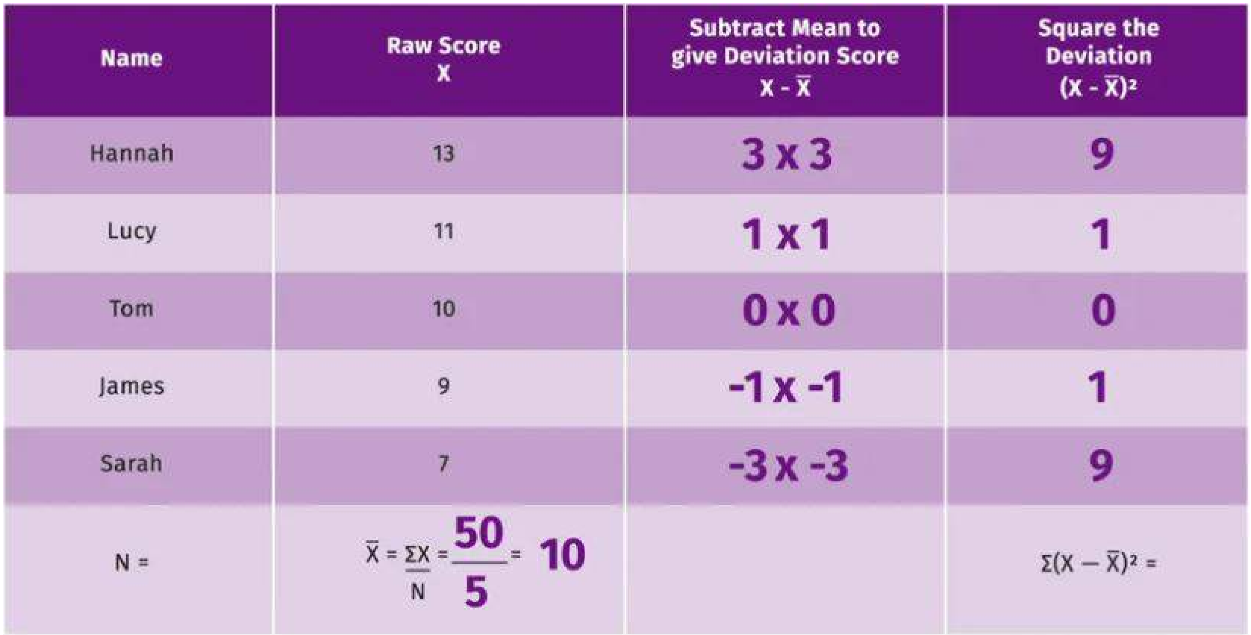
X = a raw score

$\bar{x}$ = the group mean score

N = the number in the group

$$SD = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

These candidates have completed an ability test and this column shows the number of correct responses they have achieved.



Name	Raw Score X	Subtract Mean to give Deviation Score $X - \bar{X}$	Square the Deviation $(X - \bar{X})^2$
Hannah	13	3 x 3	9
Lucy	11	1 x 1	1
Tom	10	0 x 0	0
James	9	-1 x -1	1
Sarah	7	-3 x -3	9
N =	$\bar{X} = \frac{\Sigma X}{N} = \frac{50}{5} = 10$		$\Sigma(X - \bar{X})^2 =$

Step 1 is calculating the group mean. We add up all of the raw scores, 50, and divide by the number of people in the group, which in this case is 5. This gives us a group mean of 10.

Step 2 we subtract the group mean from each raw score; for example the first candidate has 13; we takeaway the group mean and we are left with their deviation score of 3.

Step 3 You can see that we have some negative deviation scores so our next step is to square those numbers to cancel out the negatives; squaring a number means we multiply the number by itself. When we do this all negatives become positives.

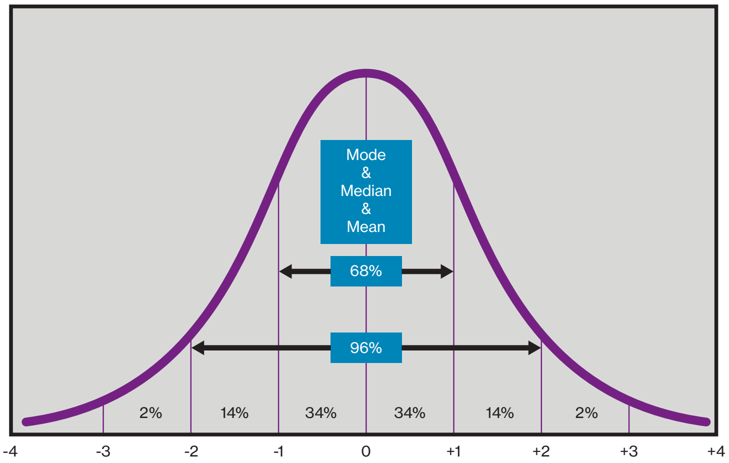
Step 4 In the final column to the right, we add up the squared deviation scores and in this example the answer is 20.

Step 5 We can now put these numbers back into our formula. The total from step 4; 20, is divided by the number candidates; 5. This gives us a value of 4.

Step 6 The final step is to find the square root; we can use a calculator to do this. The square route here is 2.

This score tells us that, on average, each raw score will differ from the group mean by 2 raw scores. This is the standard deviation, which is a statistically robust measure of spread, or variance, of scores in a group.

SDs and the Normal Distribution



To illustrate how standard deviation and the normal distribution work together, we can consider the following example using the data we have just calculated.

Let's imagine we look at a larger sample with the same group mean, 10, and standard deviation, 2. We can plot these values on the distribution curve.

Adding the mean, we can see the average or mid-point scored in this group. If a person performs better than the average by one standard deviation they will have a raw score of 12. If better by 2 standard deviations, and three standard deviation their raw score will be 14 and 16. We can also see this when candidates achieve less than the average score. When scores are normally distributed, we would expect 68% of scores to be between 8 and 12 which is 1 standard deviation above and below the mean. We would expect 96% of scores to be between 6 and 14 which is 2 standard deviations above and below the mean. Should you find much more the 68% of scores occurring 1 standard deviation below and above the mean, it looks as though there is not enough variation in the data.

We would also expect all measure of the average, the mean, median and mode should be approximately the same.

When we construct norm groups, we take account of these considerations, amongst others, to establish the suitability of the data to be used for benchmarking.

Other Ways of Describing Scores

Z Scores

By looking at units of standard deviation, we have been using a new scoring system without even knowing it! We have been using the Z score system. Z scores are a vital concept in psychometrics and form the basis of all standard score norms.

The Z score scale represents the number of standard deviation units a raw score is above or below the group mean. Z scores are equal units of measurement and (unlike percentiles) you can add Z scores from different tests to produce overall composite scores. To calculate the Z score you need to know the group mean and the group standard deviation. The Z score is calculated using the formula presented here.

Typically, the Z score scale runs from approximately +3 to -3. The average of a Z score scale is 0 and the standard deviation of a Z score scale is 1. There are some disadvantages in using Z scores to describe people's scores on tests – they are decimals and they also have positive and negative signs. This makes them difficult to interpret from the test user, line manager and candidate point of view and, whilst it is possible to produce composite scores, it is sometimes difficult to do other mathematical calculations with them.

- When we develop test norms, we look at the composition of the sample; 99% of participants should be within +/- 3SDs. We want the scores to be normally distributed so that is represents the population and we can band the sample in a useful way.
- They can be less user-friendly for feedback
- It can be difficult to do mathematical calculations with them

T Scores

To overcome the issues created by having a scale that runs through from positive to negative, we can simply transform the Z score into a T Score (Transformed Score).

The T Score scale has an average of 50 and standard deviation of 10. Whilst T scores are widely used in the interpretation and analysis of ability test scores, they aren't particularly candidate friendly and so you will rarely see them on reports.

- Use z-score formula to derive T-scores (Transformed Scores) which do not include negative numbers
- This makes them helpful for analysis but they are still less candidate-friendly in feedback

Sten scores

The Sten scale (or standard ten scale), is another way of describing scores. Stens tend to be used when measuring personality, style or motivation, because less detailed differentiation is required for these attributes. The Sten scale divides scores into 10 categories, with 1 being low and 10 being high. The mean of the Sten scale is 5.5 and the standard deviation is 2. When describing Stens, it is customary to round up the Sten to the nearest whole number. The disadvantages of using Stens are that they are broad bands, like grades, and you cannot make fine-tuned judgments about differences between scores. However, converting ability test scores to Stens can be useful in assessment or development centre situations, where other assessment exercises have been scored or rated using a 1 to 10 scale.

- Use z-score to derive stens, a standard to ten measure where 1 is low and 10 is high meaning that they can easily be compared to other 1-10 scores
- Stens tend to be used most often in personality measures as they don't give as much granularity as a hiring manager may want from an ability test

Norms

Choosing Norms

What is a norm?

A norm group is the sample group against which a candidate's scores are compared. A norm group can be regarded as a sample, from which a set of scores have been gathered to provide a representation of the population it is intended to represent (e.g. managers, graduates, call centre staff or the general population).

How important is norm size?

A large group not guaranteed to be representative and attention should be paid to the sampling method that has been used and whether there is an appropriate spread of people in the group. There are a number of different ways of collecting samples and developing norms. The main sampling methods include random, stratified and usage sampling.

The standard error of the mean (SEmean) allows us to estimate the distance of our sample mean from the population mean. It can be calculated using the mean of the sample, the SD and the sample size. As the sample size gets larger, the SEmean will reduce.

- A norm group may not reflect the reality of a whole population
- The standard error of the mean is a calculation that estimates how representative our norm group is of a population
- As the sample size gets larger, the standard error of the mean reduces and, therefore, the norm group will generally be more representative
- Here the mean and SD for the norm group are smaller than the population – raising possible questions about the representativeness of this sample
- Small norm groups (<150) are more likely to be unrepresentative
- Size matters only to a point – increasing the norm beyond 500 will make little practical difference

Norm Sampling Methods

Random Sampling

In random sampling, you randomly select people to include in your sample, typically with the entire population having equal chance of being selected – this can be challenging to achieve when your population is large and broad, but can reduce sampling bias.

- People are randomly selected from the population for the sample
- Can reduce sampling bias
- Difficult to achieve with a large population

Stratified Sampling

Stratified sampling is where you purposely select a sample that is representative of your intended population, for example if your population consists of 50% males and 50% females, you aim for the same proportions in your sample. However, even with a sophisticated methodology, it is still challenging to ensure the candidates included in the sample represent motivated candidates. Therefore, the final data from random or stratified samples are likely to be different from realistic live usage samples.

- Individuals are specifically selected for a sample to represent the population they are drawn from
- It is challenging to sample individuals that represent real test candidates

Usage Sampling

Usage sampling selects those who have previously completed a test in a real application for the sample. This is a common method of collecting norm samples; an advantage of this is that those within the sample are usually realistically motivated to complete the tests.

- Individuals are chosen from pools that have gone through real test situations
- These samples are often convenient and include realistically motivated candidates

The Saville Assessment Approach

We combine the advantages of using motivated usage data with the techniques of careful stratified sampling to create norm groups which are both representative and based on realistic test completion data.

- Saville Assessment carefully stratifies usage data

How do you choose the appropriate norm?

When choosing an appropriate norm group you should always consider the job being applied for. It would be appropriate to use a norm group of graduates for graduates entering an organisation. It would not be appropriate to compare graduates' scores on a numerical test against a group of 16-year olds, nor would it be appropriate to compare them to a group of experienced employees. It is therefore important to take into consideration things like educational level and work experience in order to ensure that your norm group is representative of your candidate population.

- Consider the role that candidates are applying for
- Take into account the education and work experience of applicants

You can find more information on our norms in our brochure and detailed norm descriptions are available in the Client Resource Area, which you will be able to access when you complete the training course.

Correlations and Reliability

Interpreting Correlations

Correlation is used to test reliability and validity; two incredibly important concepts when it comes to testing.

The correlation coefficient was first conceived by early psychologist Sir Francis Galton. Subsequently refined by statistician Karl Pearson, the main coefficient became known as Pearson Product Moment Correlation Coefficient and is universally given the letter ‘r’.

Correlation coefficients show how two variables relate to each other. To run a correlation on a group, you need to be able to measure two things about each person in the group and then you can calculate how they are related.

- Correlation is a statistical technique for establishing whether there is a relationship between two things
- The degree of relationship is expressed using a correlation coefficient (r); the further away from 0, the stronger the relationship
- Correlation is commonly used within research to establish reliability and validity
- You can calculate how they are related using the Pearson’s Product Moment Correlations Coefficient which gives you the exact correlation coefficient value; e.g. r = .8

Strength and Direction of Correlations

Correlation answers questions about a group of people, e.g. ‘Do people who do well on verbal assessments do well on numerical assessments?’ Relationships can be assessed both as a matter of direction and strength.

Direction

Relationships can be either positive or negative. If people perform well on verbal and numerical assessments, we can describe this as a positive relationship. That is, the higher your performance on the verbal test, the higher your performance on the numerical test.

Conversely, a negative relationship is where people perform well on the verbal test but poorly on the numerical test; or vice versa. Direction is indicated by a plus or minus sign.

Whether the line is angled to the left or to the right tells you about the direction of the relationship (positive or negative).

Strength

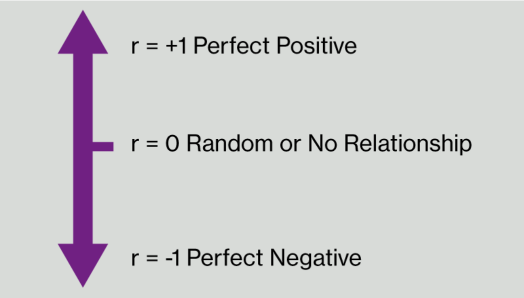
In addition, correlations vary in strength – from being perfect, where you can exactly predict scores on one variable from another, to being very weak or non-existent. The strength of the correlation is shown by the size of the r value.

How close the data is to a straight line tells you about the strength of the relationship (how close to 1).

Scattergrams

Scattergrams are a good visual representation of correlation.

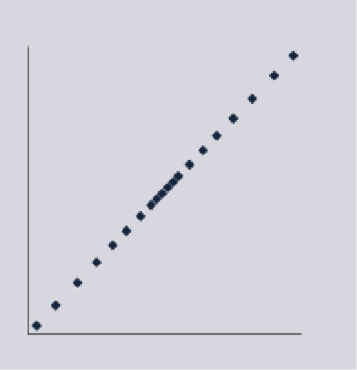
Here are some examples of scattergrams showing the relationship between verbal and numerical performance in different groups.



r = +1 Perfect Positive

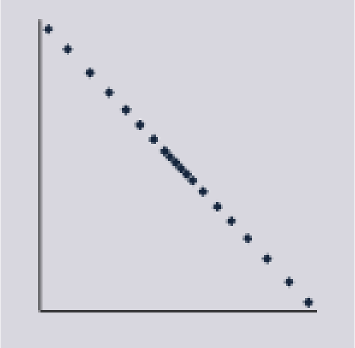
r = 0 Random or No Relationship

r = -1 Perfect Negative



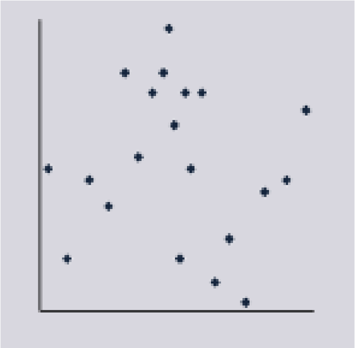
Perfect Positive

The correlation is plotted as a perfect straight line, as one variable increases, the other increases at the same rate.



Perfect Negative

This correlation is another perfect straight line but is in the other direction. That is, as one variable increases, the other variable decreases at the same rate.



No Relationship

Where there is no relationship, we cannot see a clear pattern in the data, there is no clear line; instead the points appear randomly scattered. Scattergrams are a helpful way to “eyeball” data to get a feel of the relationship between variables, however, it can be time consuming and is not specific; it doesn’t give you the correlation coefficient value.

Using Correlations

Maximizing or Minimizing Correlations

Several factors can maximize or minimize the strength of your correlations.

Maximize

- No time delay between measures; e.g. verbal and numerical test scores collected at the same time in a research setting
- A strong underlying relationship between two variables

Minimize

- Many influencing factors; e.g. external variables which can impact the relationship
- Poor measures of variables; you can't accurately measures one or both of your variables

Correlations Confidence and Causality

We looked at scattergrams earlier on and they gave us a good visual representation of the relationship between variables. However, we also want a level of confidence in the result which we can find with Statistical Significance. This explains the likelihood of whether a results has come about through chance or there is a statistically-meaningful relationship.

Are we confident?

- Statistical significance gives you a level of confidence in the result
- Lack of statistical significance indicates that the result is likely to be a chance finding

Correlation vs. Causality

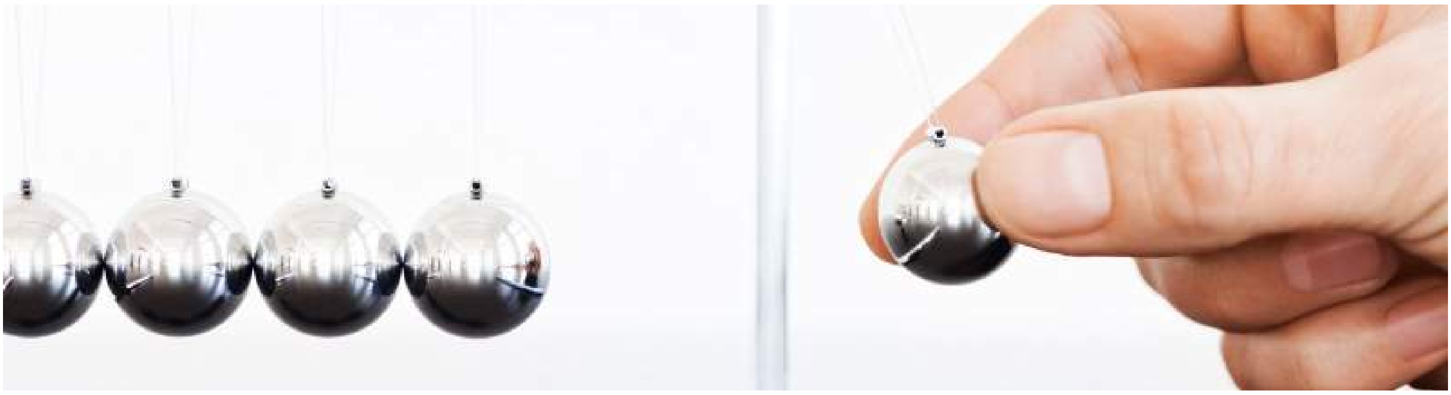
Correlations can tell us whether there is a relationship between two variables and how strong this relationship is. However, they cannot tell us whether one variable causes another. Two variables can be highly correlated, for example the sale of ice cream and increase of sunburn. They're highly correlated but one does not cause the other. However, when you do have a causal relationship, you will also find that those two variables are highly correlated.

- **Correlation does not tell us about causality**
- When variables are related, we cannot tell whether one leads to the other
- It is common that a correlation may be caused by an external factor which causes both variables to change systematically

Population Factors

Different populations may have different correlations.

- The smaller the sample, the greater the potential for sampling error and the greater the chance of obtaining a misleading result
- This means that looking for small or moderate correlations with samples of less than 100 does not give results with much meaning
- Ideally we require a sample of 200 or even higher, to get a result from one sample that has small enough confidence intervals for meaningful interpretation



Reliability

Reliability is fundamental to measurement and concerns how precise and error-free a tool is in measuring desired constructs. Any instrument that measures something in the real world needs to have a level of precision or accuracy, for example, weighing scales, a digital clock or a light meter in a camera. The greater the reliability or precision, the greater the chance that it will allow for valid decision-making.

- Nothing can be measured with absolute accuracy
- A test's reliability concerns the precision and consistency of measurement
- Classical Test Theory (CTT) assumes that:
  Observed score = true score + margin of error
- The more error in test scores the lower the reliability
- Generally, the longer the test the greater the reliability
  - Note: Use total scores on Swift tests for decision making
- Reliability is a prerequisite for Validity

Sources of Test Error

Error can creep into the testing process from a variety of sources: the candidate, the test administrator and test environment, and the test developer. The aim is to standardize the testing session to minimize error and maximize the reliability of the results.

Candidate

If the candidate feels unwell, has not prepared or reviewed practicse materials, misinterprets the test instructions or experiences severe test anxiety, these factors could all mean they do not give their best on a test.

- Feeling unwell
- Misinterpreting instructions
- Severe test anxiety

Administrator

If the test administrator has chosen a test which doesn't accurately measure what it claims to measure, e.g. a poorly constructed verbal reasoning assessment this could impact the candidate's results. Likewise, when administrators do not properly brief candidates or set up the testing environment properly, to minimize disruptions for example, this results in error which can lower the test reliability. The administrator should diligently score hard-copy answer forms, where used, and be sure to accurately interpret results; where this is not the case reliability of the results will be lowered.

- Using an unreliable test
- Poor candidate briefing
- Misinterpreting test results

Test Developer

Test Developers should be rigorous in ensuring the quality of their tests to support the reliability of their findings. This includes writing clear items which lack any ambiguity, giving straightforward instructions and being sure that they are measuring what they claim to be measuring. Reliability is about getting the test right; validity is about getting the right test. It is the test developer's responsibility to develop an accurate test and ensure it is a reliable measure.

- Ambiguous items
- Items measuring the wrong thing
- Poor instructions

An example of an ambiguous item is a really wordy numerical item. This can end up assessing verbal reasoning rather than numerical reasoning so is not a reliable test of numerical reasoning.

Types of Reliability

Reliability can be measured in a number of ways

Test-Retest Reliability

Refers to the stability of a measure over time. It is calculated by correlating scores on a measure completed by the same group of people at two points in time.

+ Gives indication that attribute is stable

- Candidates not willing to do it twice

Alternate or Parallel Form Reliability

Refers to the consistency between two versions of the same measure. This is the correlation between the results for the same group of people who complete two versions of the test.

+ Shows developer is clear/consistent on what is measured

- Has the expense of developing two forms

Internal Consistency Reliability

Relates to the internal correlations of the components of the measure, for example the relationship between the different scales within an assessment.

+ Easy to do as only requires one set of data from one time period

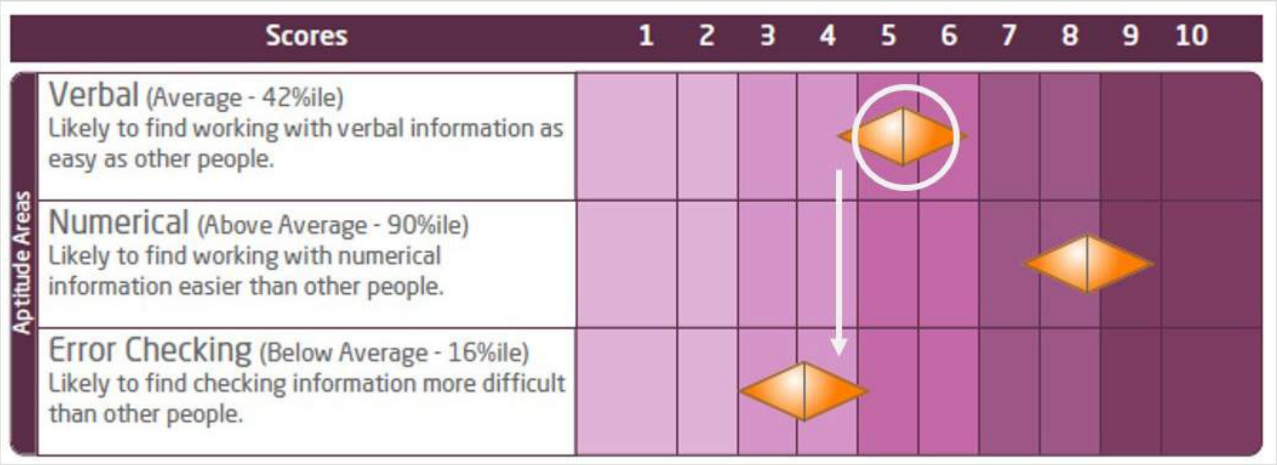
- Can be misleadingly high with repetitive item content

Margin of Error

The Standard Error of Measurement (SEm) estimates the margin of error around test scores. The greater the reliability the smaller the band of error around a test score. How do we know how large the band of error is?

Typically, test publishers will already have calculated these for you; this is displayed on the automated outputs from assessments.

- Reliability is about getting the test right; validity is about getting the right test.
- Margin of error is measured using Standard Error of Measurement
- In this example, the margin or error is shown by the breadth of the diamond in each Aptitude Area Sub-score. This estimates the amount of error there is in a test



Scores 1 2 3 4 5 6 7 8 9 10

Aptitude Areas

Verbal (Average - 42%ile)
Likely to find working with verbal information as easy as other people.

Numerical (Above Average - 90%ile)
Likely to find working with numerical information easier than other people.

Error Checking (Below Average - 16%ile)
Likely to find checking information more difficult than other people.

Calculating Standard Error of Measurement

SEm = Standard error of measurement, the band of error around scores from a test

SD = Standard deviation of scores for the reliability sample, the degree of spread of score within the group that has been assessed

r = the reliability coefficient (the reliability value of the test established through one of the reliability estimate methods, i.e. test-retest, alternate form or internal consistency)

To calculate the SEm, follow these steps:
In this example, the SD is 2 Stens and the r value is .7.

$$SEm = SD\sqrt{1 - r_{tt}}$$

1. Insert the Standard Deviation and Reliability figures into the equation.

2. Subtract the Reliability estimate from 1. e.g. (1 – 0.7).

3. Using a calculator, find out the square root of the answer to step 2.

4. Multiply the answer to step 3 by the Standard Deviation (2 x 0.50)

5. This is your SEm, i.e. 1 Sten Score

6. If we ran this calculation again with perfect reliability 1.0, the Standard Error of measurement would be 0. The better the reliability, the less error.

Applying the SEm

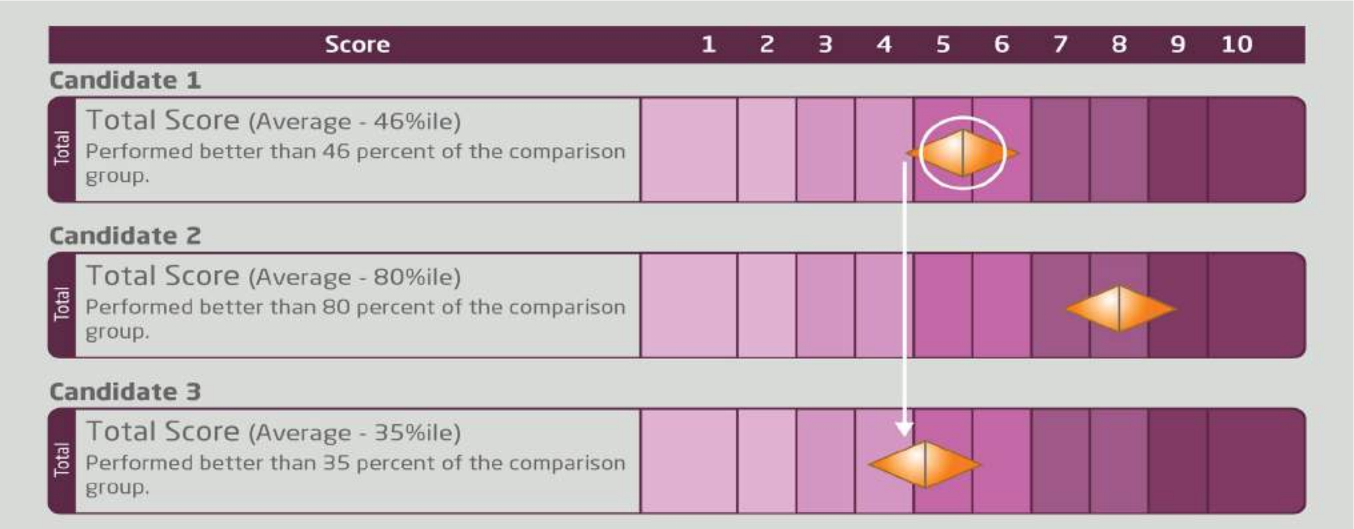
Typically, we show a band that has a 68% probability of capturing the true score within 2 Stens of a displayed score.

We can be confident that 68% of the time, a person's score will lie within plus or minus 1 Standard Error of Measurement of their score. We can be confident that 96% of the time, a person's score will lie within plus or minus 2 Standard Errors of Measurement of their score.

Reliability and Differences in Scores

The example below illustrates applying the principles of reliability and differences in scores in practical terms for three candidates who have completed the same test. You can see from this example that the band of error around each test score represents the Standard Error of Measurement (SEm). This is always given to you on the test reports in terms of number of Sten score marks either side of a candidate's score. In this case, the band of error is 1 Sten either side of the total score.

For you to be able to say there is a difference in scores, the bands of error must not overlap.



Candidate 1

Total Score (Average - 46%ile)
Performed better than 46 percent of the comparison group.

Candidate 2

Total Score (Average - 80%ile)
Performed better than 80 percent of the comparison group.

Candidate 3

Total Score (Average - 35%ile)
Performed better than 35 percent of the comparison group.

Validity and Test Utility

Validity

What is Test Validity?

A test is valid to the extent that it measures what it is designed to measure. In particular, validity is a measure of how relevant a test is to job content. This is a key aspect of using occupational tests; if the test is not valid, then there is little point in using it. You may have a highly reliable test, but if it is not measuring the particular job competency you are interested in assessing, then it is not useful. Remember, that a valid test has to be reliable in the first place.

Types of Validity

We have both, "informal" and "formal" methods of measuring validity in assessments.

Informal

Face Validity

Do the questions and reports 'look right' i.e. appear to be appropriate/job-relevant? However, we can't assume that a face-valid test is psychometrically robust and also need to consider other forms of validity.

Tests with high face validity ensure buy-in from candidates and line managers, but with face validity test choice is not based on hard evidence and is unlikely to be legally defensible if challenged. However, it may be the lack of face validity which ignites a legal challenge when candidates question the relevance of the questions they are being asked in relation to performing effectively on the job.

Faith Validity

An unfounded belief that a test is appropriate and effective; a feeling that the test works in the absence of evidence.

Faith validity can aid in getting buy-in to the use of objective assessment methods. However, lacking hard evidence of robust assessments can lead to misuse of tests and in the worst-case scenario could lead to the use of tests that are not legally defensible or valid, which don't allow for the selection of better candidates.

Formal

Consequential Validity

The intended and unintended consequences of using a test.

Test users should be mindful of how their use of assessment could impact assesses. For example, when using assessments to identify high potential there is the intended consequence of encouraging individuals to develop in relevant areas. An unintended consequence could be narrowing individuals' focus to just those areas being assessed rather than other relevant work areas.

Construct Validity

Can refer to a wide range of different sources of evidence demonstrating that a test measures an expected underlying construct, trait or theory, e.g. evidence that the test correlates with other similar measures.

Construct validity can be established by correlating a new test with existing tests designed to measure the same theoretical construct.

Criterion-related Validity

Refers to evidence that the test predicts relevant criteria (e.g. competencies or workplace outcomes).

Concurrent Validation

Studies involve the collection of test scores and job performance measures at the same time, typically with jobholders.

Predictive Validation

Studies involve collecting test scores from job applicants, waiting for a period of time while work experience is gained, and then the collection of job performance measures.

Content Validity

The extent to which the questions are actually focused on job-relevant content.

This is a more practical basis of choosing tests. Test choice is preceded by job analysis/role profiling activities and tests are chosen based on the degree to which they assess the key skills required for the job. The closer the test replicates the tasks required on the job, the better. If you have established content validity, then by definition you also have face validity. Content validity is likely to be legally defensible. All test users should aim to establish content validity when selecting aptitude assessments; tools like Job Profiler and card sorting with the Wave and Aptitude section cards can support this.

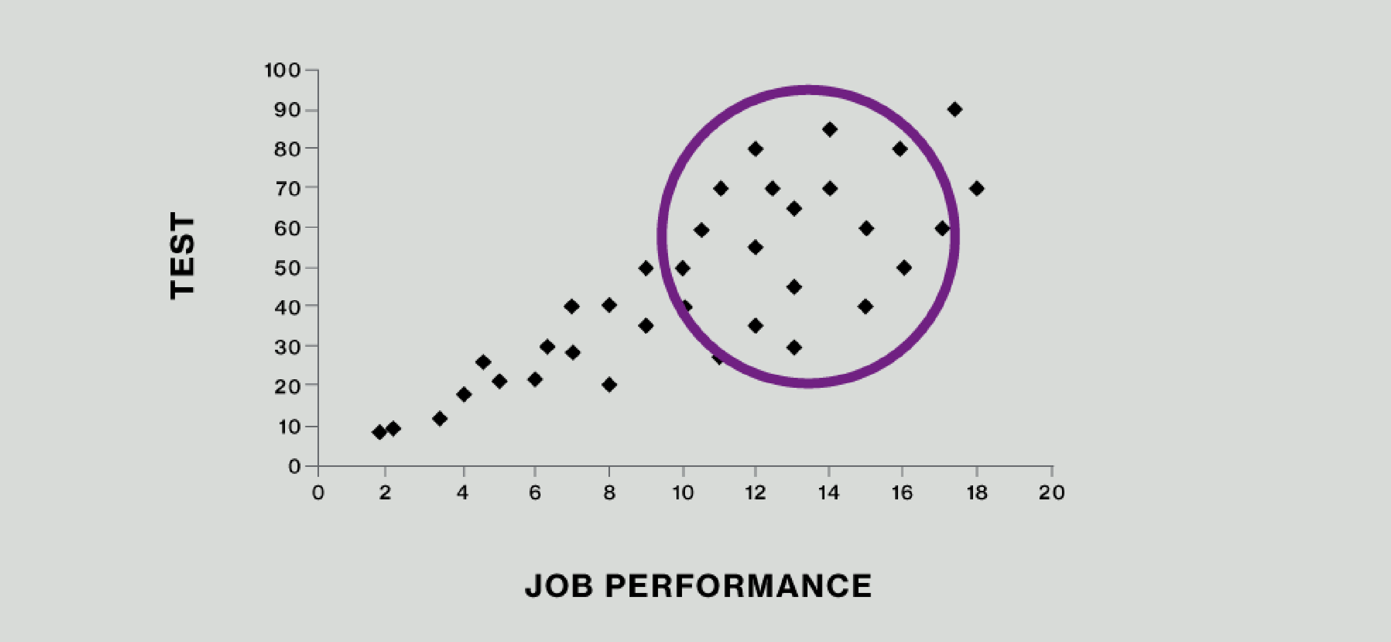
Key Problems in Validation

Whilst validation studies are important, conducting them is not without drawbacks. To ensure that you have a representative sample, it is recommended that you find at least 100 job holders for a concurrent validity study or 100 applicants for a predictive validation study.

The Criterion Problem

- It is often impossible to identify one measure which encompasses effective job performance
- Objective, quantifiable outcomes, such as sales performance or work sample tests, are preferred measures
- Performance ratings gathered from boss/self or 360 ratings can also be good measures of job performance but can be time consuming to organize and conduct
- In validation analysis, some unreliability in job criteria measures can be corrected using the correction for attenuation formula. This reduces the effects of error which can underestimate validity of assessments.

Restriction of Range



- When we conduct concurrent validation studies with jobholders, these individuals have already gone through a selection process so are likely to be performing at a higher level than a typical population. This means we are only considering a small part of the sample and the range of performance is therefore limited, which could underestimate validity correlations between assessments and job performance
- We must ensure that those taking part in validation studies are motivated and representative of populations likely to take these assessments in real selection and development settings
- Restriction of range can be corrected using a formula

Meta-Analysis and Validity Gereralization

Meta-analysis is a statistical process which combines many studies which have related research aims so as to form overarching conclusions; by increasing the sample size you minimize the chances of sampling error.

Meta-analysis has revolutionized our capability to make well-reasoned decisions based on a large number of research studies which often seem contradictory (i.e. give seemingly different results and come to seemingly different conclusions).

Validity generalization is the extrapolation of validity correlations established in research to other settings. When seeking to generalize the findings from one validity study to another situation, this should be done on the basis that the situation is similar to that of the study e.g. if the study demonstrated a numerical test was predictive of the performance of accountants, we would be more confident in applying this test to the recruitment of accountants.

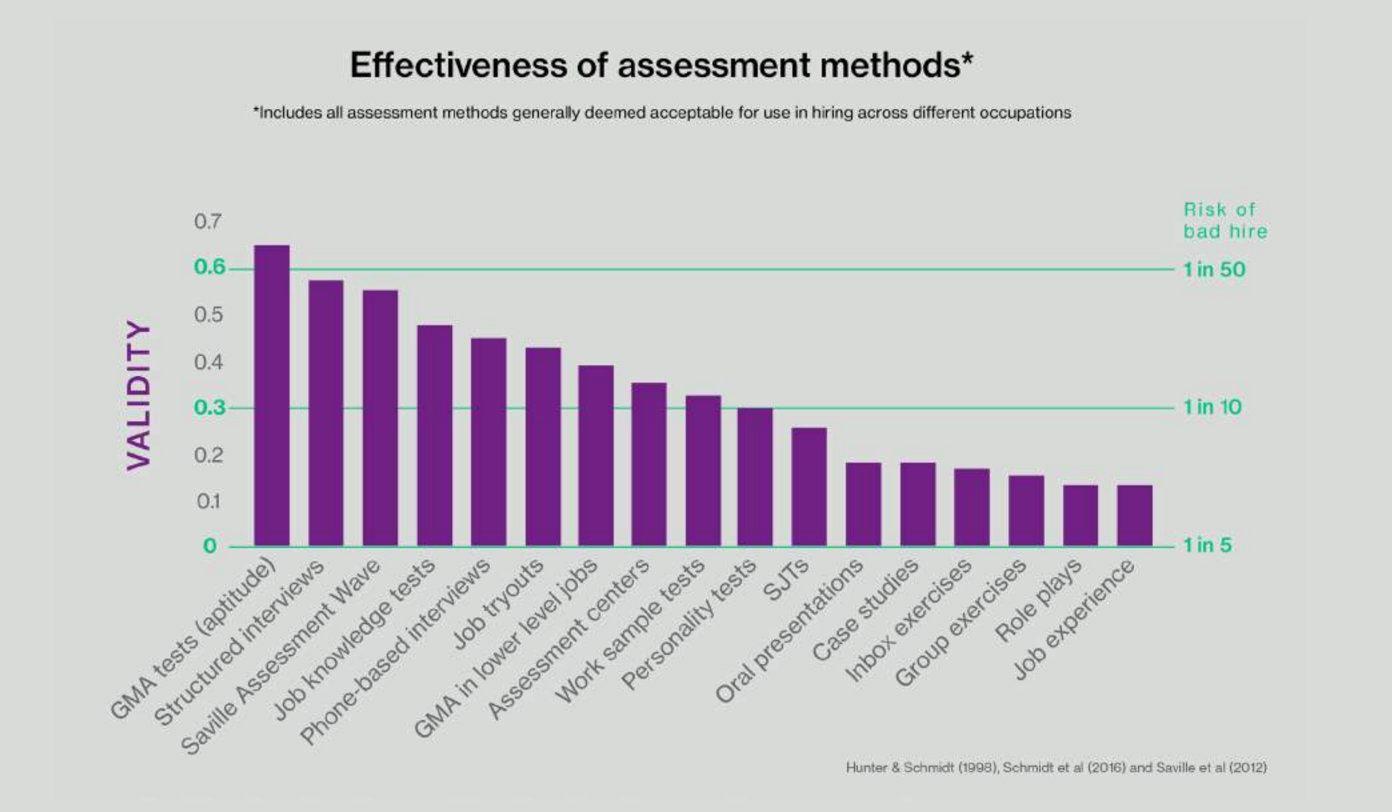
Alternatively, if the study is linked to similar criteria to what we are trying to predict, we would have more confidence in using the test e.g. commercial problem solving.

Example of Meta-analysis in Validity

As a guide, a validity score of 0.4 would be a 'strong' validity effect of these methodologies, which are quite difficult to correlate with job performance. As shown, job experience and role plays were not very predictive. Structured interviews were found to be a good predictor but we should take care when using less structured interviews as they can be less valid. Specific cognitive ability tests were the single highest predictor, above .6 validity.

When putting together a selection process you should use the most valid methods like aptitude assessments, structured interviews and Saville Assessment Wave.

- 1/5 – If you have a validity of 0 you have a 1 in 5 chance of hiring a poor performer
- 1/10 – If you have a validity of .3 you have a 1 in 10 chance of hiring a poor performer
- 1/50 – If you have a validity of .6 the risk of a poor hire is greatly reduced to 1 in 50



Effectiveness of assessment methods*

*Includes all assessment methods generally deemed acceptable for use in hiring across different occupations

Hunter & Schmidt (1998), Schmidt et al (2016) and Saville et al (2012)

Test Utility

Cost-benefit or test utility is concerned with answering the question: what is the pay-off from using tests? Ultimately this can often be viewed as the most convincing argument for using tests.

According to utility equations accounting for test cost versus the potential gains from testing.

Test utility is maximized when:

- The validity of a test is high
- There is high variability in job performance: i.e. some perform really well and some perform poorly
- The test score can be used to cut out high numbers of candidates: e.g. a cut off at the 70th percentile or above
- You select the highest performers

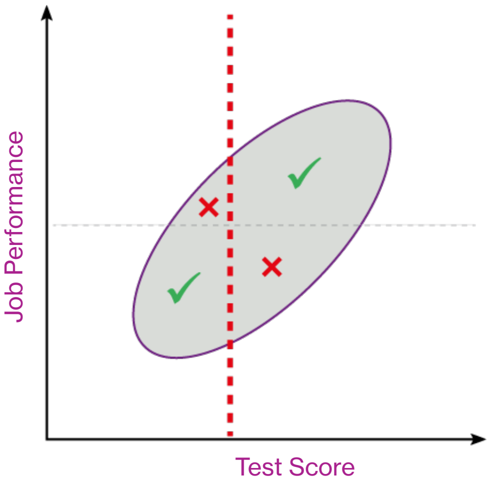
Implications for Utility

When would you use a low or high cut-off score?

Cut offs

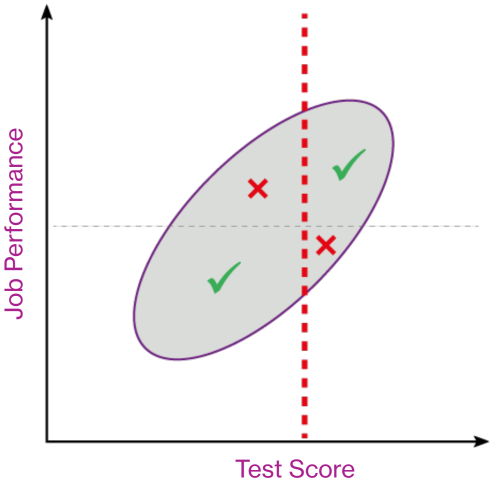
Low cut-offs

- Used when you have a similar number of applicants and jobs available
- Candidates who reach this minimum standard can be progressed, others are screened out
- The graph demonstrates how, in these instances, whilst you are likely to reject fewer people who will be successful, you may hire more poor performers
- This is likely to lead to lower overall test utility as the overall caliber of the workforce could be lower
- To combat this, hirers should seek to improve their ratio of candidates to roles available by running effective attraction campaigns, which will give them more choice about who they hire



High cut-offs

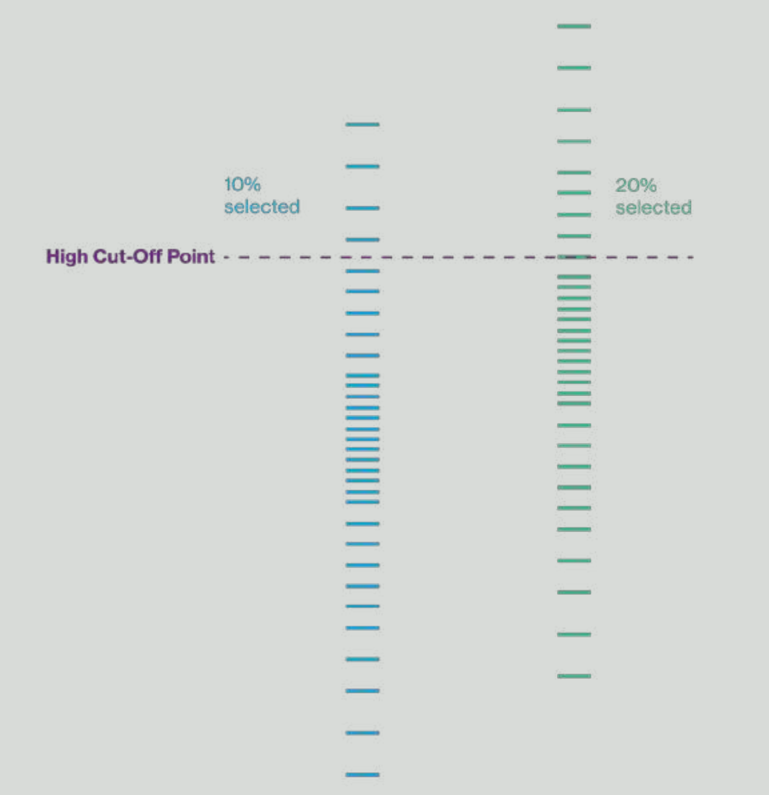
- Used when organizations can be highly selective because there are many applicants to few available jobs
- Higher cut-offs can give even greater return on investment but require more justification (i.e. evidence the job is difficult or that there is strong validity)
- Higher cut-offs can result in lower proportions of potentially disadvantaged groups being selected
- Particular care should be taken when using top-down selection (i.e. awarding positions to highest scoring candidates)
- There is an increased risk of rejecting people who would have succeeded if they were hired which is problematic for job applicants. It's increasingly important to consider this as we need to get the candidate experience right



The Four Fifths Rule

“Four-fifths” is a rough rule to use when controlling for adverse impact between groups being assessed.

- A selection rate for any protected group which is **less than four-fifths (or 80%)** of the rate for the group with the highest rate of selection is seen as **adverse impact**
- This affects where we need to set cut-off scores in assessment



Setting High Cut-Offs

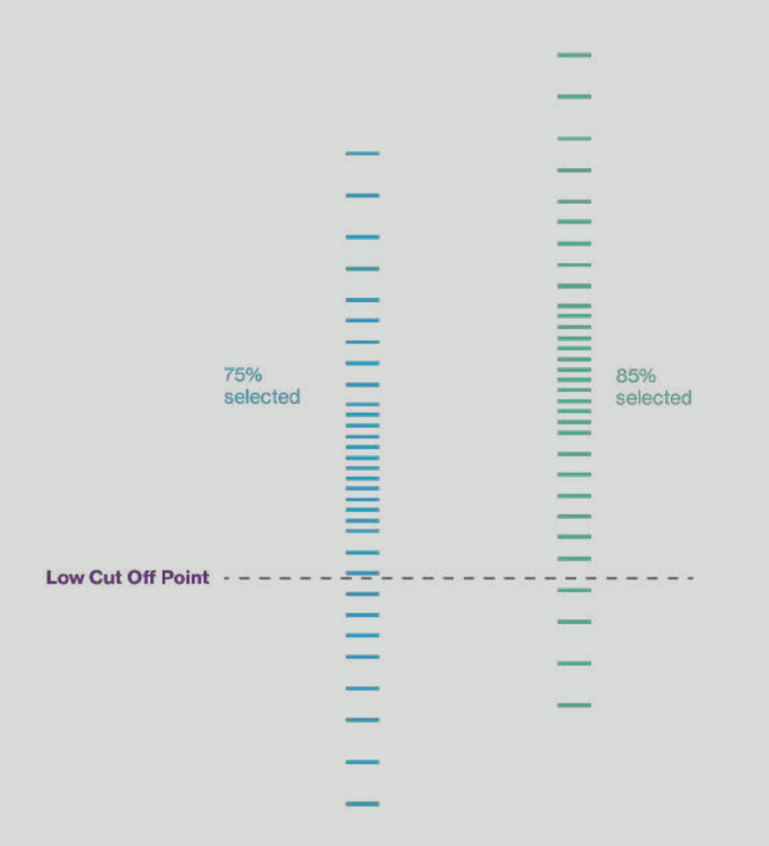
Blue group 50% of Green group:

Here, the selected Blue Group is 50% of the selected Green group. This creates an unacceptable adverse impact i.e. disproportionate. numbers of Green individuals are progressed compared to Blue individuals.

Setting Low Cut-Offs

Blue group 88% of Green group:

Acceptable Adverse Impact – more equal proportions of Blue and Green group individuals are progressed.



Summary

Using lower cut-offs can reduce the risk of adverse impact or unfairness in an assessment process. Nevertheless, it’s still important to bear in mind that every sample of people is different and any apparent evidence of adverse impact or unfairness may be because of sample-specific factors. It is therefore important to monitor the sizes of different groups progressed through each stage of assessment. Similarly, if there appears to be adverse impact in a data set, reducing the cut-off is likely to be a more effective solution. Saville Assessment monitors numbers of groups and models the potential adverse impact at different cut-off levels.

Best Practice and Ethics

Proper Data Management – GDPR

When using assessments, you need to follow these six principle of the General Data Protection Regulation (GDPR).

1. Processed lawfully, fairly and in a transparent manner. The scores should be used to make fair decisions about people. This requires the use of well chosen tests with appropriate interpretation. Ensure that candidates are provided with sufficient information about the assessment process.

2. Collected for specified, explicit and legitimate purposes and not further processed for an incompatible purpose. Ensure scores are only used for the purposes for which they were collected. To use them for other purposes requires gaining further permission from the candidate. If an assessment is completed as part of a development process it is unlikely it would be appropriate to use the results for selection or promotion decisions at another time.

3. Adequate, relevant and limited to what is necessary in relation to the purpose. Ensure only appropriate tests are used. Tests are not used unless the information is needed for a proper business purpose, e.g. making effective selection decisions, developing staff.

4. Accurate and, where necessary, kept up-to-date. Ensure that care is taken in collecting and processing data to ensure it is accurate.

5. Kept in a form which permits identification of data subjects for no longer than is necessary for the purpose. That there is a policy of deleting data once it is no longer useful. Typically test scores remain relevant for 12-24 months. After this they should be erased.

6. Processed in a manner that ensures appropriate security of the personal data. Appropriate security should be in place when storing data. Appropriate technical or organisational measures should be in place to protect against unauthorized or unlawful processing and against accidental loss, destruction or damage. Each organization should take their own legal advice with regard to their human resource activities. Saville Assessment is not in a position to advise on legal matters.

Equal Opportunities Legislation

Equal opportunities legislation has developed over time to protect more groups, with major legislative developments in the latter half of the 20th Century. This legislation has continued to strengthen and evolve to cover more protected groups.

For example, the UK Equality Act 2010 protects the following characteristics:

- age
- disability
- gender reassignment
- marriage and civil partnership
- pregnancy and maternity
- race
- religion or belief
- sex
- sexual orientation

Discrimination

Unfair treatment of any of these groups would be considered as discrimination. Discrimination may be Indirect or Direct.

Indirect Discrimination

Indirect Discrimination is the unintentional differential treatment or adverse impact that affects different groups as a result of the testing conditions imposed. Hiring managers should consider whether there is clear justification for their testing choice, for example, it would be indirect discrimination to ask one group of candidates to complete an English language test but not asking all of their candidates to do this.

- The unintentional differential treatment of candidates in different groups
- Testing decisions need to be justifiable if it could be claimed that indirect discrimination has occurred, for instance, the cut-score in a selection process negatively impacts a particular group but it is vital for selected candidates to have that level of performance in a given area
- Be sure to select tests that have minimal observed group differences

Direct Discrimination

Direct Discrimination treats people differently because of the group they belong to; this is almost universally outlawed and this is not something that any high-quality assessment is designed to do. An example of direct discrimination of assessment could be not allowing a person with a disability to complete a test as part of a selection process.

- The intentional differential treatment of people depending on a certain group they may be part of, such as gender, race or religion
- High-quality assessments are not designed to be used in this way

Fair Assessment

Respect for the Individual

It can be beneficial for the administrator to understand candidates' concerns or overall perspective of the experience. Ensure that candidates are fully briefed on the rationale and processes used to reach decisions and that you treat candidates with understanding.

Administrators need to deal with questions and problems in a patient and professional manner. Testing may be unfamiliar to candidates and they may be surprised by the formal nature of administration. There is evidence that candidates are more likely to regard decisions as fair when they are aware of the processes used to reach the decision.

- Administrators should treat candidates as they would like to be treated in the process
- Be sure that the candidates are aware of the process and why it is being used
- Demonstrate understanding of the nerves a candidate may experience

Choosing Appropriate Tests/Questionnaires

As we learned in Job Analysis and Assessment Choice, tests and questionnaires should be chosen on the basis of a thorough job analysis to ensure that decisions are being made on the basis of relevant information. To ensure assessment fairness, look for evidence of studies examining the appropriateness of the instrument with different groups.

- Assessment choice should be based upon thorough job analysis
- Consider reviewing technical summaries for evidence regarding the appropriateness of test use with different groups, e.g. validation studies

Preparing the Candidate

This is particularly important for aptitude tests. It is usually recommended to advise candidates how their data will be used, how they will be stored and whether they may be used again in the future. Candidates can access practise tests and guidance on the Saville Assessment website,and are included on the candidate dashboard.

Candidates should be briefed ahead of completing a psychometric test:

- The rationale for using the assessment
- What the assessment measure
- How their data will be stored and who will have access to their results
- Whether they require any reasonable adjustments
- Gaining informed consent from the candidate

Dealing with Language

For any psychometric measure you should consider what the impact of language needs are. Where English is not the primary language, consider whether it would be more appropriate to test in another language. Where an organization considers English to be the required language they may feel that testing candidates in English is justified. However, it is generally recommended that candidates are tested in their preferred language where possible. English proficiency assessments are available alongside aptitude assessments.

It is generally recommended to assess candidates in their language of greatest proficiency, wherever possible.

If you are not sure of the implications for testing, you can contact us.

Disability Considerations

Disability adjustments should be managed on a case by case basis, discuss any issues with candidates ahead of assessment to understand and accommodate their needs.

Some examples include:

- A candidate with dyslexia may have difficulty reading some assessment content and may need more time than other candidates to complete the task
- A candidate with sight impairment may have difficulty reading a booklet or seeing a computer screen; the candidate may need to use screen-reader software or have assistance from a sighted facilitator
- A candidate with a motor impairment may have difficulty using a mouse to fill in an answer sheet, so could instead use touch-screen functionality
- Manage candidate needs on a case-by-case basis
- Ask the candidate to provide what has been recommended for them by an appropriate specialist, e.g. An educational psychologist has recommended additional time for a person with dyslexia
- The general principle is that any adjustment should attempt to provide the individual with a comparable assessment experience to other candidates
- The assessor and assessee both have a responsibility for being as accommodating as is reasonable
- Saville Assessment online aptitude assessments are compatible with all modern computer and tablet browsers which permits a range of adjustments
- Where you are unsure of the appropriate reasonable adjustments to make, you should seek expert advice

Using Tests Responsibly

Interpreting Score

Care should always be taken to interpret an assessment correctly, being clear on what the different aptitude areas measure and what scores mean. You can use the assessment descriptions in the technical manuals to support you. Consider the appropriate scales to feedback to candidates, the most suitable comparison groups and whether any reasonable adjustments made have impacted test scores. Remember to take into account the size of error around their score and how they perform in comparison to the benchmark group.

- Make sure you know what the assessments you are using are measuring
- Be clear on how to interpret scores, their error of measurement and how best to give feedback on these to a candidate

Feedback

Candidates are likely to be interested in their results. Giving the option to have written or spoken feedback is recommended and in some regions, candidates have a legal right to access their results. This can help to increase candidates' self-awareness and better understand how their results have been used in the decision-making process. This is likely to make candidates feel more comfortable about the way in which their results are used in selection and development processes.

- Feedback may be a legal requirement based on the country in which the process takes place
- Feedback can help the candidate's self-awareness and understanding of the process

Test Use Policy

It is generally good practice for the use of tests to be guided by a test use policy. This will set out standards and local policies on a range of relevant issues. This helps ensure that minimum standards are maintained and that there is a consistency in practice across different assessment processes.

- Your organization should have and use a test use policy
- A test policy outlines the standards and requirements to be used consistently through your organization's testing processes
- A sample test use policy is available from us

Training and Responsibility of Test Users

It is important to complete training before using some assessments but, as with any skills or knowledge, over time parts may be forgotten and bad habits can develop. Equally, new developments may require updating of knowledge. Engaging with these developments to maintain up-to-date knowledge and develop skills means that you can continue making best use of assessments. It is the responsibility of the test administrator to ensure proper practice and ensure that all interpretations from the test are valid and appropriate to the context and for the person who is using the information.

- It is important to complete appropriate training ahead of using some assessments
- Test administrators should stay up to date with any new developments to ensure they are delivering best practice assessment use

Best Practice: Key Points

When using aptitude tests it is important to consider points of best practice:

Promote proper data management

- Ensure fair assessment
- Use valid tests
- Provide preparation materials to candidates
- Monitor group differences in the samples you progress at each assessment stage
- Consider candidates' needs
- Make accommodations for special requirements
- Review your testing policy

About Saville Assessment, a Willis Towers Watson Company

Our integrated approach to talent solutions helps organizations achieve their business objectives. We decrease risks and increase good opportunities associated with talent assessment and development. Representatives in over 80 countries equip us to support projects all over the world. Whether early careers recruitment or leadership development, local authority or multinational corporation, we help all our clients unlock potential and achieve results. Learn more at savilleassessment.com

savilleassessment.com

training