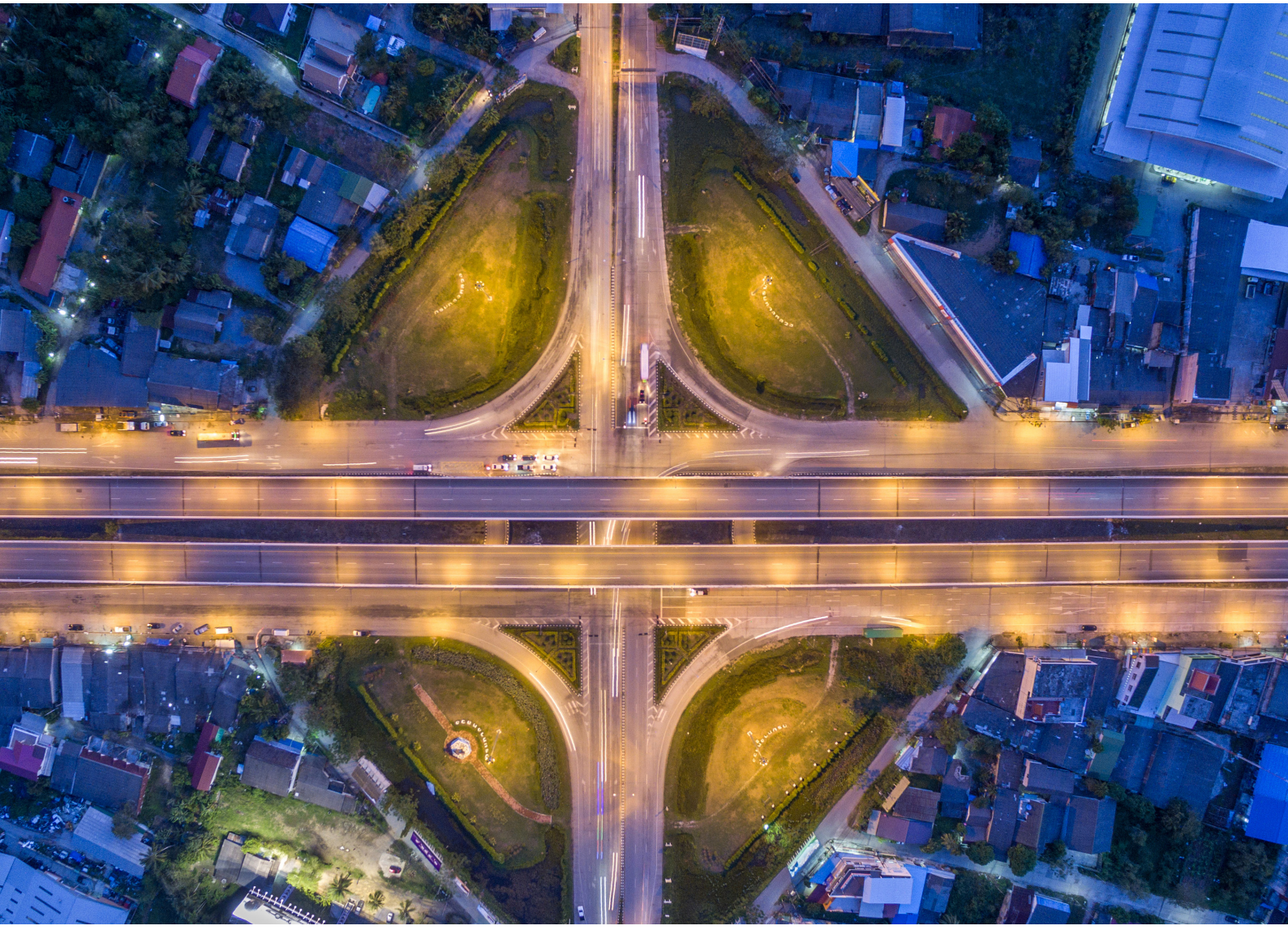
Saville Assessment

Willis Towers Watson



Swift Global Aptitude Technical Summary

Contents

1.0 Swift Global Aptitude Test Information

Key Information

- This combination test is an online measure of the cognitive ability that underpins the capacity to flexibly solve problems and apply learning in new environments at work
- The test consists of three sub-tests:
    - **Recall** involves memorizing figures and their position in a grid of cards
    - **Numbers** involves identifying patterns in a series of numbers
    - **Abstract** involves identifying logical sequences in a series of shapes and symbols
- A mobile-first test – designed from first principle for completion on mobile devices/phones
- Includes an interactive element – i.e., turning over cards in Recall
- Minimized language content to make more internationally applicable/scalable

Technical Information

- Technology supporting individual time limit for groups of four questions (testlets)
- Linear-on-the-fly (LOFT) testing
    - Fixed-length test
    - Based on Item Response Theory (IRT) methodology and scoring mechanism
    - Draws items of equivalent difficulty from a bank of items for different candidates
- Available for unsupervised used online (Invited Access, IA)
- Compatible with smartphones, tablets, laptops and desktop computers

Sub-test	No. of Testlets	Time Limit per Testlet	Total No. of Questions	Total Time Limit
Recall	**3**	1 minute (15sec + 45sec)	12	3 minutes
Numbers	**2**	2.5 minutes	8	5 minutes
Abstract	**2**	2 minutes	8	4 minutes
SGA Total			**28**	**12 minutes**

2.0 Norm Groups

The Swift Global Aptitude norm groups available at launch are as follows:

- International Graduates – All (2022, N=577)
- International Professionals & Managers (2022, N=410)
- International Individual Contributors (2022, N=365)
- International Mixed Occupational Group (2022, N=634)

Norm group descriptions and other support documentation for the above norms are available in the Client Area on the Saville Assessment website (www.savilleassessment.com).

We monitor our data regularly and seek to develop new regional and country specific norms where possible. For the latest norm availability information, please contact Saville Assessment.

3.0 Report

An example report for the test can be found in the Client Area on the Saville Assessment website.

4.0 Practice and Preparation

An online practice test is available for this test. It is designed to provide a realistic set of example questions to help familiarize the test taker with the format and style of the aptitude assessment questions, as well as additional information about the assessment process.

The online practice test also provides individual feedback on the responses given, featuring realistic time limits which replicate a real assessment scenario.

The aptitude practice and preparation materials can be found on the Saville Assessment website.

5.0 Theoretical Background

Cognitive abilities, intelligence and intellect are closely related but complex concepts that vary in their composition and definition. This was true, even near when the concepts were developed and operationalized, as reflected in the contrasting theories proposed by different psychologists since the early 20th Century. To this day, there is still debate around the degree to which we should be thinking of these as one unitary construct or 'g', or considering these as individual constructs, or even a combination where we have a unitary construct comprising multiple constructs.

Charles Spearman proposed the higher factor of general intelligence 'g' after finding those who performed well on one ability test tended to do better on other abilities too. According to this theory, under the 'g' factor lies multiple specific abilities such as verbal, numerical and spatial, and an individual who performs well on a verbal test would also score well on a numerical test.

In contrast, Louis L. Thurstone posited a multi-faceted view of intelligence that encompasses seven primary mental abilities, which are verbal comprehension, word fluency, number facility, spatial visualization, associative memory, perceptual speed and reasoning. Thurstone's theory places greater emphasis than Spearman's theory on the fact that people can have high mental ability in one area while being lower in other areas.

To a large extent, reconciling the theories of Spearman and Thurstone, Phillip Vernon proposed a hierarchical structure of intelligence with 'g' at the highest level then broken down into two factors 'academic' and 'practical'. The 'academic' factor refers to abilities including reading comprehension and arithmetic reasoning, whereas 'practical' was about mechanical and spatial abilities. These areas of ability can be split further into specific skills.

In the second half of the century, more theories of intelligence were proposed which held the view of a multi-facet construct. An example is Howard Gardner's Theory of Multiple Intelligences which suggests eight different intelligences and that some intelligences are best measured using specific skills assessments.

The Saville Assessment aptitude portfolio does not attempt to measure one single 'general mental ability', nor does it focus on an extremely broad range of ability measures, such as the Wechsler Adult Intelligence Scale measure of IQ.

Instead, we design tests of two types – 1) specifically for individual aptitude areas that are most relevant and applicable to a range of occupations and job levels in the modern world of work; 2) combination or Swift tests which combine aptitude areas that are relevant for particular occupations and occupational levels to allow for shorter completion times.

Apart from the debate on whether intelligence is a single construct or has different forms, there is also the 'nature versus nurture' debate in the field of cognitive ability. Intelligence was thought to be inherited in the theories by Spearman, Thurstone and Vernon etc., until Raymond Cattell proposed the concept of fluid and crystallized cognitive abilities. Fluid cognitive ability, defined as the ability to think flexibly and abstractly to deal with novel problems, is thought to be genetic and not susceptible to culture, prior learning, and experience. In comparison, crystallized cognitive ability was thought to be grounded in knowledge, expertise and wisdom learnt over time. Based on these definitions, aspects related to crystallized cognitive abilities are expected to correlate more strongly with educational variables than aspects related to fluid cognitive abilities.

Although existing tests in the Saville Assessment aptitude portfolio do not explicitly test prior knowledge in specific areas, they generally lean more towards the concept of crystallized cognitive abilities where a specific ability area could improve through learning and gaining experience in the area. For example, gaining more exposure to working with numerical data over time would be expected to help improve an individual's ability in numerical reasoning.

To meet our goal of developing a test that is broadly applicable and measures a universal and culture-free construct, which was recognized as a useful addition to our test portfolio, abilities closer to the concept of fluid cognitive abilities were identified as the target area for the new Swift Global Aptitude. Swift Global Aptitude was designed to measure the ability to identify patterns and structures, and the capacity to hold information for future use. These aptitude areas are essential in problem solving and learning at work, both of which are common to many occupations across different role levels.

The three components selected to be included in the test were Recall, Numbers and Abstract. Recall aims to measure the ability to memorize then recall information. Having the information available for use will be helpful in problem solving and completing tasks in the workplace. Being able to retain and then recall the newly-acquired information also helps to effectively apply learning in new environments at work.

Numbers and Abstract measure the ability to identify patterns and logical sequences presented in two different formats, in a series of numbers and diagrams respectively. They require the ability to think flexibly and abstractly, which is important when dealing with problems and making decisions in novel situations in the workplace.

An added advantage of measuring these forms of aptitude is that they require very little language, making them more applicable for multi-national or global assessment.

6.0 Development

Development Goal

The starting point for the development of any assessment must be a clear understanding of the criterion (outcome) it is designed to measure. The primary assessment goal of the Swift Global Aptitude is to measure aspects in fluid cognitive abilities, which are thought not to be influenced by prior learning, experience and culture, creating a test that has good global applicability.

Moreover, it was aimed to design the Swift Global Aptitude as a mobile-first test to make the test more accessible and inclusive, as the ownership of smartphones rapidly increases each year across the world including emerging economies (Sources: Statista; Pew Research Center). Therefore, the test was designed to contain a reduced amount of text and use mostly figures, symbols and numbers.

Development Considerations

The development of Swift Global Aptitude focused on the following features and characteristics:

High Face Validity, User-Friendliness and High-Quality Design – Users of tests should be impressed by the quality of the test materials, supporting documentation and feedback reports as the standard of these materials can influence their perception of the organization using the tests. For this same reason, test items should also appear relevant to workers. Swift Global Aptitude was designed to be modern, user-friendly on both personal computers and smartphones, and to provide a high-quality user experience.

Validity – The item types, quality and content of the tests were designed to focus on those characteristics which differentiate success in roles of the relevant level.

Fairness – Care was taken to include only materials suitable for test users regardless of their culture, country of origin, age, gender, ethnicity, sexual orientation or religious belief. This includes avoiding content which might favor one test-tasker over another (e.g., the choice of figures used in Recall).

Reliability – Swift Global Aptitude was designed to be a short, combination test yet with a good level of internal consistency. The test should consist of the right number of items that gives the optimal balance between length and reliability.

Security for Online Use – The test was designed purely for online use that could allow for the interactive elements in the Recall subtest. The test is powered by a bank of items, with different testlets of equivalent difficulty being given to different candidates for increased test security.

Appropriate Difficulty Level – Swift Global Aptitude was designed so that the majority of individuals could attempt all or most of the questions with the average number of correct answers tending to be around half or more of the items correct. This helped ensure that the test could differentiate effectively between the strongest and weakest candidates at the intended level of difficulty. If the average candidate answers about half of the questions correctly, this generally means the test is likely to have a good distribution of scores and suggests that the test was rigorous but not overly difficult. In addition, respondents are likely to feel positively about a test which is neither too difficult nor too easy.

Mixed Item Types – The test was designed to ensure that different content areas are sampled effectively, giving appropriate breadth of measurement. The items were written to sample different aspects of individual aptitude areas; this was achieved using different item sub-types within the test. The presence of different item sub-types also helps to mitigate against method effects which often arise as a result of overreliance on one particular item format.

Short Assessment – It was aimed to develop Swift Global Aptitude as a short, combination test that makes for a more efficient assessment process and a more effective use of candidates' time. To be able to use the test alongside another behavioral measure to create a comprehensive assessment was also a key consideration in the test development process.

Power and Speed – The test was based on individually timed testlets, each testlet consisting of four items. The time allowed was designed to ensure that the majority of candidates could complete the test and that the main differentiating factor for candidates was their accuracy in answering the questions, rather than their cognitive processing speed.

Item Writing, Review & Selection

Recall

Item Writing

As a new test format, the item writing phase for Recall included several stages.

1. The format of the test was first determined, which was to remember figures presented in a grid of nine cards. Several item formats were then created based on the ability to remember 1) the figure itself and 2) the location of the figure in the grid. A total of four item types were proposed, but one was discarded as it was thought to give a greater chance of guessing the correct answer. Items were then written according to the final three item types targeting at the recall of the location of the figure in the grid and/or what the figure(s) was.

2. Another important aspect in the test design was the type of figures to be used. It was decided to use figures that appeared to be modern and work relevant, rather than shapes and patterns which would be too complex to remember and similar to those used in the Abstract sub-test. Research has shown that the memory of objects is impacted by the familiarity and ability to name the object (Brodeur et al., 2010). Hence, a review was conducted with a group of nine internal staff members, including psychologists, to rate the familiarity of and 'easiness to name' for each of the 60 figures sourced from an online image database. An average combined score was then created for each figure based on the ratings on the two aspects, giving an indication of how easy the figures were to recall. Each figure was also assigned a category based on its genre, e.g., office, technology, etc. The figures were then evenly distributed into four different groups with the average combined scores and categories balanced in each group, to ensure the groups are equivalent in terms of the memorability of the figures within each group. This was an approach taken to enable variations of items to be created, i.e., different versions that used different sets of figures were created based on the same item template.

3. An internal pre-trial pilot was conducted with 16 staff members using paper cards to confirm that the format of the test (turning over cards), the item types, the choice of figures as well as the time limits were all workable.

4. Afterwards, further testlets were created. Each testlet was based on one square grid of nine cards followed by four items. Testlets had between five and nine figures in the grid with a variety of memorability scores, and a mixture of item types.

Item Review & Selection

The items written were then reviewed contextually by members in the R&D department in a mock-up format similar to how the figures will be presented on screen. As the test format has a clear right or wrong answer (whether a figure was present or not in the grid and where), no changes were required in the items. However, feedback from reviewers indicated that the original 30 seconds time limit for the 'memorizing' part of the task and 1 minute for answering the four items in a testlet were too lenient. Adjustments were then made to shorten both time limits to 15 and 45 seconds respectively.

The trial version for Recall was powered by a bank of items and four testlets (16 items in total) were drawn from the bank for each participant.

Numbers

Item Writing

Most items for the Numbers sub-test were initially written for a previous content trial. A few additional items were written to ensure there were sufficient items for the new trial. The item writers in both trials were psychologists, including experienced psychometricians. For each item written, several variations following a similar number pattern were also created to help increase the item pool size. The items were written with varying difficulty levels, intended from easy to difficult. Moreover, to create a varied bank of items, the items varied between an ascending or descending order of numbers and in terms of the position of the missing number, and some items involved negative numbers and/or decimals.

It was found during the layout design stage that items with eight or nine numbers in the sequence were too long to fit on smaller screen sizes, therefore longer sequences were shortened to have a maximum of seven numbers. In some cases, the answer options for the items had to be changed according to the updated number sequence.

Item Review & Selection

The review of the items in the pool was conducted by two psychologists, to ensure there was no ambiguity in the sequence of the numbers and each item had only one correct answer. After the first stage of review, using the data from the separate trial and feedback from the reviewers, as well as the type of sequence used, the items were sorted into testlets. Each testlet had four items with an increasing gradient of difficulty. Testlets also varied in terms of their overall difficulty levels, i.e., some testlets were overall more difficult (had a lower average percentage correct) than others. Several items were discarded as their difficulty level was thought to be lower than desired as indicated by the data collected from the previous trial. The testlets were then reviewed by a psychologist to ensure the structure of the testlets met the requirements.

The trial version for Numbers was powered by a bank of items and four testlets (16 items in total) were drawn from the bank for each participant.

Abstract

Item Writing

Items for the Abstract sub-test were adapted from the existing item bank for Abstract Reasoning Aptitude. Items that were deemed suitable for presenting on smaller screen sizes (i.e., on mobile phones) were first selected, excluding items that used small elements in the series of figures which would make the pattern difficult to decipher on a small screen. As with the Numbers sub-test, it was found that the Abstract items should only have a maximum of seven figures in the sequence to suit smaller screen sizes, therefore any items selected that had more than seven figures were shortened.

Furthermore, some pattern elements in the series used in the existing abstract reasoning tests would appear relatively small on a mobile phone, therefore a new mobile-first format was designed to separate those elements from the main shape in the series allow them to be displayed more clearly.

Item Review & Selection

The first stage of the item review was conducted by two individuals in the R&D department to ensure there was no ambiguity in the sequence, especially in the revised (shortened) items. Item exclusions were made following this review stage. The selected items were then organized into testlets based on existing item parameters used in Abstract Reasoning Aptitude. Similar to the construction of Numbers testlets, each Abstract testlet consisted of four items with an increasing difficulty gradient. Some testlets, by design, are more difficult than others, indicated by the testlet average of the item difficulty parameters.

The testlets were then created in the new format as described earlier (separating small pattern elements from the main shape) and reviewed contextually by two other individuals in the department. A few changes and further item exclusions were made at this stage.

The trial version for Abstract was powered by a bank of items and four testlets (16 items in total) were drawn from the bank for each participant.

Trialling

Content item trials took place between November 2020 and November 2021, during which 1,755 participants completed the first version of the test online. The trial version of the test was 22 minutes in length and had 48 items. Participants (including professionals, managers, graduates and A-Level students) were recruited via internal staff members and external clients.

Data Analysis

The data obtained from the content trialling phase were analyzed separately for each of the three measures based on item responses, item partials and IRT item parameters. The best items and content were selected for the final version based on their psychometric properties, i.e., the items which had the strongest partial correlations, appropriate range of percentage correct, good level of item discrimination etc., in order to produce reliable tests at the required level of difficulty.

The final test was cut down to 28 items in total with a time limit of 12 minutes.

Development of Norms

The norm groups for the final test were sampled from the participants in the content trial. The biographical data that the participants reported as part of the data collection were used to classify participants into the appropriate norm groups.

The methodology for creating specific norm groups is described below:

a. **International Norms**

The international norms are created based on the participants' geographical location. They include various countries around the world with each country accounting for no more than 25% of the group.

b. **Graduates Norm**

The participants' self-report of their highest qualifications was used to create the Graduates norm. Only individuals who reported having a first/bachelor's degree, a master's degree or a PhD/Doctorate were included.

c. **Professionals & Managers and Individual Contributors Norms**

These norms were developed based on participants' self-report of their level of management responsibility. The Professionals & Managers norm consists of individuals who reported themselves as managers, team leaders, supervisors, or who were professional individual contributors, as well as a small portion of senior managers and executives. The Individual Contributors norm consists of individuals who were professionals or non-professional individual contributors, or those who indicated 'not applicable' in relation to their management responsibility.

d. **Mixed Occupational Norm**

This norm group included participants with all levels of management responsibility, with those that reported themselves as managers taking up no more than 50% of the group.

7.0 Languages

Swift Global Aptitude is designed to contain a minimal amount of text in the test instructions and within the test, therefore less text will be required for translations. For the latest language availability information, please contact Saville Assessment.

8.0 Reliability

The internal consistency figure presented here for Swift Global Aptitude is a Separation Index. This method produces similar figures to Cronbach's Alpha (Andrich, 1982) and allows for an internal consistency calculation to be made in item-banked tests, rather than fixed-form tests.

As a Swift combination test, it is worth noting that the greatest level of reliability is found at the total score level, which is designed to be the decision-making score. The sub-test scores provide additional test-taking information, but we would not recommend that these are used in isolation for decision making.

The mean percentage correct figure broadly reflects the design aim of giving a positive candidate experience where many candidates answer above 50% of the questions correctly.

The large standard deviation value seen in the table reflects the ability of the items to differentiate performance through a wide score range. This is required to give an accurate representation of test-takers' ability.

Swift Global Aptitude Internal Consistency Reliability (N=1740)

	Mean % Correct*	SD (%)*	SEm Sten	SEm 'T'	r**
Total	66.03	15.33	1.00	5.02	**.75**

**Percentage correct based on the longer Swift Global Aptitude trial version with 48 items in total.*

***Based on the internal consistency reliability for the longer Swift Global Aptitude trial version and corrected for length with the Spearman-Brown Prophecy Formula.*

9.0 Validity

Criterion-Related Validity

We are seeking new criterion-related validation studies as part of our ongoing research for this new aptitude test. Please contact Saville Assessment for further information.

Construct Validity

The scores from the three sub-tests in Swift Global Aptitude were intercorrelated, as shown in the table below.

Intercorrelations of the three sub-tests in Swift Global Aptitude (N=1740)

	Mean Score	SD	Total	Recall	Numbers	Abstract
Total	.00	.64		.69	.78	.80
Recall	.00	.82			.27	.32
Numbers	.00	.85				.49
Abstract	.00	.85				

The sub-tests within Swift Global Aptitude do not correlate very strongly with each other, indicating that they measure separate constructs. The highest correlation came from Numbers and Abstract (.49). Both sub-tests involve identifying patterns in a series but require somewhat different reasoning skills to do so. As expected, Recall correlates less with the other two sub-tests as it is less closely related to the identification of patterns.

This analysis is based on the same trial sample used to calibrate the item parameters, hence the average aptitude (theta) scores of the group are all around the mid-point (0) of the latent ability scale.

10.0 Fairness

Gender Group Differences

Gender group differences on Swift Global Aptitude Total Score and three sub-tests

	Male (N=518)		Female (N=722)		Pooled SD Difference
	Mean	SD	Mean	SD	
Total	.08	.64	-.07	.60	.24
Recall	.04	.82	-.02	.79	.07
Numbers	.20	.84	-.17	.80	.45
Abstract	-.01	.87	-.03	.81	.03

The table above presents the gender group differences on the Total Score and the three sub-tests in Swift Global Aptitude.

Expressed in terms of raw theta (ability) scores, there was a small difference (.24 of a standard deviation) between males and female on the Total Score, where males overall slightly outperformed females on the test. This was largely driven by the difference found between the two gender groups on the Numbers sub-test with males scoring higher than females in general (.45 of an SD). However, there was no notable gender group difference on the Recall or Abstract sub-tests. This helped reduce the advantage for males on the Numbers sub-test resulting in a much smaller gender group difference on the Total Score.

Age, Ethnic and Other Group Differences

In our trial sample, there were insufficient participants from some of the demographic subgroups (subgroup N<500), therefore we currently are unable to present both age group and ethnic group differences on the test.

Apart from age group and ethnic group differences, we are also looking to analyze whether educational attainment level is associated less with performance on Swift Global Aptitude given that it is a test measuring ability closer to the concept of fluid cognitive abilities.

We continue to collect biographical data from test usage and will publish the data as the subgroup sizes are sufficient to provide stable estimates of these group differences.

Group Differences Summary

The performance differences reported here are at the group level, rather than being reflective of specific individuals. In all cases, the average group-levels of performance represent largely overlapping performance distributions, with greater variation in performance within any group than between groups. Based on these average group-level data, it is inaccurate and inappropriate to make any predictions or decisions about any given individual's performance as a result of their membership of a particular demographic group.

It is also important to bear in mind that each sample of individuals is different and group differences should not be generalized beyond these specifically-reported samples in an excessively broad manner.

As a measure of cognitive ability, Swift Global Aptitude will occasionally reveal small to moderate differences between groups. To ensure that any group differences shown are meaningful, relevant and fair, it is important to make sure that the use of such tests can be justified. This is especially true when using the test in selection with a cut-off score. Justifying the use of any test involves making sure that the skills being assessed by the test are relevant and valid and that the level of any cut-off applied is demonstrably appropriate. The use of job analysis and, where possible, local validation studies is particularly important for demonstrating the link between a test and the job it is being used to select for.

In particular, the use of high cut-offs (e.g., above the 50th percentile) may require additional justification and analysis to ensure that this does not lead to adverse impact against any group. A further precaution is to use a behavioral measure, e.g., Work Strengths or Match 6.5, alongside aptitude to create a weighted overall fit score which can be expected to mitigate against the potential for adverse impact.

It is one thing for an assessment to be designed to be fair and valid, and another for it to be used fairly. The clearer and more consistent the structure and process presented for aligning tests to a job and agreeing consistent criteria for decision making based on the tests, the less likely it is that the assessments will be unfairly applied by using different standards for candidates in different groups.

In general, the differences between gender groups are small and we do not therefore advise that specific differences in profile interpretation should be warranted when considering test results from different groups defined according to these variables. We do not, unless local legal frameworks permit or mandate such an approach, recommend using separate norms for gender groups. For further information, please contact Saville Assessment directly.

11.0 Appendix 1: Internal Consistency Reliabilities for Sub-Tests in Swift Global Aptitude

The following table shows the internal consistency reliability coefficients for the sub-tests in Swift Global Aptitude.

Swift Global Aptitude Internal Consistency Reliability (N=1740)

Sub-Test	Mean % Correct*	SD (%)*	SEm Sten	SEm 'T'	r**
Recall	70.57	19.65	1.28	6.39	.59
Numbers	60.79	20.45	1.33	6.66	.56
Abstract	66.54	20.78	1.35	6.76	.54

**Percentage correct based on the longer Swift Global Aptitude trial version with 16 items in each sub-test.*

***Based on the internal consistency reliability for the longer Swift Global Aptitude trial version and corrected for length with the Spearman-Brown Prophecy Formula.*

12.0 References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95-104.

Brodeur M.B., Dionne-Dostie E., Montreuil T., & Lepage M. (2010). The Bank of Standardized Stimuli (BOSS), a New Set of 480 Normative Photos of Objects to Be Used as Visual Stimuli in Cognitive Research. *PLoS ONE* 5(5): e10773. doi:10.1371/journal.pone.0010773

Cattell, R.B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22.

Gardner, H. (1983). Frames of Mind: The Theory of Multiple Intelligences. New York: Basic Books.

Spearman, C.E. (1904). General Intelligence, Objectively Determined and Measured. *American Journal of Psychology, 15*, 201-292.

Thurstone, L.L. (1946). Theories of Intelligence. *The Scientific Monthly*, 62(2), 101–112.

Vernon, P.E. (1964). The structure of human abilities. London: Methuen.

Websites:

Statista: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world#sources

Pew Research Center, February 2019, "Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally"

Saville Assessment

CI Tower
1st Floor
St George's Square
New Malden
KT3 4HG
United Kingdom

Tel +44(0)20 8619 9000

info@savilleassessment.com

Saville Assessment

Willis Towers Watson