

TECHNICAL MANUAL & USER GUIDE

GENERAL ANALYTICAL APTITUDE TEST (GAAT)

This manual has been created by the psychologists from the Research and Development department of Central Test.

*The assessment **GAAT (GENERAL ANALYTICAL APTITUDE TEST)** and the Central Test logo are registered trademarks.*

No reproduction, modification or adaptation is permitted without the written consent of the publisher Central Test.

Table of Contents

INTRODUCTION	3
1. GENERAL DESCRIPTION AND APPLICABILITY	4
1.1. Factor Descriptions	4
1.2. Main Areas of Application	6
1.3. Administration and Questionnaire Structure	6
1.4. Technical and Security Issues	7
2. SCORING	8
2.1. The Relationship between Raw Scores and Norms	8
2.2. Transforming Raw Scores to Standardised Scores	8
2.3. Link between Scores and Report Comments	9
3. PSYCHOMETRIC PROPERTIES	10
3.1. Sample Descriptives	10
3.2. Norms	11
3.3. Reliability	17
3.4. Validity	17
3.5. GAAT and Equal Opportunities	19
4. ADDITIONAL INFORMATION AND CONSIDERATIONS	23
REFERENCES	24

Introduction

How do individuals think? What mechanisms do individuals use to solve problems, and can these mechanisms be learned? Although many aspects of human cognition are still a mystery, psychologists have begun to flesh out the strategies used to think in organized ways to analyse and solve problems. This systematic style of thinking is generally referred to as “critical thinking”.

Critical Thinking has been studied since at least the 1910s when John Dewey first published his landmark book, “How We Think” (1910/1933), and it is included in many models of skills that are important for workforce success (Trilling & Fadel, 2009). Reasoning, logic, and judgment are all widely understood to be useful cognitive skills in the workplace, and as these are important components of critical thinking, it becomes clear why the inclusion, as well as the development of a critical thinking assessment, the **General Analytical Aptitude Test (GAAT)**, was necessary to aid organisations to assess and recruit the right individual.

1. General Description and Applicability

Critical Thinking is defined as the ability to judge a complex situation or an argument reflectively and purposefully. This judgment is usually based on clearly identified data and allows an individual to reach a logical conclusion, to decide what to believe or do. In any context, the use of critical thinking implies asking questions such as: What is the core of the problem? What approach should I use to solve it? Will this data enable me well, to reach a conclusion? Is this argument not in opposition to the other? If this is true, what does that entail? And so on.

Therefore, **GAAT** is designed to assess these critical thinking skills of an individual through four crucial factors – *Evaluating Arguments, Critical Analysis, Deductive Reasoning* and *Inductive Reasoning*.

Some benefits of the **GAAT** that most distinguish it from other similar tools in the market are the following:

- Summary of the individual's overall performance through a Critical Thinking Quotient
- Overall score on a scale of zero to twenty, including the number of correct responses by the individual
- Personalised interpretation of the individual's overall score
- In-depth description of Critical Thinking, its history as well as its measurement in **GAAT**
- A graphical representation of scores on each of the four factors – *Evaluating Arguments, Critical Analysis, Deductive Reasoning* and *Inductive Reasoning* – with reference to the population
- Detailed comments on each factor's score
- Test profile with appropriate solutions on incorrect responses

1.1. Factor Descriptions

GAAT is constructed in such a way as to measure the process of analysing, synthesising, conceptualising, applying, and/or evaluating information from various sources via four important components - *Evaluating Arguments, Critical Analysis, Deductive Reasoning* and *Inductive Reasoning*.

Evaluating Arguments – Evaluating Arguments refers to the ability to understand, formulate and deconstruct arguments, distinguishing between important and superficial information.

When trying to convince others, multiple arguments to support our viewpoints are used. An effective argument should compel others towards our conclusions based on an agreement with the information we provide. Knowing the order in which to present arguments skillfully can make our speech more convincing. During a debate, it is vital to

be able to respond to your opponent by effectively deconstructing their argument by recognising and articulating the weaknesses of their assertions.

These skills are not only important in a variety of occupations, such as law or sales, but are also useful in daily life, for instance, when leading a negotiation, and when collaborating and discussing topics or opinions with others.

Critical Analysis – Critical Analysis refers to the ability to understand and interpret information objectively and systematically in order to draw reasoned conclusions.

It is important to use this skill to analyse the logical connections between ideas and actively appraise them in order to make a rational judgement on their value. This allows us to make decisions about our actions based on a variety of options.

Critical Analysis is a skill that is used regularly, for instance, in conversation with others, when reading information and judging its reliability, and when problem-solving and reflecting on our own actions in our daily lives.

Deductive Reasoning – Deductive Reasoning refers to the ability to logically draw certain conclusions from laws in multiple premises of information. This reasoning is commonly applied using syllogisms, or 'if-then' statements. If $A = B$ and $B = C$, then deductive reasoning enables us to logically conclude that $A = C$.

If deductive arguments are created accurately, deductive reasoning allows us to draw conclusions that must be true. For this, an argument must be both deductively valid (the conclusion is true as long as the premises are true), and sound (the content of the premises are true themselves).

This type of reasoning is used in many occupational fields, such as finding practical applications of biological laws for the production of medicine, and in marketing strategies and human resources practices.

Inductive Reasoning – Inductive Reasoning refers to the ability to make broader inferences and generalisations based on specific premises of information. For example, as all the swans that I've seen have been white, I generalise that all swans are white.

This type of reasoning is also used when decisions are made using experiential information. By determining the links within presented data, using knowledge obtained from previous experiences, general conclusions are reached about this information using inductive reasoning.

Inductive Reasoning is used frequently in daily and working life, commonly in most situations that involve prediction or forecasting, using existing knowledge to inform decisions or information. This type of reasoning is key in science and research, where

scientists gather data, experiment, and make hypotheses for further research.

1.2. Main Areas of Application

Several studies, including that of Schmidt and Hunter (1996), show that general mental ability (intelligence) is one of the main predictors of job performance. Indeed, research affirms this ability to understand, and to do so quickly, various intricacies as well as the ability to continually adapt and improve one's expertise. As a consequence, in the professional field, the measurement of critical thinking is imperative as it is an indispensable quality to make the right decisions quickly.

GAAT assesses the individual's abilities by placing them in situations close to reality. The precise comprehension, as well as analyses of the questions in the assessment, reflects how they perceive information in real situations. Therefore, being a direct evaluation of the ability to understand and make decisions.

Hence, **GAAT** provides a norm-based objective assessment of the individual's level of aptitude in critical thinking as used in business settings. This test is recommended for the recruitment of employees whose functions involve the use of critical thinking and decision-making. **GAAT** is equally applicable for the recruitment and pre-training of individuals in an administrative, executive and/or managerial position as well as for students as part of entrance examinations or in organisations, as a prelude to personal development programs.

GAAT is specially reserved for professionals in the Human Resource departments of organisations, schools and colleges, as well as trainers.

1.3. Administration and Questionnaire Structure

Remote Administration of GAAT

GAAT can be administered under non-supervised settings; however, appropriate care should be taken by the test administrator to minimise the possibility of getting external aid on the part of the test-taker.

Format of GAAT

- Number of Questions: 30
- Assessment Format: Multiple Choice Questionnaire
- Duration: 40 minutes (Timed)
- Administration: Online

1.4. Technical and Security Issues

The following points serve as guidelines for securely administering **GAAT**. These are useful for ensuring that **GAAT** is administered objectively and fairly. Security specifications are also important for protecting the candidate's information, both during administration and after the completion of the assessment.

Respecting the Ethical Charter

As a publisher of online assessment tools, *Central Test* respects the legal conventions of French labour law and the regulatory provisions resulting from the law of 6 January 1978 on information technology, files and freedoms: Declaration N°1834067. Additionally, as of 25th May 2018, *Central Test* complies with the GDPR (General Data Protection Regulation) regulations made by the European Parliament & Council.

Test results and personal data cannot be shared with third parties without explicit consent from the candidate. Candidates have the right to access their results and the assessor must grant them access when requested. Furthermore, if the candidate requests it, all data must be permanently deleted; to do so, the candidate must contact either the organisation or *Central Test* directly.

Confidentiality and Security

Central Test guarantees the security and confidentiality of data generated during the administration of online tests: access codes are encrypted in the database and cannot be accessed by anyone. Passwords are computationally generated and not visible to companies. A URL rewriting module (secure link generator) has been installed and has transparent parameter passing for the user.

Browsing and data exchanged via the *Central Test* application can be secured and represented by an icon in the form of a padlock in the internet browser. The *Central Test* application has a "GeoTrust SSL TRUE BusinessID" digital signature, issued and certified by GeoTrust®. This certificate protects all *Central Test* assessments from phishing attacks and serves as proof of our dedication to providing a secure platform. Invitations to take tests are made via secure links that are of limited use (in terms of numbers and duration). The tests are timed and IP addresses are tracked.

2. Scoring

A raw score is the total number of responses endorsed by a candidate for each factor measured. In the context of a work-based assessment, raw scores mean little as they do not provide effective or appropriate comparisons between candidates across a specific interest. Norms enable HR practitioners, managers and recruiters to evaluate an individual's score relative to the score of other delegates who completed the same assessment.

2.1. The Relationship between Raw Scores and Norms

The best norm group is one that is most representative of the characteristics of the individuals taking a test. Applied assessment assumes personality traits, abilities, interests, motivations, values, attitudes and perceptions are distributed along a normal probability or bell-shaped curve. Therefore, norm groups effectively identify individual scores across this distribution. In any given population, levels of a target attribute may be found. Therefore, for a given attribute or trait, the normal curve represents the norm or average performance of a population and the scores that are above and below the average for that population (Kaplan & Saccuzzo, 2008). For norms to be meaningful, the sample used must reflect the intended population of test takers. If a test is developed for engineers, the norm sample must consist of engineers in the population. Similarly, if an assessment was developed for nurses, it is not appropriate for use in the selection of chemists. Norm groups can range in size from a few hundred to a hundred thousand. However, the larger the norm group sample size, the closer the approximation to a normal population distribution.

Key to the creation of norms is the conversion of raw scores into derived scores. This process provides a system of common metrics that readily facilitate the interpretation of test scores. For GAAT, the forms of score standardisation used were: 1) Mean; and 2) Standard deviation. Table 1 describes the various types of derived scores used in the validation of GAAT.

2.2. Transforming Raw Scores to Standardised Scores

For the 30 items in **GAAT**, the minimum and maximum overall score for each factor measured is 0 and 10, respectively. The conversion of raw scores to standardised scores is achieved using a three-step process, as explained below:

Step 1. Raw scores for each attribute are calculated for each scale by summing all responses for each attribute. For each response, the possible scores are 0 and 1, depending on which option on the Likert scale was endorsed.

Step 2. Scores are repositioned on a scale from 0 to 10 by using the following formula:

$$Score = \left(\frac{raw\ score - (min)}{(max) - (min)} \right) * 10,$$

whereby max refers to the maximum possible score of each scale and min refers to the minimum possible score of each scale.

Step 3. To enable comparisons between individuals, scores for each scale are standardised using the mean and standard deviation across respondents. This process assesses what percentage of the norm group the respondent's raw score was higher than and standardises the respondent's score to reflect this.

2.3. Link between Scores and Report Comments

Comments on **GAAT** reports reflect the scoring process. An individual receives a score of 0 to 10 (as demonstrated above). The comments provide an extensive interpretation of all the factors that are within the four dimensions of the test. They are chosen from six levels according to the score of the candidate, and detail individual attributes according to the individual's score. Higher scores are accompanied by stronger descriptors, weaker scores are accompanied by weaker descriptors, and so on. For example, a candidate with a low score on a factor may be described as 'not appearing to have a tendency' in the relevant factor. A high score on the other hand, may describe an individual as having 'a strong tendency' in the relevant factor. Scores should be interpreted in relative terms as the order of the targeted profiles is more important than their absolute score.

3. Psychometric Properties

At *Central Test*, assessments are continuously subjected to revision to ensure applicability as well as efficiency, and such is also the case with **GAAT**. To ensure that the assessment continues to meet the criteria for reliability and validity established by internationally recognised psychological institutions, analyses were conducted with a range of male and female professionals across the globe. The sample comprised 1832 English-speaking professionals across 99 countries to establish the questionnaire as an international measure of critical thinking ability as well as to establish that the items are accurate, relevant, clear, and concise and that they reflected the targeted construct.

3.1. Sample Descriptives

Table 1. Descriptives based on Gender.

Gender	Percentage (%)	N
Female	40.91	749
Male	51.99	952

Table 2. Descriptives based on Age Group.

Age Group	Percentage (%)	N
18 and below	8.08	148
18-25	17.80	326
26-33	46.75	856
34-41	15.46	283
42-49	7.37	135
52-56	2.24	41
57 and above	0.98	18
Not Defined	1.31	24

Table 3. Descriptives based on Study Level.

Study Level	Percentage (%)	N
Level 1: Master's Degree, PhD, Engineering or Business School, MBA	30.42	557
Level 2: Bachelor's Degree, First-year Masters	48.06	880
Level 3: 2 years of Vocational or University Studies	3.28	60
Level 4: High School Diploma	1.86	34
Level 5: Trades Training, No High School Diploma	1.31	24
Level 6: Students	1.42	26
Not Defined	13.65	250

3.2. Norms

Norms help us to compare an individual's score on an assessment with the scores of others who complete the same assessment, usually under similar conditions. These individuals are representative of the population for which an assessment is intended, and are therefore referred to as the norm group. The closer the match between the characteristics of the sample and the target population, the more accurate the distribution will be as a ranking guide. With norms, test developers and psychologists are able to establish comparisons between individuals and make more objective decisions. Norm groups enable us to attribute meaning to test scores and quantify an individual's performance, level of an attribute or trait relative to his/her peers.

Tables 4 to 11 detail the Mean, Standard Deviation (SD), Median, Minimum (Min), Maximum (Max), Range and Standard Estimate (SE) scores for the English-speaking sample on the factors as well as Critical Thinking Quotient for **GAAT**. These tables are divided into categories based on gender, age group and study level.

Table 4. Critical Thinking Quotient Norms for the Whole Sample.

Quotient	Mean	SD	Median	Min	Max	Range	SE
Critical Thinking	102.93	14.05	102.12	74.16	141.74	67.58	0.33

Table 5. Factor Score Norms for the Whole Sample.

Factor	Mean	SD	Median	Min	Max	Range	SE
Evaluating Arguments	5.55	1.89	5.87	2.15	10.00	7.85	0.04
Critical Analysis	5.71	1.89	5.81	3.05	10.00	6.95	0.04
Deductive Reasoning	5.92	1.92	6.36	2.75	10.00	7.25	0.04
Inductive Reasoning	6.20	1.91	6.35	3.10	10.00	6.90	0.04

Table 6. Critical Thinking Quotient Norms based on Gender.

Gender	Mean	SD	Median	Min	Max	Range	SE
Females	100.86	13.65	100.23	74.16	141.74	67.58	0.50
Males	102.83	13.75	101.93	74.16	141.74	67.58	0.45

Table 7. Factor Score Norms based on Gender.

Factor	Mean	SD	Median	Min	Max	Range	SE
Females							
Evaluating Arguments	5.33	1.84	4.66	2.15	10.00	7.85	0.07
Critical Analysis	5.40	1.83	5.81	3.05	10.00	6.95	0.07
Deductive Reasoning	5.74	1.91	5.12	2.75	10.00	7.25	0.07
Inductive Reasoning	6.02	1.88	6.35	3.10	10.00	6.90	0.07
Factor	Mean	SD	Median	Min	Max	Range	SE
Males							
Evaluating Arguments	5.52	1.86	5.87	2.15	10.00	7.85	0.06
Critical Analysis	5.79	1.87	5.81	3.05	10.00	6.95	0.06
Deductive Reasoning	5.91	1.91	6.36	2.75	10.00	7.25	0.06
Inductive Reasoning	6.11	1.86	6.35	3.10	10.00	6.90	0.06

Table 8. Critical Thinking Quotient Norms based on Age Group.

Age Group	Mean	SD	Median	Min	Max	Range	SE
18 and below	113.74	13.03	115.88	77.26	141.74	64.47	1.07
18-25	103.84	12.41	102.64	74.16	136.67	62.51	0.69
26-33	101.43	14.02	100.37	74.16	141.74	67.58	0.48
34-41	101.99	13.89	101.96	74.16	136.30	62.14	0.83
42-49	101.36	13.17	102.01	74.16	133.89	59.73	1.13
50-56	103.82	15.78	101.89	74.16	141.74	67.58	2.46
57 and above	97.80	15.59	98.48	74.16	122.22	48.06	3.67

Table 9. Factor Score Norms based on Age Group.

Factor	Mean	SD	Median	Min	Max	Range	SE
18 and below							
Evaluating Arguments	6.78	1.87	7.08	2.15	10.00	7.85	0.15
Critical Analysis	6.77	1.83	7.13	3.05	10.00	6.95	0.15
Deductive Reasoning	6.77	1.86	6.36	2.75	10.00	7.25	0.15
Inductive Reasoning	7.69	1.76	8.48	3.10	10.00	6.90	0.15

Factor	Mean	SD	Median	Min	Max	Range	SE
18 – 25							
Evaluating Arguments	5.55	1.66	5.87	2.15	10.00	7.85	0.09
Critical Analysis	5.81	1.86	5.81	3.05	10.00	6.95	0.10
Deductive Reasoning	6.21	1.82	6.36	2.75	10.00	7.25	0.10
Inductive Reasoning	6.19	1.84	6.35	3.10	10.00	6.90	0.10

Factor	Mean	SD	Median	Min	Max	Range	SE
26 – 33							

Evaluating Arguments	5.38	1.91	5.87	2.15	10.00	7.85	0.07
Critical Analysis	5.52	1.84	5.81	3.05	10.00	6.95	0.06
Deductive Reasoning	5.81	1.94	5.12	2.75	10.00	7.25	0.07
Inductive Reasoning	6.03	1.87	6.35	3.10	10.00	6.90	0.06

Factor	Mean	SD	Median	Min	Max	Range	SE
---------------	-------------	-----------	---------------	------------	------------	--------------	-----------

34 – 41

Evaluating Arguments	5.47	1.88	5.87	2.15	9.59	7.44	0.11
Critical Analysis	5.70	1.97	5.81	3.05	10.00	6.95	0.12
Deductive Reasoning	5.72	1.83	5.12	2.75	10.00	7.25	0.11
Inductive Reasoning	6.09	1.89	6.35	3.10	10.00	6.90	0.11

Factor	Mean	SD	Median	Min	Max	Range	SE
---------------	-------------	-----------	---------------	------------	------------	--------------	-----------

42 – 49

Evaluating Arguments	5.53	1.81	5.87	2.15	9.59	7.44	0.16
Critical Analysis	5.68	1.70	5.81	3.05	10.00	6.95	0.15
Deductive Reasoning	5.47	1.87	5.12	2.75	9.87	7.12	0.16
Inductive Reasoning	6.02	1.78	6.35	3.10	10.00	6.90	0.15

Factor	Mean	SD	Median	Min	Max	Range	SE
---------------	-------------	-----------	---------------	------------	------------	--------------	-----------

50 – 56

Evaluating Arguments	5.60	2.01	5.87	2.15	10.00	7.85	0.31
Critical Analysis	5.71	2.06	5.81	3.05	10.00	6.95	0.32
Deductive Reasoning	6.08	1.94	6.36	2.75	10.00	7.25	0.30
Inductive Reasoning	6.37	2.03	6.35	3.10	10.00	6.90	0.32

Factor	Mean	SD	Median	Min	Max	Range	SE
---------------	-------------	-----------	---------------	------------	------------	--------------	-----------

57 and above

Evaluating Arguments	5.17	1.90	5.87	2.15	8.29	6.14	0.45
-----------------------------	------	------	------	------	------	------	------

Critical Analysis	5.09	1.76	5.10	3.05	8.46	5.41	0.42
Deductive Reasoning	5.07	2.18	5.12	2.75	8.73	5.98	0.51
Inductive Reasoning	5.84	1.98	5.80	3.10	9.51	6.41	0.47

Table 10. Critical Thinking Quotient Norms based on Study Level.

Study Level	Mean	SD	Median	Min	Max	Range	SE
Level 1: Master's Degree, PhD, Engineering or Business School, MBA	106.29	14.60	105.63	74.16	141.74	67.58	0.62
Level 2: Bachelor's Degree, First-year Masters	100.46	11.46	99.65	74.16	136.95	62.79	0.39
Level 3: 2 years of Vocational or University Studies	97.64	12.47	94.36	74.16	124.11	49.95	1.61
Level 4: High School Diploma	91.76	8.88	90.85	74.16	112.16	38.00	1.52
Level 5: Trades Training No High School Diploma	105.16	14.94	107.90	74.16	125.21	51.05	3.05
Level 6: Students	107.35	12.28	109.08	85.55	130.30	44.75	2.41

Table 11. Factor Score Norms based on Study Level.

Factor	Mean	SD	Median	Min	Max	Range	SE
Level 1: Master's Degree, PhD, Engineering or Business School, MBA							
Evaluating Arguments	5.88	1.97	5.87	2.15	10.00	7.85	0.08
Critical Analysis	6.19	1.98	5.81	3.05	10.00	6.95	0.08
Deductive Reasoning	6.18	1.95	6.36	2.75	10.00	7.25	0.08
Inductive Reasoning	6.56	1.92	6.35	3.10	10.00	6.90	0.08

Factor	Mean	SD	Median	Min	Max	Range	SE
Level 2: Bachelor's Degree, First-year Masters							
Evaluating Arguments	5.27	1.62	4.66	2.15	10.00	7.85	0.05
Critical Analysis	5.39	1.69	5.81	3.05	10.00	6.95	0.06

Deductive Reasoning	5.75	1.78	5.12	2.75	10.00	7.25	0.06
Inductive Reasoning	5.91	1.71	6.35	3.10	10.00	6.90	0.06

Factor	Mean	SD	Median	Min	Max	Range	SE
<i>Level 3: 2 years of Vocational or University Studies</i>							
Evaluating Arguments	4.75	1.67	4.66	2.15	8.29	6.14	0.22
Critical Analysis	5.21	1.63	5.81	3.05	8.46	5.41	0.21
Deductive Reasoning	5.52	1.81	5.12	2.75	9.87	7.12	0.23
Inductive Reasoning	5.63	1.83	5.24	3.10	9.51	6.41	0.24

Factor	Mean	SD	Median	Min	Max	Range	SE
<i>Level 4: High School Diploma</i>							
Evaluating Arguments	4.26	1.51	4.66	2.15	7.08	4.93	0.26
Critical Analysis	4.64	1.19	4.38	3.05	7.13	4.08	0.20
Deductive Reasoning	4.55	1.35	3.98	2.75	7.50	4.75	0.23
Inductive Reasoning	5.13	1.55	5.24	3.10	9.51	6.41	0.27

Factor	Mean	SD	Median	Min	Max	Range	SE
<i>Level 5: Trades Training, No High School Diploma</i>							
Evaluating Arguments	6.12	2.56	5.87	2.15	9.59	7.44	0.52
Critical Analysis	5.88	2.07	5.81	3.05	9.79	6.74	0.42
Deductive Reasoning	5.74	1.68	6.36	2.75	8.73	5.98	0.34
Inductive Reasoning	6.59	1.71	6.35	3.10	9.51	6.41	0.35

Factor	Mean	SD	Median	Min	Max	Range	SE
<i>Level 6: Students</i>							
Evaluating Arguments	6.33	1.67	5.87	3.36	9.59	6.23	0.33
Critical Analysis	6.04	1.44	5.81	3.05	8.46	5.41	0.28
Deductive Reasoning	6.20	2.02	6.36	2.75	9.87	7.12	0.40

3.3. Reliability

Reliability refers to the extent to which an assessment is free from error and the degree to which items on a scale measure the same construct. For a measure to be reliable, the same results must be consistently generated over time. Reliability is important for an assessment as it enables us to have confidence that scores on a test are really typical of an individual's behaviour, aptitude or attitude. A primary concern when developing and using tests, is the extent to which random measurement errors influence test performance. If a test is unreliable, it is impossible for it to be valid as the accuracy of results (validity) depends on the consistency with which the same results are produced. High reliability of an assessment helps to guard against biases that may be caused by errors in human judgment during the recruitment and selection process.

Reliability values vary from 0 to 1, with higher values being more desirable. As a rule of thumb, a reliability of 0.70 or higher is recommended (Nunnally, 1978), with values greater than 0.9 (excellent); 0.7 to 0.9 (good); 0.6 to 0.7 (acceptable); 0.5 to 0.6 (poor) and values less than 0.5 (unacceptable).

Internal Reliability

Internal reliability measures the extent to which items on a scale are internally consistent and all measure the same construct. For **GAAT**, **internal consistency** was measured using Cronbach's coefficient alpha. High values of Cronbach's coefficient alpha indicate a greater degree of consistency in scores, and generally more homogeneous scale content.

GAAT has an alpha reliability of 0.77, indicating that the overall scale has a good level of consistency and is reliable.

3.4. Validity

Validity refers to the extent to which an assessment measures what it claims to measure, and that this is well-founded and accurately corresponds to actual behaviour. In psychometrics, validity relates to the degree to which practical evidence and scientific theory support the interpretations of test scores. Whilst the reliability of an assessment relates to the degree to which results are consistent over time, validity tells us the degree to which results of an assessment are accurate and enable us to confidently make predictions about a target outcome. Therefore, a test can be reliable but not valid (one can consistently get the wrong answer), but it cannot be valid without being reliable, as confidence in the accuracy of results depends on the consistency with which the same results are obtained. Established validity protects practitioners from faulty, costly, unethical and legally risky decisions.

Construct Validity

Construct validity measures the degree of the appropriateness of inferences made on the basis of observation or measurement. For work-based assessments, construct validity is particularly important as it: a) adds legitimacy and acceptability to the standardised measurement of employee behaviours, b) facilitates understanding of organisational implications of findings, and c) enables organisations and individuals to tailor interventions to meet desired performance targets. For **GAAT**, construct validity was established using an inter-factor correlations approach.

The factors in **GAAT** measure critical thinking, however, they all measure different factors of the skill. In other words, if someone scores highly on one factor, it does not mean that they will score highly on all factors.

The correlation coefficient is a numerical way to quantify the relationship between the variables, and ranges between -1 and 1. The closer the value to 1, the stronger the relationship. The correlation coefficient does not imply causality; it merely signifies the association between variables. Traditionally, factor correlations are measured using Pearson correlation (1895).

Table 12. Inter-factor Correlations.

Factor	Deductive Reasoning	Evaluating Arguments	Critical Analysis	Inductive Reasoning
Evaluating Arguments	0.42			
Critical Analysis	0.42	0.40		
Inductive Reasoning	0.50	0.48	0.48	
Critical Thinking Quotient	0.75	0.76	0.75	0.81

The results displayed in Table 12 show that a positive relationship exists between the four factors. Between each of the 4 factors – *Evaluating Arguments*, *Critical Analysis*, *Deductive Reasoning* and *Inductive Reasoning*, there exists a weak uphill (positive) linear relationship.

A positive correlation is to be expected in an assessment like **GAAT** as all four factors contribute towards the measurement of a second-order quotient factor or in other words, an overall critical thinking capacity.

3.5. GAAT and Equal Opportunities

Organisations are increasingly under pressure to ensure that tools used to inform employee recruitment, selection, development, promotion and performance measurement decisions are fair, objective and accurate. This pressure stems from a progressively knowledgeable applicant pool, a more sophisticated labour force and stringent local and international employment legislation and testing standards. Equal opportunities therefore represent a critical consideration for any modern-day work assessment, and an assessment is only considered to be truly valid when it is appropriate for use with different groups. **GAAT** is a fair and accurate measure of critical thinking, and HR practitioners and psychologists can be confident of its application.

Measurement Equivalence

Measurement equivalence (ME) refers to the consistency of measurement across some specified group demarcation such as nationality, ethnicity and gender. The focus of ME is to ensure that individuals from different groups with the same standing on an attribute have the same set of observations on an assessment (Drasgow, 1987; McLeary & Cruise, 2012). ME helps to determine the extent to which respondents are interpreting items similarly, and identified differences across groups reflect actual differences in the trait being measured (Drasgow, 1987). One approach to assessing measurement equivalence is differential item functioning analysis.

Differential Item Functioning (DIF)

An item functions differently, or displays DIF, if respondents from different groups have different probabilities of endorsing the item, despite having the same level of the attribute or trait. When a different probability of endorsing an item remains constant across different levels of the attribute or trait, the DIF is said to be uniform. Here, the item of interest consistently gives one group an advantage across all levels of the attribute or trait. Conversely, when a different probability of endorsing an item varies along the level of the attribute or trait, the DIF is characterised as non-uniform (Clauser & Mazor, 1998; Hanson, 1998). Rather than a consistent advantage being given to the reference group across the ability continuum, an advantage to one group over another exists at different levels of the trait. For instance, an item may give the reference group an advantage at the higher end of the continuum, while a disadvantage at the lower end.

DIF Analysis: Gender

DIF analysis was performed across gender to ensure that there was not a bias towards either males or females, and therefore that **GAAT** is a fair test for both genders. For this analysis, the norm sample was divided according to gender; 749 females and 952 males.

Table 13. Coefficients for Significant DIF and Effect Size for DIF Items.

Factor	Item	Chisq-value	P-value	Effect size
Critical Analysis	9	6.607	0.0368	Negligible
Critical Analysis	18	7.380	0.0250	Negligible
Critical Analysis	20	16.84	0.0002	Negligible
Deductive Reasoning	10	10.13	0.0063	Negligible
Deductive Reasoning	19	7.419	0.0245	Negligible
Deductive Reasoning	25	13.90	0.0010	Negligible
Inductive Reasoning	23	7.933	0.0189	Negligible

The analysis flagged 7 of the 30 items for DIF, based on gender. Table 14 shows the DIF coefficient as well as the R^2 effect size. According to Zumbo and Thomas' (1997) classification of effect size, any R^2 under 0.13 is considered negligible. Out of the 3 items flagged for DIF, none showed an effect size of 0.13 or over. Therefore, it can be considered that **GAAT** does not contain any item that functions differently across genders.

T-tests and ANOVAS

Although 7 items were flagged with DIF, their impact on the overall assessment is negligible, meaning that the assessment does not bias towards gender. Therefore, any significant difference can be attributed to real differences between the groups, rather than biases in the assessment. To further explore the differences found between the groups, an independent sample t-test was performed between the groups; t-tests and ANOVAs were performed between genders, age groups and study levels (see Tables 15 to 17).

Gender

In order to check for differences in scores between genders, an independent sample t-test was conducted to compare all **GAAT** factors between females (n = 749) and males (n = 952) (see Table 12).

Table 14. T-tests For Factors by Gender.

Factor	Female	Male	T.	df	P
Evaluating Arguments	3.95	4.03	-0.53	831	0.60
Critical Analysis	2.81	3.16	-2.46	831	0.01
Deductive Reasoning	3.71	3.74	-0.25	831	0.80
Inductive Reasoning	3.96	3.87	0.59	831	0.55

The t-tests found significant differences between males and females on *Critical Analysis*, with males scoring higher than females.

Age Group

In order to check for differences in scores between age group, an ANOVA was conducted to compare all **GAAT** factors between seven age groups: 18 and below (n = 148), 18-25 (n = 326), 26-33 (n = 856), 34-41 (n = 283), 42-49 (n = 135), 50-56 (n=41) and 57 and above (n = 18) (see Table 16).

Table 15. ANOVA by Age Group.

Factor	>18	18-25	26-33	34-41	42-49	50-56	57<	F	df	p-value
Evaluating Arguments	6.78	5.55	5.38	5.47	5.53	5.60	5.17	1.17	5	0.32
Critical Analysis	6.77	5.81	5.52	5.70	5.68	5.71	5.09	1.18	5	0.32
Deductive Reasoning	6.77	6.21	5.81	5.72	5.47	6.08	5.07	2.73	5	0.02
Inductive Reasoning	7.69	6.19	6.03	6.09	6.02	6.37	5.84	0.88	5	0.49

The ANOVA found significant differences for *Deductive Reasoning*. Post-hoc analysis was conducted using pairwise t-tests (Bonferroni correction) in order to see where these differences occurred. Those individuals between ages 18-25 (M = 6.21) scored higher than those between ages 26-33 (M = 5.81, p = 0.0085) and 42-49 (M = 5.47, p = 0.0269). Those individuals between ages 50-56 (M = 6.08) scored higher than those between ages 26-33 (M = 5.81, p = 0.0244) and 42-49 (M = 5.47, p = 0.0154).

Study Level

In order to check for differences in scores between study level, an ANOVA was conducted to compare all **GAAT** factors between six study levels: Level 1 - Master's Degree, PhD, Engineering or Business School, MBA (n = 557); Level 2 - Bachelor's Degree, First-year Master's (n = 880); Level 3 - 2 years of Vocational or University Study (n = 60); Level 4 - High

School diploma (n = 34); Level 5 - Trades Training, No High School Diploma (n = 24); and Level 6 - Students (n = 26) (see Table 17).

Table 16. ANOVA by Study Level.

Factor	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	F	df	p-value
Evaluating Arguments	5.88	5.27	4.75	4.26	6.12	6.33	2.27	5	0.000
Critical Analysis	6.19	5.39	5.21	4.64	5.88	6.04	1.293	5	0.000
Deductive Reasoning	6.18	5.75	5.52	4.55	5.74	6.20	7.115	5	0.000
Inductive Reasoning	6.56	5.91	5.63	5.13	6.59	6.71	8.861	5	0.000

The ANOVA found that there were significant differences for all factors. Post-hoc analysis was conducted using pairwise t-tests (Bonferroni correction) in order to see where these differences occurred.

For *Evaluating Arguments*, those individuals in Level 1 (M = 5.88) scored higher than those in Level 2 (M = 5.27, $p = 0.000$), Level 3 (M = 4.75, $p = 0.000$) and Level 4 (M = 4.26, $p = 0.008$).

For *Critical Analysis*, those individuals in Level 1 (M = 6.19) scored higher than those in Level 2 (M = 5.39, $p = 0.000$), Level 3 (M = 5.21, $p = 0.000$) and Level 4 (M = 4.64, $p = 0.008$). Those individuals in Level 4 (M = 4.64) scored lower than those in Level 6 (M = 6.04, $p = 0.029$).

For *Deductive Reasoning*, those individuals in Level 1 (M = 6.18) scored higher than those in Level 2 (M = 5.75, $p = 0.000$) and Level 4 (M = 4.55, $p = 0.000$). Those individuals in Level 4 (M = 4.64) scored lower than those in Level 2 (M = 5.75, $p = 0.005$), Level 3 (M = 5.52, $p = 0.011$) and Level 6 (M = 7.12, $p = 0.041$).

For *Inductive Reasoning*, those individuals in Level 1 (M = 6.56) scored higher than those in Level 2 (M = 5.91, $p = 0.000$), Level 3 (M = 5.63, $p = 0.004$) and Level 4 (M = 4.55, $p = 0.000$).

4. Additional Information and Considerations

GAAT was developed primarily to guide Human Resource decisions for recruitment, but its usage is manifold for organisational settings. It is important that the appropriate norms, scores and interpretations are used in order to aid selection, career development, coaching, and performance management. Accurate interpretation of psychometric tests helps both administrators and candidates to: a) put the assessment into context; b) effectively compare candidates across target capabilities; c) accurately identify areas for development; d) maximise fit between job requirements and candidate potentials; and e) promote the organisation as a fair and objective place to work. The current chapter gives guidelines for Human Resource practitioners to understand and successfully interpret candidate scores of **GAAT** for different organisational objectives.

Report generation software for **GAAT** provides graphs and personalised comments for all four factors that relate to the candidate's score. The report should only be interpreted by professionals who have completed Central Test's accredited training course and are familiar with this manual. Training in and knowledge of the assessment process are vital as scores should be interpreted in light of information regarding reliability, validity, standard error of measurement, and any accommodations to the assessment or assessment session that were made. Scores should also be presented within the context of the assessment and in a clear, meaningful and appropriate way for the target audience. It is also important to note that all the assessments by Central Test have been developed for Human Resource purposes only. Therefore, no clinical conclusions can be drawn from the results – not even from extreme scores.

Recruitment and selection is an important organisational function which can be very time-consuming and expensive. Furthermore, bad recruitment decisions have additional effects on organisational productivity, capability and sustainability. Therefore, it is recommended that prior to administering **GAAT**, recruiters have a detailed job description and person specifications based on a job analysis.

Furthermore, **GAAT** can be used in line with performance evaluations, which are a key component of employee development and success. Performance reviews should be fair and balanced assessments of an employee's achievements and capabilities, and should give the opportunity to discuss expectations and accomplishments.

Whilst **GAAT** is effective in identifying areas for development, it should not be used to directly track an employee's performance progress. Therefore, it is recommended to use objective performance assessments or other evaluations in conjunction with **GAAT**.

References

- Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice*, 17(1), 31-44. <http://doi.org/10.1111/j.1745-3992.1998.tb00619.x>
- Dewey, J. (1910/1933). *How we think: A restatement of the reflective thinking to the educative process*. Boston: Heath.
- Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72(1), 19-29. <http://doi.org/10.1037/0021-9010.72.1.19>
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 56-83.
- Hanson, B. A. (1998). Uniform DIF and DIF Defined by Differences in Item Response Functions. *Journal of Educational and Behavioral Statistics*, 23(3), 244-253. <http://doi.org/10.3102/10769986023003244>
- Holtzman, S., & Vezzu, S. (2011). *Confirmatory Factor Analysis and Structural Equation Modeling of Noncognitive Assessments using PROC CALIS*. <http://www.lexjansen.com/nesug/nesug11/sa/sa07.pdf>
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods*, 6(1), 53-60.
- Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. doi.org/10.1080/10705519909540118
- Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language. *Scientific Software International*.
- McLeary, C., & Cruise, P. (2012). Modeling Organisational Trust: Addressing Challenges of Measurement Invariance. In *Proceedings of the 11th European Conference on Research Methods* (p. 470). Academic Conferences Limited.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Pearson, K. (1895). Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London: Mathematical, Physical and Engineering Sciences*, 186, 343-414.

<http://doi.org/10.1098/rsta.1895.0010>

Schmidt, F. L., & Hunter, J. E. (1996). Intelligence and Job Performance: Economic and Social Implications. *Psychology Public Policy and Law*, Vol. 2, 447-472.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors*. In annual meeting of the Psychometric Society, Iowa City, IA, 758.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times: Learning for life in our times*. San Francisco: John Wiley & Sons.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia.

CONTACT US

Client Service: customer@centraltest.com

Training and Consulting: consulting@centraltest.com



www.centraltest.com

Central Test – Empowering Talent © 2022

All rights reserved.