# Watson-Glaser™

# Critical Thinking Appraisal

## User-Guide and Technical Manual

***UK Supervised and Unsupervised Versions 2012***

TalentLens

ALWAYS LEARNING

PEARSON

Printed in the United Kingdom

www.talentlens.co.uk

# Watson- Glaser™ Critical Thinking Appraisal

## *User-Guide & Technical Manual*

## Contents

## Tables

# Figures

# Introduction

The Watson-Glaser Critical Thinking Appraisal® (W-GCTA) is designed to measure important abilities and skills involved in critical thinking. It has been used in organisations as a selection and development tool and in academic settings as a measure of gains in critical thinking skills. Originally developed in the USA, the test has been adapted and translated for use in languages other than English. A *Mental Measurement Yearbook* review noted that the Watson-Glaser is distinguished by its voluminous research and validity studies (Geisenger, 1998).

The Watson-Glaser Critical Thinking Appraisal - Supervised and the Watson-Glaser Critical Thinking Appraisal – Unsupervised are the latest versions. They are hereafter referred to as the W-GCTA Supervised and the W-GCTA Unsupervised. The supervised version is a paper-and-pencil test, whilst the unsupervised version is administered online.

This manual includes both technical background and user guidance for the test. The first section describes the history and development of the test, its psychometric properties and validation evidence. The second section is a straightforward and practical guide to using the W-GCTA (both supervised and unsupervised) for trained test users. It includes details of interpretative reports that are available as well as case studies illustrating its use.

## Instrument Overview

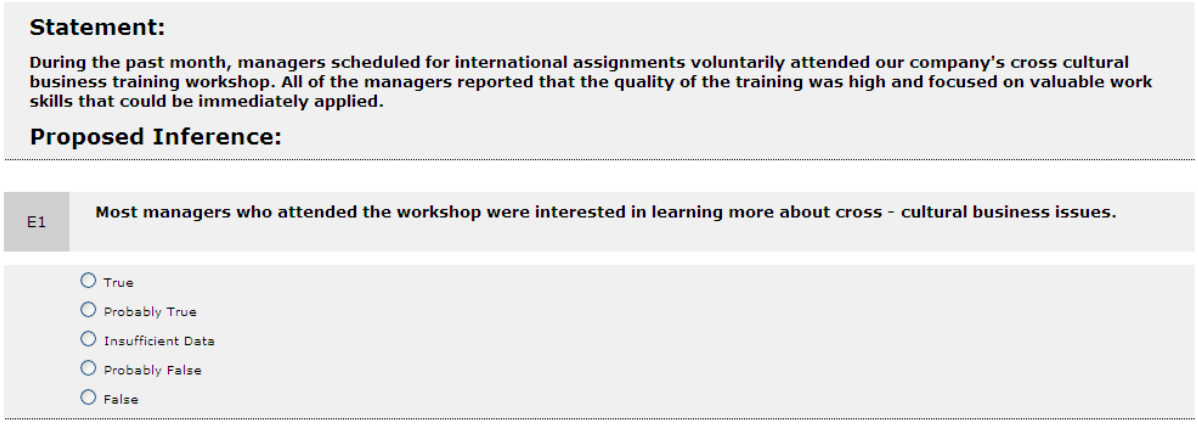
The W-GCTA is a psychometric test of critical thinking and reasoning. It measures skills related to problem solving and decision making in a variety question types. Critical thinking can be defined as the ability to identify and analyse problems as well as seek and evaluate relevant information in order to reach an appropriate conclusion. The questions are of varying difficulty and format in order to measure all areas of critical thinking ability. It is a test of power which means that it measures the quality and depth of critical reasoning rather than the speed at which an individual can perform. It can be administered untimed or with a generous time limit. It is appropriate for use with both general and high ability populations including university graduates.

The W-GCTA is a multi-faceted measure of critical thinking. The five subtests require different, though interdependent, applications of analytical reasoning in a verbal context with scores reported on three subscales.

### *Subtest 1: Inference*

Rating the probability of truth of inferences based on given information.

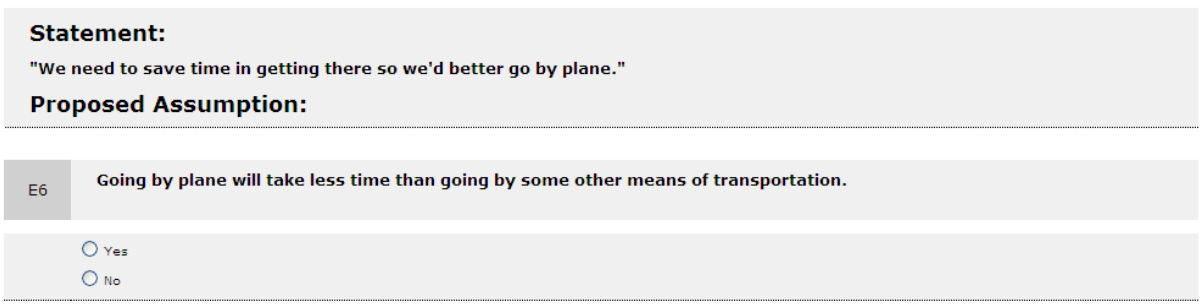
**Figure 1:** Example Inference item




### *Subtest 2: Recognition of Assumptions*

Identifying unstated assumptions or presuppositions underlying given statements.

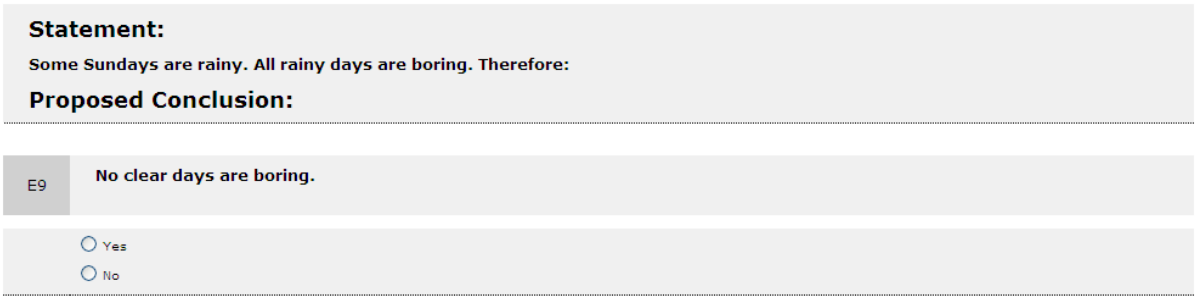
**Figure 2:** Example Recognition of Assumptions item



### *Subtest 3: Deduction*

Determining whether conclusions follow logically from given information.

**Figure 3:** Example Deduction item




### *Subtest 4: Interpretation*

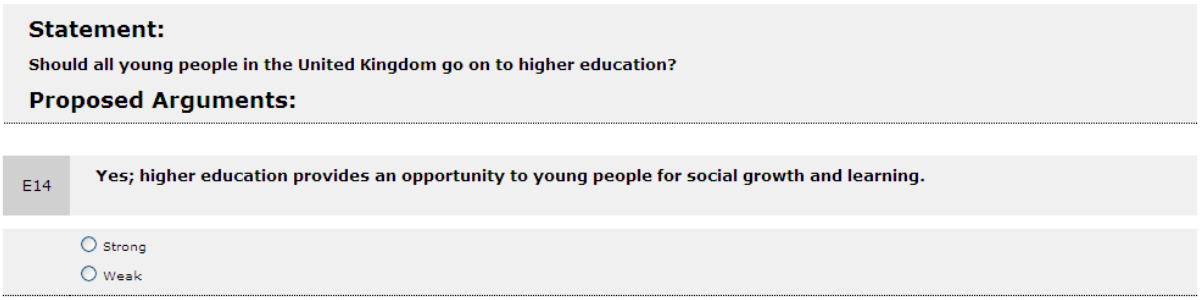
Weighing evidence and deciding if generalisations or conclusions based on data are warranted.

**Figure 4:** Example Interpretation item


### *Subtest 5: Evaluation of Arguments*

Evaluating the strength and relevance of arguments with respect to a particular question or issue.

**Figure 5:** Example Evaluation of Arguments item




Each W-GCTA subtest is composed of reading passages or scenarios that include problems, statements, arguments, and interpretations of data similar to those encountered on a daily basis at work, in study and in newspaper or magazine articles. There are a variety of topics and content is typical of that found in business and the media which requires critical evaluation and cannot

necessarily be accepted unquestioningly. Each scenario is accompanied by a number of items to which the participant responds.

The W-GCTA Supervised and Unsupervised Versions bring Watson-Glaser up-to-date with testing needs in the 21st Century whilst still retaining its robustness as a measure of critical thinking ability.

The Unsupervised W-GCTA can be administered in this way as the test is not a static, fixed-form test in which every test-taker completes the same questions. Instead, an item-banking system is used. An 'item-bank' is a large pool of questions (or 'items') from which tests are randomly generated with constraints to ensure different tests are equivalent. This is a much more secure way of administering tests online in an unsupervised environment since test-takers are unlikely to have seen the items before or complete the same items as someone else taking the same test.

The new unsupervised (item-banked) W-GCTA is scored using item-response theory (IRT). This form of scoring can adapt for minor differences in difficulty between test versions so that scores are equivalent. The online version, using the same item bank, can also be administered under supervised conditions with greater control over test conditions, candidate identity and behaviour. The supervised paper-and-pencil version of the W-GCTA is also available and this test has good psychometric equivalence to tests generated from the item-bank. The test can also be administered online but must always be supervised to maintain test security. Versions of the W-GCTA are available in other languages including: UK and US English, Chinese, Dutch, German, French, Japanese, Korean, Spanish and Swedish.

The online test versions will be automatically scored with a variety of norms and computer generated interpretive reports available. Paper and pencil versions are hand scored using acetate overlays or a bureau scoring service is available.

**What does the W-GCTA Measure?**

The W-GCTA measures the fundamental cognitive ability of critical thinking. Critical thinking is an organised and disciplined way of thinking. It is logical and approaches ideas with clarity and precision. It entails questioning assumptions, making evaluations that are fair and accurate and requires the ability to identify and focus on relevant information when reaching conclusions.

Cognitive ability has been shown to underlie performance in both work and study. They have consistently been shown to be the single best predictor of educational achievement and the most effective tool for selecting effective employees (Schmidt and Hunter, 1998, 2004; Salgado et al., 2003). Critical thinking is required to understand issues and situations solve problems and reach appropriate decisions.

Watson and Glaser (Glaser, 1937; Watson & Glaser, 1994) considered that critical thinking includes:

- attitudes of inquiry that involve an ability to recognise the existence of problems and an acceptance of the general need for evidence in support of what is asserted to be true,
- knowledge of the nature of valid inferences, abstractions, and generalisations in which the weight or accuracy of different kinds of evidence are logically determined, and
- skills in employing and applying the above attitudes and knowledge.

Consistent with this conceptualisation, the W-GCTA has maintained the same approach to measuring critical thinking. Each W-GCTA subtest is composed of reading passages or scenarios that include problems, statements, arguments, and interpretations of data similar to those encountered on a daily basis at work, in the classroom, and in newspaper or magazine articles. Each scenario is accompanied by a number of items to which the participant responds.

There are two types of scenario/item content: *neutral* and *controversial*. Neutral scenarios and items deal with subject matter that does not cause strong feelings or prejudices, such as the weather, scientific facts, or common business situations. Scenarios and items having controversial content refer to political, economic, and social issues that frequently provoke emotional responses.

As noted in the critical thinking research literature, strong attitudes, opinions, and biases affect the ability of some people to think critically (Klaczynski, Gordon, & Fauth, 1997; Nickerson, 1998; Sa, West, & Stanovich, 1999; Stanovich & West, 1997, 2008; West, Tolplak, & Stanovich, 2008). Though controversial scenarios are included throughout the W-GCTA, the majority are included in the Evaluate Arguments subtest. Evaluate Arguments scores are, therefore, expected to reflect people's ability to think critically about controversial issues.

**RED Model**

The W-GCTA introduces one notable change to Watson and Glaser's original work. Factor analyses of the existing instrument (Forms Short, A, B) consistently revealed a structure in which three scales, Inference, Deduction and Interpretation—all related to drawing conclusions—factored together. Recognition of Assumptions and Evaluation of Arguments remained as independent factors. The W-GCTA is therefore based on the RED model of critical thinking (Watson & Glaser, 2009). The three-factor model has logical appeal and provides interpretational ease. The three factors are listed below and can be seen as the keys to Critical Thinking.

### *Recognise Assumptions*

Assumptions are statements that are assumed to be true in the absence of proof. Identifying assumptions helps in the discovery of information gaps and enriches views of issues. Assumptions can be unstated or directly stated. The ability to recognise assumptions in presentations, strategies, plans, and ideas is a key element in critical thinking. Being aware of assumptions and directly assessing their appropriateness to the situation helps individuals evaluate the merits of a proposal, policy, or practice.

### *Evaluate Arguments*

Arguments are assertions that are intended to persuade someone to believe or act a certain way. Evaluating arguments is the ability to analyse such assertions objectively and accurately. Analysing arguments helps in determining what weight to put on them and what actions to take. It includes the ability to overcome a confirmation bias—the tendency to look for and agree with information that confirms prior beliefs. Emotion plays a key role in evaluating arguments as well. A high level of emotion can cloud objectivity and the ability to accurately evaluate arguments.

### *Draw Conclusions*

Drawing conclusions consists of arriving at conclusions that logically follow from the available evidence. It includes evaluating all relevant information before drawing a conclusion, judging the plausibility of different conclusions, selecting the most appropriate conclusion, and avoiding over-generalisation beyond the evidence.

The five subtests of W-GCTA map directly onto these areas with Drawing Conclusions consisting of three subtests: Inference, Interpretation and Deduction.

# Applications

The W-GCTA can be used in employment contexts to assist with selection, development and career counselling. It also has applications in educational contexts for development of critical reasoning and career counselling. The use of the test is restricted to individuals with an appropriate qualification in test use.

**Selection**

The W-GCTA can be used to predict success in jobs that require critical thinking skills.

Tests of reasoning ability have been shown to account for 25% of the variance in people's job performance and training success (Schmidt and Hunter, 1998). This means that by using reasoning tests to help make employment decisions, you can make more informed decisions on an applicant's ability to do a job and succeed in it, thus reducing error associated with recruitment decisions.

Many organisations use psychometric tests to screen-out unsuitable applicants, to rank order applicants by merit and/or to complement other selection information used to help find the most suitable applicant. Results from the W-GCTA may be used to rank order applicants or in combination with other assessment methods to provide a full profile of an applicant.

The W-GCTA provides a single score which represents a broad measurement across the verbal critical thinking domain. Scores should always be used in combination with other assessment techniques.

Before using the test as part of the selection process, organisations should ensure that the test is relevant to the role. Using inappropriate tests or relevant tests in an inappropriate manner can result in poor and unfair decisions. Job analysis and validation of the tests in the context should be carried out.

**Case Study**

There are two applicants for the role of sales manager. Sharon, who has a lively and persuasive manner, impressed the selectors at interview with her broad perspective. Martin responded well to the questions but was a little hesitant. The test results showed that Sharon's critical reasoning was quite weak (30$^{th}$ percentile) whereas Martin was well above average compared to other sales managers at 70$^{th}$ percentile. The selectors agreed that problem solving and decision making were key elements of the role and considering the test results there was no doubt that Martin was the better candidate all round.

If you require assistance on validation, contact TalentLens.

## Development

Psychometric tests can be helpful in better understanding a person's strengths and weaknesses so that appropriate development goals and activities can be set.

The W-GCTA allows a broad and in-depth analysis of a person's critical thinking skills. Scores can be broken down into sub-scores to permit a full exploration of the strengths and weaknesses within this skill.

Knowledge of critical thinking ability allows people to better understand their own strengths and weaknesses. They can then consider ways to build on their strengths and minimise the impact of their weaknesses. This might be through appropriate career choices or identifying projects and tasks where it is possible to work on areas which need development. There are also training courses which help people develop critical thinking skills. Awareness of one's limitations allows a person to take actions which will mitigate their impact. Someone who is weak in an area of critical thinking might learn to discuss important or complex decisions through with a colleague who is stronger in this area before taking action.

It is recommended that these tests be used with other assessments to create a clear and complete picture of an individual, for example, alongside personality inventories or 360 appraisals. Tests over- or under-interpreted or used in isolation of other assessments can lead to poor advice being given as the 'whole person' is not taken into account.

**Case Study**

Jasmine has been working for the environment office of her local authority since she completed her degree three years ago. She participated in a programme run by her employer to help her develop her career. As part of this she took the W-GCTA. Her profile showed that while her overall W-GCTA score was good (65th percentile) she was much weaker on the 'Recognition of Assumptions' subtest. Her mentor suggested they discuss some of the data and information she used on a day-to-day basis to identify underlying assumptions and consider their implications in order to develop this skill.

## Outplacement and Career Guidance

The W-GCTA can be used in outplacement or career guidance, for example when someone is facing redundancy, a change of circumstances, or experiencing a lack of opportunity in the current role or profession and seeking an alternative. The purpose of the assessment process is to provide a wide perspective on suitable career paths and to help individuals choose options which best suit their own abilities, needs and interests.

The W-GCTA assesses an aptitude for assimilating information and problem-solving, providing a broad assessment to explore individuals' potential, without pre-judging their suitability to a given role. This can help people develop an awareness of their own potential. This is useful for those who do and those who do not have a clear idea about what to do.

The W-GCTA produces two levels of information:

***Career scope and potential for analytical work***

Performance on the tests can indicate the scope the test taker could expect in a career. High scorers may be particularly suited to roles where there is a high need for analytical thinking and critical evaluation of data. These could include professional areas and high level strategic roles. Low scorers may be better suited to roles that do not rely heavily on these skills. These might include more operational rather than strategic roles and jobs where there is a much greater focus on interpersonal relationships or practical skills.

***Type of analytical work to which the individual might be better suited***

Generally speaking, those in financial or scientific roles may not need such a high level of performance on the W-GCTA 1 as those employed in roles that require critical thinking using language. Therefore the level of scores on the W-GCTA together with information on interests and other skills can help to identify the type of work individuals will be suited to. For example, if an individual achieves a rather low score on the W-GCTA then it can be inferred that they may be less suited to roles with a large verbal critical thinking component, such as, for example, marketing, legal advice and HR roles.

Care should be taken to avoid over-interpretation of test scores and differences between the test scores. An individual's interests, motivations and circumstances will also be important factors in making career choices.

**Case study 1**

Corinne was studying Business at University. She attended the careers service to find out about graduate recruitment schemes in general management. The advisor reviewed her W-GCTA scores and noted that she was above the 90th percentile on the test in comparison to UK Management Trainees. He suggested that these scores were high enough to consider applying to high flier schemes run by graduate recruiters. Corinne was shocked, but very excited by this possibility and began to investigate employers running these schemes.

**Case Study 2**

Mal had been employed as a Production Line Supervisor in a factory for 30 years after being promoted from a Production Line Operator. Mal is currently being made redundant. As part of the outplacement scheme Mal completed the W-GCTA. The results showed that his verbal skills were very high (75th percentile), in comparison to the UK population. Mal was surprised as he did not expect to do so well on the tests. After discussing these results with the outplacement consultant, Mal decided that he would be interested in re-entering education, after 35 years. On exploration with the outplacement consultant Mal found that he was interested in counselling, indicated by an interest in working with people and the skills he utilised in his current role. Mal started volunteering with a local support group to gain some experience in the area and then applied for a full time course using his redundancy pay out to fund his studies.

The W-GCTA will not directly help an individual realise their ambition, but can provide critical information for evaluating possibilities available and an indication of potential success.

### Training in Critical Thinking

Critical thinking is often taught in business and educational settings. The W-GCTA may be used to assess the extent to which pupils of these courses have mastered the critical thinking skill. In this context the tests should be administered prior to and following completion of the training course. The time interval between the testing occasions should be carefully considered.

If you require any assistance with this please contact TalentLens.

Candidates are able to practice their critical thinking skills via an iPhone Think-O-Meter App. Here they can think through scenarios and test their ability to separate reliable facts from assumptions, focus on the relevant information, and think critically to get the right answer. This is available for free from the ITunes App Store.

# History and Development of the W-GCTA

### History of Watson-Glaser

The Watson-Glaser Critical Thinking Appraisal has a distinguished history, dating back to its initial development in the 1920s. It was designed to measure important abilities and skills involved in critical thinking with careful consideration of the theoretical background. Since then it has been used in thousands of private and public sector organisations as a selection and development tool and in academic settings to track the development of critical reasoning skills. It has been translated into many languages and is used around the globe.

There have been a number of refinements and developments to the test since its launch, see Figure 6. These revisions were undertaken to incorporate enhancements requested by customers while maintaining the qualities that have made the Watson- Glaser the leading critical thinking appraisal for nearly a century.

Both Watson (1925) and Glaser (1937) were working on the measurement of critical thinking from early on in their careers. In 1964 two 100 item parallel forms (Ym and Zm) were published under the name Watson-Glaser Critical Thinking Appraisal in the USA (Watson & Glaser, 1964). The forms were revised in 1980 to update the language, improve clarity and eliminate racial and gender stereotypes (Watson & Glaser, 1980). The new forms, A and B, were shorter at 80 items but otherwise retained the basic test structure.

The first UK adaptation – Form C (Watson & Glaser, 1991) was based on the US form B which had already been in wide use in the UK with senior managers and high value occupational settings. American-English vocabulary and usage was replaced and content changed where not appropriate for a UK test taker. In 2000, minor revisions were made to form C and an extensive UK norming and standardisation exercise was also undertaken. The result was the 80 item WGCTA$^{UK}$ (Watson, Glaser & Rust, 2002).

A shorter 40 item form was developed in 1994 for use in employment related training and career development contexts. This used a subset of the Form A items and was published initially as Form S (Watson & Glaser, 1994; Watson & Glaser, 2008).

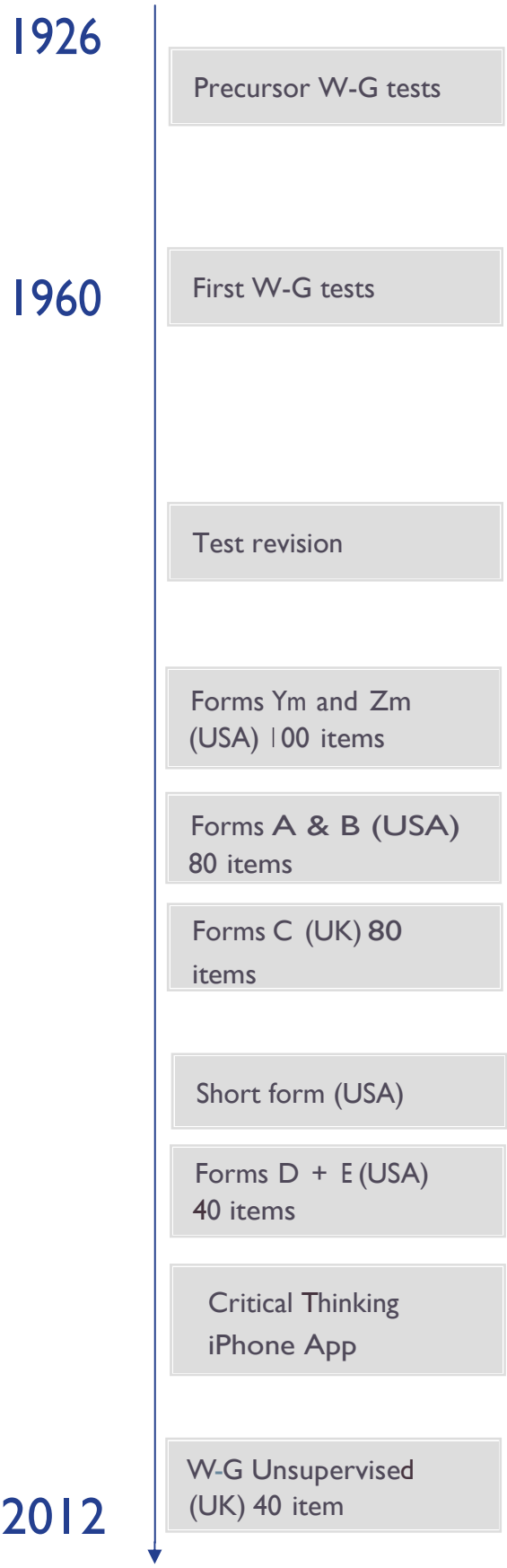
Historical and test development information for the Short Form is available in the Watson-Glaser, Short Form Manual, 2006 edition and historical and test development information for Forms A and B is available in the Watson-Glaser, Forms A and B Manual, 1980 edition.

The latest version of the W-GCTA builds on and develops these existing forms. The next section describes how the underlying model was developed and theoretically grounded. With more demand for the test around the world and shorter forms, two parallel short forms were developed. Form D is a development of the original Short Form with items chosen to be internationally appropriate and amenable to translation into other languages. A parallel 40 item version, Form E, was developed from the original form B test for use in the USA. Form D is hereafter referred to as the W-GCTA Supervised.

In order to support unsupervised internet based testing a new item banked test version has been added to the W-GCTA family of tests. In summary, there are three current versions of the test in the UK:

**WGCTA$^{UK}$** – An 80 item paper and pencil test, also known as Form C.
**The W-GCTA Supervised** – A 40 item paper and pencil test, also known as Form D.
**The W-GCTA Unsupervised** – An online 40 item test, which can be administered unsupervised.

**Figure 6:** Time line of the development of the W-GCTA



## Development of the W-GCTA

A substantial review and revision of existing Watson-Glaser test forms was undertaken leading to the publication of the W-GCTA Supervised in 2010. This introduced the clearer, more theoretically grounded, RED model and new computer generated forms for the internet age.

This revision was undertaken to incorporate enhancements requested by customers while maintaining the qualities that have made the Watson-Glaser the leading critical thinking appraisal over the last 85 years. Specific enhancements include:

- More contemporary and business relevant items.
- Better international face validity and applicability of items.
- Increased discrimination of individuals with high level critical thinking skills while maintaining discrimination at lower levels of ability.
- Shorter forms and testing times while maintaining psychometric properties.
- Internet delivered tests with very many equivalent forms.
- Enhanced computer generated reporting including a basic Profile Report, Interview Report, and Development Report.
- Improved subscale structure and interpretability.

### *Factor Structure*

In developing the W-GCTA the existing test forms and all the data that had been amassed over the years in their use was the starting point. Development of the latest version of the W-GCTA began with investigation of the factor structure of Forms A, B, and Short. A series of exploratory factor analyses were conducted using Form A, B, and the Short Form based on testlet scores. A testlet is 1 scenario and a set of 2 to 6 questions. Testlet scores were generated by summing the number of correct responses for items associated with each scenario. Evidence suggests that testlet based factor structures are more robust than those based on individual items (Bernstein & Teng, 1989).

A maximum likelihood extraction method with oblique rotation was used to analyse the Watson-Glaser Short Form (N = 8,508), Form A (N = 2,844), and Form B (N = 2,706). Initial exploration resulted in three stable factors and additional factors (four or five) that could not be interpreted. These additional factors included psychometrically weak testlets and were not stable across forms. Follow-up analyses that specified three factors revealed the configuration of Recognise Assumptions, Evaluate Arguments, and Draw Conclusions (i.e., Inference, Deduction, and Interpretation loaded onto one factor). Given this evidence, and logical appeal and interpretational ease, the three factor model was proposed for the revised W-GCTA.

An exploratory factor analysis with a UK sample of 714 who completed the WGCTA$^{UK}$ replicated the separation of Recognise Assumptions factor although the separation between Evaluate Arguments and Draw Conclusions was not as clear.

***Confirmatory Factor Analysis***

Confirmatory factory analysis (CFA) can be used to determine how well a specified theoretical model explains observed relationships among variables. Common indices used to evaluate how well a specified model explains observed relationships include the goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), and the root mean squared error of approximation (RMSEA). GFI and AGFI values each range from 0 to 1, with values exceeding .9 indicating a good fit to the data (Kelloway, 1998). RMSEA values closer to 0 indicate better fit, with values below .10 suggesting a good fit to the data, and values below .05 a very good fit to the data (Steiger, 1990). CFA can also be used to evaluate the comparative fit of several models. Smaller values of chi-square relative to the degrees of freedom in the model indicate relative fit.

During the tryout stage, a series of confirmatory models were compared: Model 1 specified critical thinking as a single factor; Model 2 specified the three factor model; and, Model 3 specified the historical five-factor model. The results, which are presented in Table 1 and Figure 7, indicated that Model 1 did not fit the data as well as the other two models. Both Model 2 and Model 3 fit the data, and there was no substantive difference between the two in terms of model fit. However, the phi coefficients in the five factor model were problematic and suggest that the constructs are not meaningfully separable. For example, the phi coefficient was 1.18 between Inference and Deduction and .96 between Deduction and Interpretation. Given this evidence, the three factor model was confirmed as the optimal model for the W-GCTA.

During standardisation there was an opportunity to replicate the confirmatory factor analyses that were run during the tryout stage. A sample of 636 people participated in the validity studies. The results of the confirmatory factor analysis supported the three factor model (GFI = .97; AGFI = .96; RMSEA = .03), providing further evidence for the three scales of the W-GCTA.

Interpretability could be improved by organising W-GCTA subscale scores according to the empirically verified three-subscale structure. This required adjusting the number of items in each subscale to improve the reliability of the shorter subscales. Specifically, each subscale is composed of a minimum of 12 items (Recognise Assumptions and Evaluate Arguments) and a maximum of 16 items (Draw Conclusions).

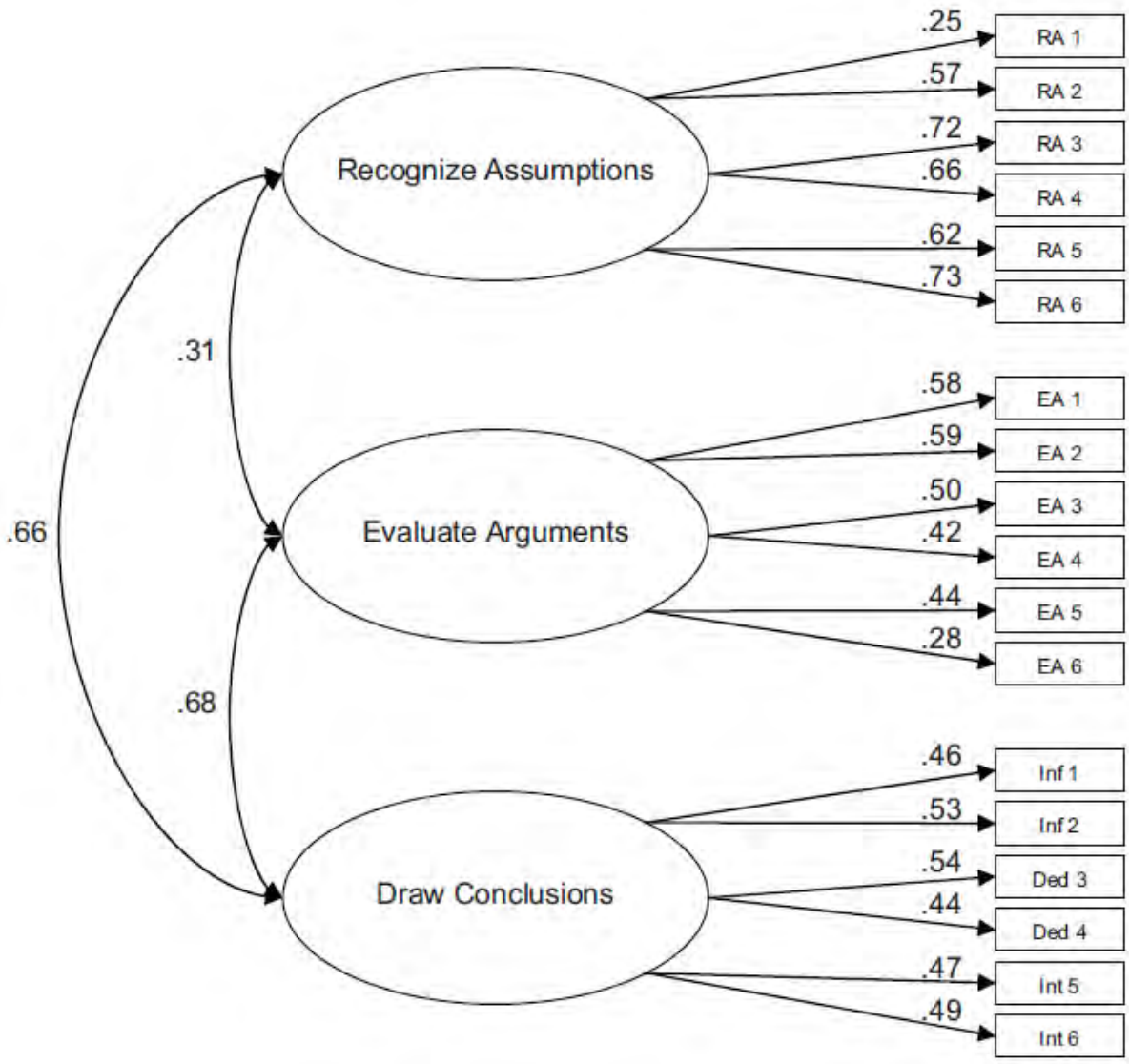
Once the revised structure of the test had been determined work commenced on extending the pool of items to support the desired new test forms.

**Table 1:** Confirmatory Factor Analyses of the W-GCTA Supervised (n = 306)

| Model | Chi-square | *df* | GFI | AGFI | RMSEA |
|---|---|---|---|---|---|
| **Model 1** | 367.16 | 135 | 0.85 | 0.81 | 0.08 |
| **Model 2** | 175.66 | 132 | 0.94 | 0.92 | 0.03 |
| **Model 3** | 159.39 | 125 | 0.95 | 0.93 | 0.03 |

*Note. The Chi-square values are Maximum Likelihood Estimation Chi-squares from SAS 9.0. See text for explanation of fit indices.*

**Figure 7:** Three Factor Model (Model 2) For Subtests and Testlets (*N* = 306)




Notes: Testlet scores were used as the unit of analysis. RA = Recognise Assumptions; EA = Evaluate Arguments; Inf = Infer; Ded = Deduce; Int = Interpret

## Developing the item pool

As well as revising the fixed test forms (W-GCTA supervised) a larger bank of items was required to support internet delivery with a large number of parallel forms.

The development of the W-GCTA item bank occurred in three main stages: item writing, trialling and analysis. These are described briefly below. The following criteria were important to address and meet:

- Ensure the new questions were equivalent to the more established items.
- Ensure that the item-banked test is reliably measuring critical thinking skills.
- Ensure that content is relevant for a global market.
- Ensure that content is appropriate for all demographic groups.
- Ensure that the item-banked test is continuing to measure the same critical thinking constructs as other versions of the Watson-Glaser.

### *Item Writing*

Existing items from the different forms were reviewed for acceptability. This included content reviews for business relevance and for international acceptability. New items were written by experienced items writers, occupational psychologists or psychometricians with at least 15 years experience. They received training and worked to an item writing specification and guide. Items underwent multiple reviews and revisions by both the item writing team and specialists in the Watson-Glaser from both the UK and the US. This ensured that new items were suitable for use in both countries. Reviews also looked at fairness issues and items were rejected if they contained topics or language which would be unequally familiar to different groups.

A total of 349 items survived the review process and were included in the first phase item trials. A similar pool of items was written and is being trialled for the second phase of the item bank development. Only results from the first phase are reported below.

### *Item Pilots*

The phase 1 new items were assigned to one of five 120 item forms. These forms included anchor items from the existing W-GCTA Supervised and WGCTA$^{UK}$ and were structured like all the Watson-Glaser tests with all five item types presented in the same order as the operational tests.

The phase 2 items were assigned to a further five forms which are just completing the trial phase at the time of writing.

All item forms were delivered via an online platform and the majority of participants undertook a supervised administration, although a small part of the sample completed the test unsupervised via a web link. Supervised testing had a time limit of one hour but unsupervised testing was not time limited.

**Sample Phase 1**

1563 people completed one or more of the 5 phase 1 pilot forms. This included just over 1200 postgraduate students on professional vocational courses who were asked to complete the test by their institutions. In addition, other university students were recruited to complete the pilot tests as a practice opportunity in preparation for other assessments. This student population is an appropriate trial group for the W-GCTA tests, which are used predominantly in the recruitment and development of graduate and other high level staff.

Additional cases were removed from the data set where there was evidence that candidates had not completed the test seriously. These included people that had completed less than half the presented items, those with very low scores and those who completed the test in less than 10 minutes.

In addition to the pilot data, a sample of 714 job candidates who had completed WGCTA$^{UK}$ as part of a recruitment process and 169 candidates who had completed the W-GCTA Supervised as part of employment evaluation were included in the final data set for analysis.

***Psychometric Analysis***

There were three stages to analysing the passages and items for inclusion in the item-bank:

Stage 1: Classical item analysis

Stage 2: Factor analysis

Stage 3: IRT analysis

***Stage 1:***

Each new form was analysed separately. Item difficulty and discrimination was examined for all items and a distractor analysis was performed for the Inference subtest where items have five response options. Scoring keys were verified and items which were not performing well were removed from the item pool. This included items which did not differentiate well between high and low performers on the test as a whole and those for which a substantial proportion of high scorers chose an incorrect response.

***Stage 2:***

In preparation for the IRT analysis it was necessary to determine whether the item pool should be analysed together, or whether the analysis should be carried out within the structure of the RED model with the three factors analysed separately. The RED model factors are correlated and therefore a combined or separate analysis might be appropriate. Each pilot form was analysed separately using principle components analysis. Examination of scree plots and parallel analysis was used to determine the number of factors. Analyses were performed at the item and testlet level.

Scree plots suggested the extraction of one or two factors for most data sets while parallel analysis suggested more factors, particularly for the trial forms which included poorer items. Pooling items into testlets resulted in fewer factors. In all cases the first factor was over double the size of other factors.

Because the results did not clearly support a uni-dimensional structure, rotation to simple structure was examined. This indicated a separation between the R items and the E and D ones. As a result the IRT analyses were performed on the pooled item sets and on separate pools of R items, and E and D items together and the results compared.

**Stage 3:**

The BILOG-MG programme was used to estimate item parameters for a number of IRT models and evaluate their fit. It was not possible to estimate parameters for some of the weaker items in the pool for some models and items were progressively dropped until a common item set was found for all the analyses for comparison purposes.

Two and three parameter uni-dimensional models were estimated for the full item pool and the R items separately from the E&D items. The criteria for selecting between IRT models included goodness of fit indices for items and the model as a whole, parameter errors of estimate and test-retest reliability for modelled parallel forms based on different parameterisations. Using these criteria, a 3 parameter model with fixed guessing showed better fit than a two parameter model. This is likely to be because the majority of the items have only two response options meaning that there is a 50% chance of answering correctly through random guessing. The current data set with less than 350 respondents for many items is insufficient for estimating the full 3PL model. Overall results were better when all items were estimated as a single pool than when R, E and D items were estimated separately. For pooled item estimation alternate form reliability ranged from 0.80 to 0.88 for different pairs of forms for the combined estimation. For separate estimation values were substantially lower for the same test pairs ranging from 0.71 to 0.82.

The final model chosen was therefore a three parameter model with a fixed guessing parameter treating all items as a uni-dimensional set. Of the 509 items included in the first pilot phase, including the W-GCTA Supervised and WGCTA$^{UK}$ items, 376 remained in the final item bank relating to 100 different passages.

***Item Bank Configuration***

The item bank was configured in order to ensure that all test takers receive equivalent tests. The number of questions in each subtest is constrained to be equal in any test form. In addition, the number of easy and more difficult questions is controlled, and each test includes questions on a variety of topics including some, but not all business related passages.

**Table 2:** Test configuration by Item Type

| | Number of questions | | |
|---|---|---|---|
| **Subtest** | **Computer Generated** | **W-GCTA Supervised** | **WGCTA$^{UK}$** |
| **Recognition of Assumptions** | 12 | 12 | 16 |
| **Evaluation of Arguments** | 12 | 12 | 16 |
| **Inference** | 5 | 5 | 16 |
| **Deduction** | 5 | 5 | 16 |
| **Interpretation** | 6 | 6 | 16 |
| **Total Number of items** | 40 | 40 | 80 |

It is difficult to calculate the exact number of different tests that can be generated from the item bank because of the different constraints that apply, but even including only the phase 1 items, and with conservative estimates of the impact of the constraints there are over 1 trillion possible tests. This makes it very unlikely that any two test takers will receive the same test version.

While there are many millions of tests that can be generated from the item bank, some of these tests will only differ in minor ways. It is therefore important to consider the likely overlap between two test versions generated. Ignoring the impact of the constraints, the average overlap between pairs of 40 item tests randomly chosen is one question relating to a common passage. This is a very low level of overlap between tests.

With computer generated tests forms it is important to be sure that there is a high level of equivalence between forms. Data on this topic is presented in the next chapter.

## Reliability

The reliability of a test is a measure of the consistency of scores; that is the extent to which the two people of the same ability or the same person tested on different occasions will receive the same score. Reliability is expressed as a coefficient which ranges from zero to one. The closer the reliability coefficient is to 1.00, the more reliable the test and the less measurement error there is associated with test scores. When tests are used in employment contexts reliabilities above .89 are generally considered excellent, .80–.89 good, and .70–.79 adequate. Values below .70 suggest the test may have limited applicability. For example, it might be used to provide developmental feedback but would not be appropriate for making hard selection or promotion decisions.

A number of methods are used to estimate test reliability. These include internal consistency of the test items (e.g. Cronbach's alpha coefficient and split-half), test-retest (the stability of test scores over time) and alternate forms analysis (the consistency of scores across alternate forms of a test).

Since repeated testing always results in some variation, no single test event ever measures an examinee's actual ability with complete accuracy. We therefore need an estimate of the possible amount of error present in a test score, or the amount that scores are likely to vary if an examinee were tested repeatedly with the same test. This value is known as the standard error of measurement (SEM). The SEM decreases as the reliability of a test increases; a large SEM denotes less reliable measurement and less reliable scores.

The SEM is a quantity that is added to and subtracted from an examinee's test score to create a confidence interval or band of scores around the obtained score. The confidence interval is a score range that, in all likelihood, includes the examinee's hypothetical "true" score which represents the examinee's actual ability. Since the true score is a hypothetical value that can never be obtained because testing always involves some measurement error, any obtained score is considered only an estimate of the examinee's "true" score. Approximately 68% of the time, the observed score will lie within +1.0 and –1.0 SEM of the true score; 95% of the time, the observed score will lie within +1.96 and –1.96 SEM of the true score.

### Internal Consistency Reliability

Internal consistency reliability considers the degree to which responses to different parts of the test are consistent with the overall score. For example if a test is reliable we would expect those who scored highest on one half of the test to have the highest scores on the other half of the test. Cronbach's Alpha is the preferred measure of internal consistency reliability for a test like the W-GCTA which is not completed under time pressure.

Reliability coefficients are sample dependent with tests showing higher reliability in more heterogeneous samples and lower reliability in groups which are all at similar levels of ability.

Table 3 shows reliability results for the W-GCTA Supervised and WGCTA$^{UK}$ for a number of samples. The reliability is above 0.8 for all but the smallest sample with the W-GCTA Supervised. The low SEM suggests that the lower reliability in this sample is due to the lower variance. For the W-GCTA Unsupervised internal consistency reliability was calculated for four possible forms to provide a range of values. The mean and standard deviation are quoted using a standardised scale which has a mean of zero and a standard deviation of 1 across all versions and participants in the pilot, since raw scores are not appropriate for comparison across test versions.

Table 3 also includes estimates of the T Score SEM. This is used to provide a band of error around a score after standardisation.

**Table 3:** Internal Consistency Reliability statistics

| | Mean | Sd | Reliability | SEM | T score SEM | Sample |
|---|---|---|---|---|---|---|
| **W-GCTA Supervised** | 30.4 | 5.0 | 0.75 | 2.5 | 5.0 | UK occupational sample, online completion n=169 |
| | 57.2 | 8.3 | 0.81 | 3.6 | 4.4 | UK standardisation sample n=1546 |
| **WGCTA$^{UK}$** | 60.1 | 11.8 | 0.92 | 3.4 | 2.8 | UK occupational sample. N=714 |
| | 57 | 13 | 0.93 | 3.6 | 2.6 | UK postgraduate students on a legal vocational course n=182 |
| | -.17 | .80 | .84 | .32 | 4.0 | UK Pilot Sample subset n=355 |
| **W-GCTA Unsupervised** | -.12 | .82 | .82 | .35 | 4.2 | UK Pilot Sample subset n=355 |
| | -.06 | .89 | .84 | .36 | 4.0 | UK Pilot Sample subset n=318 |
| | -.07 | .92 | .86 | .35 | 3.7 | UK Pilot Sample subset n=318 |

**Test-Retest Reliability**

Cognitive ability is a stable trait (Deary, Whalley, Lemmon, Crawford, & Starr, 2000), and prior versions of the Watson-Glaser have demonstrated an acceptably high level of test-retest reliability. In light of this evidence, we refer to previous research. Form B was administered twice to a group of 96 students with a three month testing interval with a correlation of 0.73.

In 1994, a study investigating the test-retest reliability of the Watson-Glaser Short Form was conducted using a sample of 42 adults who completed the Short Form two weeks apart. The test-retest correlation was .81 and the difference in mean scores between the first testing and the second testing was statistically small (d = 0.16).

In 2006, test-retest reliability was evaluated using a sample of 57 job incumbents drawn from various organisational levels and industries. The test-retest intervals ranged from 4 to 26 days, with a mean interval of 11 days. As shown in Table 4, the Watson-Glaser Short Form total score demonstrated acceptable test-retest reliability (r = .89). The difference in mean scores between the first testing and the second testing was statistically small (d = 0.17).

The last 2 studies in the table show the results of an equivalency study in 2005 between paper and pencil and computer administered versions of the Short Form. These studies are described in full in Watson and Glaser, 2006. Test-retest reliability across modes yielded similarly high results to test-retest studies within modes supporting the equivalence of paper and pencil and computer generated versions of the test.

Overall stability is high when retesting over a few weeks and with different modes. It is moderate over longer periods. There is a small increase in scores on second testing but the effect does not reach 0.3, the effect size usually considered 'small'.

**Table 4:** Test Retest Reliability

| | Mean | Sd | Reliability | Cohen's d | Sample |
|---|---|---|---|---|---|
| **Form B (precursor of WGCTA$^{UK}$)** | | | 0.73 | 0.07 | 96 students, (reported in Watson, Glaser & Rust 2002) |
| **First test** | 57.4 | 8.1 | | | |
| **Second test** | 56.8 | 8.4 | | | |
| **Test interval** | 3 months | | | | |
| **Short Form (precursor of the W-GCTA Supervised)** | | | 0.81 | 0.16 | 42 US adults, 1994 |
| **First test** | 30.5 | 5.6 | | | |
| **Second test** | 31.4 | 5.9 | | | |
| **Test interval** | 2 weeks | | | | |
| **Short Form (precursor of the W-GCTA Supervised)** | | | 0.89 | 0.17 | 57 US Job Incumbents, 2006 |
| **First test** | 29.5 | 7.0 | | | |
| **Second test** | 30.7 | 7.0 | | | |
| **Test interval** | 4 to 26 days | | | | |
| **Short Form (precursor of the W-GCTA Supervised)** | | | 0.88 | 0.13 | 108 US adults |
| **First computerised administration** | 28.8 | 5.7 | | | |
| **Second Paper and Pencil administration** | 29.5 | 5.5 | | | |
| **Short Form (precursor of the W-GCTA Supervised)** | | | 0.86 | 0.09 | 118 US adults |
| **First Paper and Pencil administration** | 30.1 | 5.7 | | | |
| **Second computerised administration** | 30.6 | 5.5 | | | |

### Alternate Forms Reliability

Alternate Form reliability considers whether similar scores are returned from different test forms. This is particularly important for the W-GCTA unsupervised, where candidates each receive a randomly designed test form with different items. To evaluate this in developing the W-GCTA

unsupervised, pairs of test forms were constructed from within the different item pilot pools so that scores could be compared. Two pairs of tests were developed that were fully equivalent. In addition two pairs of tests were developed that were maximally different in difficulty to check whether the IRT calibration could control for variation in item difficulty. These latter tests were either too easy or too difficult to meet the build constraints of the test generation system but provided a more extreme test of effectiveness

Table 5 shows that the equivalent test had alternate form reliabilities above 0.80 and that even for the non equivalent forms alternate form reliability is around 0.80 with very small differences in average scores between forms.

**Table 5:** Alternate Forms Reliability – W-GCTA Unsupervised

| | Mean | Sd | Reliability | Cohen's d | Sample |
|---|---|---|---|---|---|
| **Equivalent Pair A** | | | 0.82 | 0.06 | UK Pilot Sample subset n=355 |
| **Test 1** | -.17 | .80 | | | |
| **Test 2** | -.12 | .82 | | | |
| | | | | | |
| **Equivalent Pair B** | | | 0.88 | 0.01 | UK Pilot Sample subset n=318 |
| **Test 3** | -.06 | .89 | | | |
| **Test 4** | -.07 | .92 | | | |
| | | | | | |
| **Non Equivalent Pair C** | | | 0.78 | 0.04 | UK Pilot Sample subset n=308 |
| **Easy test** | -.17 | .89 | | | |
| **Difficult Test** | -.21 | .90 | | | |
| | | | | | |
| **Non Equivalent Pair D** | | | 0.80 | 0 | UK Pilot Sample subset n=282 |
| **Easy test** | -.24 | .91 | | | |
| **Difficult Test** | -.24 | .89 | | | |

Three studies of Alternate Forms reliability with earlier fixed form tests are shown in Table 6 overleaf, see Studies 1 and 2. These studies are reported in more detail in Watson & Glaser (1980) and Watson and Glaser (1991).

**Correlations between the W-GCTA Unsupervised and fixed form versions**

Research has been undertaken to show links between the W-GCTA Unsupervised and the W-GCTA Supervised and the WGCTA$^{UK,}$ see Table 6. A correlation of .77 (after correcting for range restriction in the W-GCTA Supervised scores: sample SD = 3.09, general population SD = 6) was found between scores on the W-GCTA Unsupervised, completed online and unsupervised, and the W-GCTA Supervised, with a group of 172 graduate applicants. See Study 4.

A correlation of .73 (corrected for range restriction in the WGCTA$^{UK}$ scores: sample SD = 4.41, general population SD = 8.33) was found between scores on the W-GCTA Unsupervised, completed online and unsupervised and the WGCTA$^{UK}$, completed under supervised conditions. See Study 5. The test takers were 284 graduate applicants.

**Table 6:** Alternate Forms Reliability – W-GCTA Unsupervised and Fixed Forms

| | **Mean** | **Sd** | **Reliability** | **Sample** |
|---|---|---|---|---|
| **Study 1** | | | 0.75 | 288 12th Grade US students |
| **Form A** | 46.8 | 9.8 | | |
| **Form B** | 46.6 | 9.3 | | |
| | | | | |
| **Study 2** | | | 0.71 | 53 sixth form students |
| **Form B** | 56.8 | 8.3 | | |
| **Form C** | 57.4 | 9.5 | | |
| | | | | |
| **Study 3** | | | 0.85 | 636 US trial participants |
| **W-GCTA Supervised *** | 27.1 | 6.5 | | |
| **Short Form*** | 29.2 | 5.7 | | |
| | | | | |
| **Study 4** | | | 0.77 | 172 UK graduates |
| **W-GCTA Supervised** | 34.1 | 3.1 | | |
| **W-GCTA Unsupervised** | .81 | .40 | | |
| | | | | |
| **Study 5** | | | 0.73 | 284 UK graduates |
| **WGCTA$^{Uk}$** | 68.4 | 4.4 | | |
| **W-GCTA Unsupervised** | .88 | .49 | | |

*Participants completed the joint item pool at one sitting.

**Reliability of Factor Scores**

As well as the overall test score, the W-GCTA can provide sub-scores for the three elements of the Red Model. In developmental contexts users may wish to compare scores for different parts of the model to understand the relative strengths and weakness of an individual's critical thinking skills. The reliability of sub-scores will be lower than for the full test as they are based on fewer items and this should be taken into account in interpreting profiles. Table 7 shows the reliability and T-score SEM values (in brackets) of the sub-scores for the W-GCTA Supervised and WGCTA$^{UK}$. The reliability of the sub-scores is good for the longer WGCTA$^{UK}$ but is moderate to low for the shorter 40 item W-GCTA Supervised. This is not surprising when there are as few as 12 items for some of the sub scores. This means that a longer test form is needed if interpretation at the sub score level is important. The final column provides the reliabilities and T-score SEM figures of the sub-scores for the item banked test based on the long 120 item pilot forms. The reliability for the 40 item forms will be somewhat lower.

**Table 7:** Sub score Internal Consistency Reliability

| **Reliability (sem)** | **W-GCTA Supervised** | **W-GCTA Supervised** | **WGCTA$^{UK}$** | **WGCTA Unsupervised Trials** |
|---|---|---|---|---|
| **Sample** | US standardisation sample N=1011 | UK occupational sample N=169 | UK occupational sample N=714 | Student Sample N=2446 |
| **Score** | | | | |
| **Full test** | 0.83 (2.6) | 0.75 (2.5) | 0.92 (3.4) | 0.90 (4.3) |
| **Recognise Assumptions** | 0.80 (1.3) | 0.66 (1.4) | 0.83 (1.4) | 0.81 (5.8) |
| **Evaluate Arguments** | 0.57 (1.5) | 0.43 (1.4) | 0.75 (1.6) | 0.66 (4.4) |
| **Draw Conclusions** | 0.70 (1.7) | 0.60 (1.5) | 0.86 (2.7) | 0.81 (3.0) |

The tables 8 and 9 show the standard error of difference in raw score, sten and T-score units between pairs of subtest scores. These show the smallest difference between pairs of scores required for 65% confidence of a real difference. Double these values provide 96% confidence. For example if a person has T scores of 55, 60 and 67 on WGCTA$^{UK}$ for Recognise Assumptions, Evaluate Arguments and Draw Conclusions respectively then only the difference between Recognise Assumptions and Draw Conclusions is greater than twice the standard error of difference (5.6 x 2), therefore we can say with 96% confidence that the person is better at Drawing Conclusions than Recognising Assumptions.

The difference between Evaluate Arguments and Draw Conclusions is 7 T score points which is greater than the standard error of difference of 6.2, therefore we can say with 65% confidence that the person is better at drawing conclusions than evaluating arguments. However the difference between Recognising Assumptions and Evaluating Arguments is only 5 points which is less than the relevant standard error of difference (6.5) therefore we cannot conclude that there is no difference in the person's ability in these two areas.

**Table 8:** Standard Error of Difference between subtest scores on the W-GCTA Supervised

| **W-GCTA Supervised Raw / Sten/ T-score** | **Evaluate Arguments** | | | **Draw Conclusions** | | |
|---|---|---|---|---|---|---|
| | **Raw Score** | **Sten Score** | **T Score** | **Raw Score** | **Sten Score** | **T Score** |
| **Recognise Assumptions** | 1.9 | 1.9 | 9.5 | 2.0 | 1.7 | 8.6 |
| **Evaluate Arguments** | | | | 2.0 | 2.0 | 9.8 |

**Table 9:** Standard Error of Difference between subtest scores on the WGCTA$^{UK}$

| **WGCTA$^{UK}$ Raw / Sten** | **Evaluate Arguments** | | | **Draw Conclusions** | | |
|---|---|---|---|---|---|---|
| | **Raw Score** | **Sten Score** | **T Score** | **Raw Score** | **Sten Score** | **T Score** |
| **Recognise Assumptions** | 2.1 | 1.3 | 6.5 | 3.0 | 1.1 | 5.6 |
| **Evaluate Arguments** | | | | 3.1 | 1.2 | 6.2 |

## Validity

Validity refers to the degree to which specific data, research, or theory support the interpretation of test scores (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). "Validity is high if a test gives the information the decision maker needs" (Cronbach, 1970). To establish the utility of the W-GCTA, components of construct validity, including content validity, internal factor structure, and convergent and discriminate validity are presented.

### Content Validity

Evidence of content validity exists when the content of a test includes a representative sample of tasks, behaviours, knowledge, skills, or abilities of the identified construct. The critical thinking skills measured by the W-GCTA were articulated many years ago by Watson and Glaser (Glaser, 1937;

Watson & Glaser, 1952), and they still correspond to critical thinking skills articulated in current models of critical thinking (Facione, 1990; Fisher & Spiker, 2004; Halpern, 2003; Paul & Elder, 2002).

Watson and Glaser (Glaser, 1937; Watson & Glaser, 1994) considered that critical thinking includes:

- attitudes of inquiry that involve an ability to recognise the existence of problems and an acceptance of the general need for evidence in support of what is asserted to be true,
- knowledge of the nature of valid inferences, abstractions, and generalisations in which the weight or accuracy of different kinds of evidence are logically determined, and
- skills in employing and applying the above attitudes and knowledge.

W-GCTA passages contain stimulus material similar to that encountered on a daily basis at work, in the classroom, and in newspaper or magazine articles. Respondents are required to show critical thinking in identifying valid and invalid inferences from passages, identifying underlying assumptions and evaluating the strength of arguments. Therefore the nature of the task is that it will require critical thinking with relevant contextual material.

In addition the W-GCTA includes both *neutral* and *controversial material*. This means that scores reflect individuals' ability to reason effectively whether or not they have strong feelings regarding the matter. As noted in the critical thinking research literature, strong attitudes, opinions, and biases affect the ability of some people to think critically (Klaczynski, Gordon, & Fauth, 1997; Nickerson, 1998; Sa, West, & Stanovich, 1999; Stanovich & West, 1997, 2008; West, Tolplak, & Stanovich, 2008).

In employment settings, the principal concern is with making inferences about how well the test samples a job performance domain—a segment or aspect of job performance which has been identified and about which inferences are to be made (Lawshe, 1975). Because most jobs have several performance domains, a standardised test generally applies only to one segment of job performance. Thus, the judgment of whether content-related evidence exists depends upon an evaluation of whether the same capabilities are required in both the job performance domain and the test (Cascio & Aguinis, 2005). In an employment setting, evidence based on test content should be established by demonstrating that the jobs for which the test will be used require the critical thinking abilities and skills measured by the W-GCTA. In classroom and instructional settings the course content and objectives of such instructional programs should correspond to the constructs measured by the W-GCTA.

**Construct Validity**

The construct validity of a test is the extent to which the test measures the theoretical construct or trait it is designed to assess. Construct validity can be demonstrated through many types of evidence and a variety of studies are desirable to support the construct validity of a test. This will

include results from content and criterion related validity studies described in other sections of this chapter. Studies specifically designed to evaluate construct validity include factor analytic studies looking at internal relationships between the parts of a test and correlations with other measures. Other evidence can come from experimental designs evaluating hypotheses related to test performance.

A number of studies provide construct validity evidence for the Watson-Glaser family of tests and for the revised W-GCTA more specifically. Exploratory and Confirmatory factor analysis results were described in an earlier section (page 15). They explored and confirmed the internal structure of the test.

***Subscale Inter-correlations***

Correlations among the W-GCTA subscales are shown in Tables 10, 11 and 12 below. The correlations are attenuated by the unreliability of the subscales and therefore are lower for the shorter W-GCTA Supervised which is less reliable than the longer UK form. The W-GCTA Unsupervised results are based on the 120 item pilot forms rather than operational 40 items tests and are therefore higher. Values in brackets show the values corrected for unreliability in the subscale scores. The highest correlation is between Evaluate Arguments and Draw Conclusions. Overall there are moderate to high corrected correlations between the subscales which are consistent with their representing three facets of the critical thinking construct.

**Table 10:** Inter-correlations among subscale scores: W-GCTA Supervised

| Scale | Mean | sd | 1 | 2 | 3 |
|---|---|---|---|---|---|
| **1. Total** | 30.4 | 5.0 | | | |
| **2. Recognise Assumptions** | 8.9 | 2.3 | .76 | | |
| **3. Evaluate Arguments** | 8.6 | 1.8 | .73 | .32 (.60) | |
| **4. Draw Conclusions** | 8.1 | 1.5 | .67 | .28 (.44) | .40 (.79) |

N=169

**Table 11:** Inter-correlations among subscale scores: WGCTA$^{UK}$

| Scale | Mean | sd | 1 | 2 | 3 |
|---|---|---|---|---|---|
| **1. Total** | 60.1 | 11.8 | | | |
| **2. Recognise Assumptions** | 12.8 | 3.3 | .74 | | |
| **3. Evaluate Arguments** | 11.5 | 3.2 | .78 | .40 (.51) | |
| **4. Draw Conclusions** | 35.8 | 7.2 | .95 | .58 (.69) | .66 (.82) |

N=714

**Table 12:** Inter-correlations among subscale scores: W-GCTA Unsupervised

| Scale | Mean | sd | 1 | 2 | 3 |
|---|---|---|---|---|---|
| **1. Total** | -0.01 | 0.95 | | | |
| **2. Recognise Assumptions** | 0.00 | 0.94 | 0.86 | | |
| **3. Evaluate Arguments** | 0.00 | 0.84 | 0.73 | 0.47 (0.64) | |
| **4. Draw Conclusions** | 0.02 | 0.95 | 0.86 | 0.58 (0.71) | 0.63 (0.86) |

N=2446

***Correlations with other measures***

Correlations with other measures provide evidence of convergent and divergent validity. Convergent validity is shown when a measure correlates strongly with other measures of the same or similar constructs. Divergent validity is shown when a measure has low correlations with measures of different constructs. Overall the pattern of correlations with other measures should reflect the degree of similarity between them.

***Measures of Intelligence and Achievement***

Over the years a number of studies have demonstrated that the Watson-Glaser tests correlate with other cognitive ability measures, including nonverbal, verbal, and numerical reasoning, achievement and critical thinking. A summary of US studies can be found in Watson and Glaser, (2010). Correlations with achievement tests ranged from 0.39 to 0.51. Correlations with other reasoning tests ranged from 0.48 to 0.70. A recent study with 62 university students who completed the WAIS-IV and the W-GCTA Supervised found a correlation of 0.52 between the Watson-Glaser total score and the WAIS full scale IQ. The strongest W-GCTA sub-score correlation was with Draw Conclusions. Three of the WAIS index scores correlated moderately with the W-GCTA but the Processing Speed Index did not reach statistical significance. The correlation with a Fluid Reasoning Complex was 0.60. An earlier study of Form A with the WAIS found a correlation of 0.41 with Full Scale IQ and 0.55 with Verbal IQ on a sample of 49 managers and executives. Overall these results suggest that the W-GCTA is related to IQ but that it is more strongly related to reasoning power than to speed of processing.

A study of 41 varied job incumbents and the Watson-Glaser Short form found a correlation of 0.53 with Raven's Advanced Progressive Matrices (Pearson TalentLens, 2006).

***Measures of Personality***

In terms of the Big 5 model of personality, the Watson-Glaser as a measure of intellectual functioning would be expected to correlate the strongest with the Openness Factor. Achievement motivation from the Conscientiousness factor might have small but significant correlations with the Watson-Glaser and Neuroticism or Emotionality might correlate negatively, particularly with the Evaluate Arguments subscale, which contains more items related to controversial topics.

Several studies have found significant relationships between Watson-Glaser scores and these personality characteristics. For example, the Watson-Glaser correlated .34 with an Openness scale on the Personality Characteristics Inventory (Impelman & Graham, 2009) in a study with a group of 171 executive and director level leadership candidates, .36 with an Openness to Experience composite (derived from the CPI Achievement via Independence and Flexibility scales; Spector, Schneider, Vance, & Hezlett, 2000) in a study of 429 management development assessment centre participants, and .33 with the Checklist of Educational Views that measures preferences for contingent, relativistic thinking versus "black-white, right-wrong" thinking (Taube, 1995) in a study of 198 US university graduates.

**Criterion-Related Validity**

One of the primary reasons tests are used is to predict a test taker's potential for future success. Criterion-related validity evidence takes place when a statistical relationship exists between scores on the test and one or more criteria. By collecting test scores and criterion scores (e.g. job performance ratings, grades in a training course, supervisor ratings), one can determine how much confidence may be placed on test scores in predicting outcomes such as job success. Provided that the conditions for a meaningful validity study have been met (sufficient sample size, adequate criteria, etc.), these correlation coefficients are important indices of the utility of the test.

Cronbach (1970) characterised criterion-related validity coefficients of .30 or better as having "definite practical value." The U.S. Department of Labor (1999) provides the following general guidelines for interpreting validity coefficients: values above .35 are considered "very beneficial," .21–.35 are considered "likely to be useful," .11–.20 "depends on the circumstances," and below .11 "unlikely to be useful". It is important to point out that even relatively lower validities (e.g. .20) may justify the use of a test in a selection program (Anastasi & Urbina, 1997). The practical value of the test depends not only on the validity, but also other factors, such as the base rate for success (i.e. the proportion of people who would be successful in the absence of any selection procedure). If the base rate for success is low (i.e. few people would be successful on the job), tests of low validity can have considerable utility or value. When the base rate is high (i.e. selected at random, most people would succeed on the job), even highly valid tests may not contribute significantly to the selection process.

***Prior Evidence of Criterion-Related Validity***

Previous studies of the Watson-Glaser have demonstrated a positive relationship between Watson-Glaser scores and various job and academic success criteria. A complete summary of these studies is provided in the manuals for the Watson-Glaser Short Form (Watson and Glaser, 2006) and Forms A and B (Watson and Glaser, 1980), respectively. A few selected findings are highlighted here and summarised in Table 13.

Using a sample of 142 job incumbents, research highlighted in Watson and Glaser (2006) found that Watson-Glaser scores (Short Form) correlated .33 with supervisory ratings of Analysis and Problem Solving behaviours, and .23 with supervisory ratings on a dimension made up of Judgment and Decision Making behaviours.

The Watson-Glaser is correlated with organisational success. For example, one study (see Watson and Glaser, 2006) found that for 2,303 job incumbents across 9 industry categories, Watson-Glaser Short Form scores correlated .33 with job success as indicated by organisational level achieved. The Watson-Glaser was also correlated with potential to advance, job performance, and specific job performance capabilities related to thinking, problem solving, analysis, and judgment.

With a sample of 64 analysts from a US government agency, research discussed in Watson and Glaser (2006) showed Watson-Glaser Short Form scores to correlate .40 with supervisory ratings on each of two dimensions composed of (a) Analysis and Problem Solving behaviours and, (b) Judgment and Decision Making behaviours, and correlated .37 with supervisory ratings on a dimension composed of behaviours dealing with Professional/Technical Knowledge and Expertise. The Watson-Glaser scores correlated .39 with "Total Performance" and .25 with Overall Potential.

Using a sample of 71 leadership assessment centre participants, Kudish and Hoffman (2002) reported that Watson-Glaser 80 Item (US form) scores correlated .58 with ratings on Analysis and .43 with ratings on Judgment. The participant group included 60 individuals from a retail/home improvement chain and 11 from a utility service, both based in the US. Ratings on Analysis and Judgment were based on participants' performance across assessment centre exercises including a coaching meeting, in-basket exercise or simulation, and a leaderless group discussion.

Spector, Schneider, Vance and Hezlett (2000) evaluated the relationship between Watson-Glaser scores and assessment centre exercise performance for managerial and executive level assessment centre participants (N = 189 to 407). They found that Watson-Glaser scores significantly correlated with six of eight assessment centre exercises, and related more strongly to exercises involving cognitive problem-solving skills (e.g. r = .26 with in-basket scores) than exercises involving interpersonal skills (e.g. r = .16 with in-basket coaching exercise). Scores also correlated .28 with "Total Performance," a sum of ratings on 19 job performance behaviours, and .24 with ratings on a single-item measure of Overall Potential.

In the educational domain, Behrens (1996) found that Watson-Glaser scores correlated .59, .53, and .51 respectively, with semester GPA for three freshmen classes in a Pennsylvania nursing program. Similarly, Gadzella, Baloglu, & Stephens (2002) found Watson-Glaser subscale scores explained 17% of the total variance in GPA (equivalent to a multiple correlation of .41) for 114 Education students. Williams (2003), in a study of 428 educational psychology students, found Watson-Glaser total

scores correlated .42 and .57 with mid-term and final exam scores, respectively. Taube (1995) found Watson-Glaser scores to be significantly associated with GPA scores in a sample of US students (r = .30). Finally, Gadzella, Stephens and Stacks (2004) reported a significant correlation between Watson-Glaser scores and both GPA (.28) and course grades (.42) with a group of 139 educational psychology students.

**Table 13:** Previous studies showing evidence of Criterion-Related Validity

| | | Watson-Glaser | | | Criterion | | | |
|---|---|---|---|---|---|---|---|---|
| **Group** | **N** | **Form** | **Mean** | **SD** | **Description** | **Mean** | **SD** | **R** |
| Job incumbents across multiple industries (Watson and Glaser, 2006) | 142 | Short | 30.0 | 6.0 | Supervisory Ratings:<br>Analysis and Problem Solving<br>Judgement and Decision Making<br>Total Performance<br>Potential | <br>37.5<br>31.9<br>101.8<br>3.1 | <br>6.7<br>6.1<br>16.9<br>1.2 | <br>.33**<br>.23**<br>.28**<br>.24** |
| Job applicants and incumbents across multiple industries (Watson and Glaser, 2006) | 2,303 | Short | 31.0 | 5.6 | Organisation Level | 3.1 | 1.2 | .33** |
| Analysts from a government agency (Watson and Glaser, 2006) | 64 | Short | 32.9 | 4.4 | Supervisory Ratings:<br>Analysis and Problem Solving<br>Judgement and Decision Making<br>Professional/Technical Knowledge and Expertise<br>Total Performance<br>Potential | <br>38.3<br>32.8<br>17.1<br>100.4<br>3.2 | <br>6.6<br>5.8<br>2.4<br>14.3<br>1.2 | <br>.40**<br>.40**<br>.37**<br>.39**<br>.25* |
| Leadership assessment centre participants from a national retail chain and a utility service (Kudish and Hoffman, 2002) | 71 | 80 Item (US version of WGCTA[UK]) | - | - | Assessor Ratings:<br>Analysis<br>Judgement | <br>-<br>- | <br>-<br>- | <br>.58*<br>.43* |
| Middle-management assessment centre participants (Spector, Schneider, Vance and Hezlett, 2000) | 189-407 | - | 66.5 | 7.3 | Assessor Ratings:<br>In-basket<br>In-basket Coaching<br>Leaderless Group<br>Project Presentation<br>Project Discussion<br>Team Presentation<br>Openness to Experience (CPI Personality Trait) | <br>2.9<br>3.1<br>3.0<br>3.0<br>2.9<br>3.1<br>41.8 | <br>.7<br>.7<br>.6<br>.7<br>.6<br>.6<br>6.4 | <br>.26*<br>.16*<br>.19*<br>.25*<br>.16*<br>.28*<br>26* |
| First year students on a Pennsylvania (US) nursing program (Behrens, 1996) | 41<br>31<br>37 | 80 Item (US version of WGCTA[UK]) | 50.5<br>52.1<br>52.1 | - | Semester 1 GPA<br>Semester 1 GPA<br>Semester 1 GPA | 2.5<br>2.5<br>2.5 | -<br>-<br>- | .59**<br>.53**<br>.51** |
| Education students (Gadzella, Baloglu and Stephens, 2002) | 114 | 80 Item (US version of WGCTA[UK]) | 51.4 | 9.8 | GPA | 3.1 | .51 | .41** |
| Educational psychology students (Williams, 2003) | 158-164 | Short | - | - | Exam 1 Score<br>Exam 2 Score | -<br>- | -<br>- | .42**<br>.57** |
| Education students (Taube, 1995) | 147-194 | 80 Item (US version of WGCTA[UK]) | 54.9 | 8.1 | GPA | 2.8 | .51 | .30* |
| Educational psychology students (Gadzella, Stephens and Stacks, 2004) | 139 | 80 Item (US version of WGCTA[UK]) | - | - | Course Grades<br>GPA | -<br>- | -<br>- | .42**<br>.28** |

Previous research has also been carried out to study the criterion-related validity of the W-GCTA Supervised. The relationship between the W-GCTA Supervised and job performance was examined using a sample of 65 managers and their supervisors from the claims division of a national US insurance company (Watson and Glaser, 2010). Managers completed the test and supervisors of these participants rated the participants' job performance across thinking domains (e.g., Creativity, Analysis, Critical Thinking, Job Knowledge) and Overall Performance and Potential.

Table 14 presents means, standard deviations, and correlations. Results showed that the W-GCTA Supervised total score correlated .44 with supervisory ratings on a scale of core critical thinking behaviours and .39 with ratings of overall potential.

The pattern of relationships at the subscale level indicated that Draw Conclusions correlated significantly with all performance ratings, and Recognize Assumptions correlated significantly with Core Critical Thinking Behaviours ($r$ = 33). Evaluate Arguments was significantly related only to Job Knowledge ($r$ = .30).

**Table 14:** Descriptive Statistics and Correlations for W-G II Scores and Performance Ratings

| | | Watson-Glaser II Form D Score | | | Performance Ratings | | |
|---|---|---|---|---|---|---|---|
| **Supervisory Performance Criteria** | **Total Score** | **Recognise Assumptions** | **Evaluate Arguments** | **Draw Conclusions** | **Mean** | ***SD*** | ***n*** |
| Core Critical Thinking Behaviours | .44 (.30)* | .33 (.23)* | .17 (.11) | .48 (.33)** | 161.1 | 17.7 | 65 |
| Evaluating Quality of Reasoning and Evidence | .43 (.29)** | .32 (.22)* | .17 (.12) | .46 (.32)** | 81.0 | 7.1 | 65 |
| Bias Avoidance | .36 (.25)* | .31 (22)* | .20 (.14) | .30 (.21)* | 22.3 | 4.1 | 65 |
| Creativity | .38 (.26)* | .25 (.17) | .15 (.10) | .45 (31)* | 23.2 | 4.7 | 65 |
| Job Knowledge | .34 (.24)* | .14 (.10) | .34 (.24)* | .30 (.21)* | 22.0 | 3.6 | 65 |
| Overall Performance | .17 (.12) | .03 (.02) | .04 (-.03) | .37 (.26)* | 4.9 | 0.7 | 65 |
| Overall Potential | .39 (.27)* | .13 (.09) | .21 (.14) | .53 (37)** | 3.2 | 0.9 | 65 |
| **W-G II Form D** | | | | | | | |
| Mean | 28.4 | 8.8 | 7.8 | 11.8 | | | |
| *SD* | 5.4 | 2.3 | 2.3 | 2.6 | | | |
| *n* | 65 | 65 | 65 | 65 | | | |

*Note. *p,.05 (one-tailed) ** p< .01 (one-tailed). Correlations were corrected for unreliability (using an average reliability of .52, based on Viswesvaron, Ones & Schmidt, 1996) and restriction of range (using normal distribution to estimate the unrestricted population variance, based on Alexander, Alliger & Hanges, 1984). Uncorrected correlations with performance ratings are shown in parentheses.*

### *Recent Validity Studies*

A study of the validity of the WGCTA$^{UK}$ for predicting outcomes on the Bar Professional Training Course (BPTC) was undertaken in 2010. This is a post-graduate vocational course for those wishing to practice as a barrister in England and Wales. 123 students on the course completed the WGCTA$^{UK}$ for the purposes of the study during a one year vocational training course. At the end of the course the average final exam grade for students was collected. The final exam grade included both more traditional written exams and ratings on vocational exercises such as writing opinions and arguing a case. The correlation between course results and the WGCTA was 0.62.

This is a very strong correlation and indicates that the WGCTA is a very good predictor of performance on this vocational course.

A larger sample of students from the course was required to complete pilot forms of the W-GCTA Unsupervised in 2011. These were calibrated as described in the development section of this manual to provide scores on a standardised scale across test versions. Course exam results were available for 988 of the 1501 students who completed the test. Further details of the sample appear in the next chapter. The mean and standard deviation for the W-GCTA and test results are shown in Table 13. The correlation between the test scores and exam results was 0.51. This is another highly statistically significant and strong result showing the ability of the W-GCTA to predict performance in a vocational training context.

**Table 15:** W-GCTA and Exam results for Law students

| | Mean | Standard Deviation | N |
|---|---|---|---|
| **W-GCTA Unsupervised** | -0.13 | 0.88 | 988 |
| **Course Exam Result** | 72.3 | 8.1 | 988 |

**Differential Validity**

In the study of students on a vocational law course described above, differential validity analyses were undertaken. This analysis compares whether there are difference in correlation and prediction for different subgroups of a sample. In particular these analyses address the question of whether groups with lower scores on the test perform in the way that would be predicted from those scores or whether the test scores underestimate their final performance level suggesting bias in the test scores. The following table summarises the results.

**Table 16:** Differential validity results summary

| Group | Results |
|---|---|
| **Gender** | No evidence of test bias and the test predicts for both groups. |
| **Ethnic origin** | No evidence of bias against those groups who scored lower on the test on average. The test predicted exam performance well for both groups. |
| **Age** | A statistically significant difference was found with the test marginally favouring younger candidates but the effect was so small as to have no meaningful impact. |
| **Primary Language** | The difference in course results was larger than that predicted by the test scores. Therefore test scores of those with English as a second language did not underestimate their level of performance on the course. |
| **Disability** | Correlation of 0.52 for non disabled group and 0.5 for disabled group suggests test predicts well for both groups. |

## Fairness and Group Comparisons

In any assessment process it is important to consider issues of fairness and equality of opportunity for both legal and ethical reasons. In particular test users need to be aware of whether members of some groups are likely to perform less well on an assessment and if so to consider whether the use of that measure can be justified. For example it is commonly found that women score less well on tests of spatial reasoning on average than men. This will have the effect, where the test is used in making selection decisions, that fewer women will be successful. The questions of fairness are whether the lower average score for women is accurately reflecting underlying ability levels of men and women rather than an artefact of the test and whether the ability measured is relevant to job performance.

A difference can be an artefact of the test where the content of the test or the way questions are asked particularly favours one group over another. For example if the W-GCTA contained passages of information that were mostly more of interest to men than women this could lead to an unfair difference. If the W-GCTA test was used to select people for a role where high level critical thinking was not relevant, it would result in unfair decisions if some groups performed less well on the test. For this reason users should ensure that the test is relevant to their context.

If the test can be shown to be relevant but group differences are found, then care should be taken not to require a higher standard of test performance than can be justified with respect to the job. This is because there will be greater adverse impact with higher cut scores than with lower scores. In other words the percentage of those selected who are from the higher scoring group will be higher the higher the cut off score that is used.

In the UK adverse impact with respect to gender, ethnic origin, age, disability, religion or sexual orientation may be illegal unless it can be clearly justified in terms of the requirements of the job.

The tables in this chapter show the results of comparing performance on the W-GCTA for different groups. They will provide an indication regarding whether group difference are likely to be expected when the test is used. However users should remember that all groups are different and even if there are score differences in one circumstance this does not mean they will always occur. For example a female led company may be more attractive to women applicants and therefore may receive more strong women candidates than a company with few high level women which has a poor track record in the equal opportunities field. High test scores from women in this situation may reflect the fact that the company attracts the best women rather than any unfairness in the test.

The comparisons in the tables below show performance on various forms of the W-GCTA broken down by various demographic factors for a variety of different samples. The difference between groups is expressed as a Cohen's d statistic. This expresses the difference in standard deviation units and therefore the statistic can be compared against results from different forms of the test or where scores are expressed on different scales. Raw score differences are not comparable in the same way. Values of Cohen's d above 0.7 are considered large, above 0.5 moderate and above 0.3 small. Below this level values can be considered negligible. Differences which reach statistical significance, i.e. those that are unlikely to be just chance findings are marked with an asterisk. It should be remembered that very small differences can reach statistical significance in large samples, but they will have little impact on relative success rates.

The data come from a number of different UK samples. The W-GCTA Supervised data is a mixed occupational group of 169 people who completed the test online for selection or development purposes. The UK standardisation sample for WGCTA$^{UK}$ is a large mixed occupational group of 1546 respondents and includes people from commercial, industrial and public organisations. Job levels range from clerical workers and security staff through to senior managers and professionals. Just over half were educated to university degree level. Another large sample consisted of 2321 job applicants to a government department. Over 80% were external applicants but there were some internal applicants and the majority were educated to graduate level. In addition there are three groups of law students. The first is a group attending a particular college who completed the WGCTA$^{UK}$. The second group (n=182) also completed the WGCTA$^{UK}$ as part of a pilot validation of the test for use in selection to the Bar Professional Training Course. The third group of 1501 BPTC students completed one of the 5 new item pilots which were used to develop the bank for the W-GCTA Unsupervised, under both supervised and unsupervised conditions. Their scores are reported here on a raw theta scale which has an approximate mean of zero and a standard deviation of 1. They also completed the test for research purposes. The results presented in Tables 15 to 23 below are based only on those people who answered the relevant demographic questions and therefore sample sizes in the tables do vary.

**Gender Comparisons**

**Table 17:** W-GCTA score gender comparisons

| Test Form and Group | Male | | | Female | | | Cohen's d |
|---|---|---|---|---|---|---|---|
| | **Mean** | **Sd** | **n** | **Mean** | **sd** | **n** | |
| **W-GCTA Supervised, mixed occupational group** | 31.5 | 4.5 | 69 | 29.7 | 5.2 | 68 | 0.37* |
| **WGCTA$^{UK}$, standardisation sample** | 50.6 | 10.3 | 1019 | 50.0 | 9.3 | 520 | 0.01 |
| **WGCTA$^{UK}$, government department** | 60.2 | 9.3 | 997 | 58.4 | 8.9 | 1269 | 0.20* |
| **WGCTA$^{UK}$, law students 1** | 60.2 | 10.1 | 311 | 60.8 | 10.9 | 48 | -0.06 |
| **WGCTA$^{UK}$, law students 2** | 57.8 | 14.8 | 65 | 57.2 | 10.5 | 82 | 0.05 |
| **W-GCTA Unsupervised, law students 3** | -0.05 | 0.90 | 449 | -0.20 | 0.82 | 545 | 0.18* |

There are only small or negligible differences between men and women in terms of average scores on these different test versions. The largest difference is for the mixed occupational group on the W-GCTA Supervised, but this remains only a small effect. Where differences do reach statistical significance it is the men who are scoring slightly higher than the women on average.

**Ethnic Origin Comparisons**

**Table 18:** W-GCTA score ethnic origin comparisons

| | White | | | Black | | | Asian | | | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Mean** | **Sd** | **n** | **Mean** | **Sd** | **n** | **Mean** | **Sd** | **n** | |
| **WGCTA$^{UK}$,** | 58.0 | 8.1 | 1332 | 50.1 | 6.9 | 60 | | | | 0.99* |
| **standardisation** | 58.0 | 8.1 | 1332 | | | | 52.1 | 8.1 | | .73* |
| **sample** | | | | | | | | | | |
| **WGCTA$^{UK}$,** | 61.8 | 7.7 | 1484 | 51.3 | 9.6 | 245 | | | | 1.31* |
| **government** | 61.8 | 7.7 | 1484 | | | | 52.1 | 8.1 | 313 | 1.25* |
| **department** | | | | | | | | | | |
| **WGCTA$^{UK}$,** | 63.4 | 7.7 | 370 | 54.5 | 13.7 | 21 | | | | 1.09* |
| **law students 1** | 63.4 | 7.7 | 370 | | | | 55.2 | 12.4 | 71 | .95* |
| **WGCTA$^{UK}$,** | 61.7 | 6.7 | 46 | 55.1 | 7.1 | 11 | | | | 0.98* |
| **law students 2** | 61.7 | 6.7 | 46 | | | | 53.7 | 6.6 | 21 | 1.20* |
| **W-GCTA** | 0.14 | 0.85 | 533 | -0.49 | 0.78 | 82 | | | | 0.75* |
| **Unsupervised,** | 0.14 | 0.85 | 533 | | | | -.52 | 0.72 | 265 | 0.82* |
| **law students 3** | | | | | | | | | | |

Large, statistically significant differences are found for all the group comparisons. Black groups consisting of those of African or Afro-Caribbean background tended to have the lowest scores although not for some of the legal groups. Differences ranged from three quarters of a standard deviation to one and a quarter standard deviations. These are large differences but similar to those found on other cognitive ability tests (e.g. Hough, Oswald and Ployhart, 2001). There may be some confounding with people who have English as a second language but most of those from minority ethnic background have English as their primary language.

Validation results for the third law sample (see previous chapter) showed that the W-GCTA was valid for each ethnic group. Therefore there is no evidence that these score difference reflect unfair bias in the test itself. However users should be aware that there will be adverse impact against the lower scoring groups when the test is used in decision making and care should be taken not to put more weight on the test scores than can be justified in terms of the importance of critical reasoning for the role.

**Age Comparisons**

The table below shows test performance for three different age groups. For most samples the youngest test takers perform slightly better than the oldest test takers, but this is reversed for the W-GCTA Supervised sample where the youngest test takers have the lowest average test score. The d statistic and significance test is a comparison of the youngest and oldest groups and two largest groups in each sample where these are different since the age profiles differ across samples. The differences vary between around zero and almost half a standard deviation. This suggests that differences are not due to bias in the test per se but to sampling factors. For example older applicants for entry level jobs may be less able than younger applicants. Whereas all young people would expect to start in entry level roles, the more able older workers are likely to have been promoted to more senior positions and when they apply for new jobs apply for higher level ones. Therefore there may be a tendency for older applicants for junior positions to be less able on average than a comparable group of younger applicants. For this reason, higher scores among younger groups do not necessarily mean that test scores decline with age.

**Table 19:** W-GCTA score age comparisons

| | 16-24 | | | 25-44 | | | 45+ | | | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Sd | n | Mean | Sd | n | Mean | Sd | n | 16-24 vs. 45+/ largest groups |
| **W-GCTA Supervised, mixed occupational group** | 28.6 | 5.9 | 15 | 30.9 | 4.8 | 105 | 30.7 | 5.5 | 16 | -0.37 / 0.04 |
| **WGCTA$^{UK}$, standardisation sample** | 62.6 | 9.3 | 380 | 58.1 | 11.8 | 182 | 61.9 | 6.9 | 25 | 0.08 / 0.44* |
| **WGCTA$^{UK}$, government department** | 60.6 | 8.5 | 671 | 58.7 | 9.1 | 1354 | 57.7 | 10.4 | 269 | 0.21*/ 0.32* |
| **WGCTA$^{UK}$, law students 2** | 59.5 | 10.0 | 98 | 54.1 | 16.7 | 46 | 59 | 7.8 | 3 | 0.05 / 0.43* |
| **W-GCTA Unsupervised, law students 3** | -0.12 | 0.84 | 647 | -.11 | 0.95 | 280 | -0.39 | 0.76 | 59 | 0.32* / -.01 |

### Disability Comparisons

There were no significant differences in performance for candidates with and without disabilities and the Cohen's d statistics are all in the negligible range. Because of the small numbers of test takers with disabilities, results have been pooled across different disabilities and it could be that these overall results mask greater differences for specific disabilities. The most common disability noted for the law students was dyslexia. The 31 students who described themselves as dyslexic had an average test score of 0.03. This is higher than the non disabled students. Where appropriate, accommodations have been made for people with disabilities in carrying out the testing. This suggests that with appropriate accommodation the W-GCTA is accessible to people with a range of disabilities.

**Table 20:** W-GCTA score disability comparisons

| Test Form and Group | No Disability | | | Disabled | | | Cohen's d |
|---|---|---|---|---|---|---|---|
| | Mean | Sd | n | Mean | sd | n | |
| **W-GCTA Supervised, mixed occupational group** | 30.6 | 4.9 | 126 | 31.5 | 5.5 | 6 | -0.18 |
| **WGCTA$^{UK}$, government department** | 59.1 | 9.1 | 2220 | 58.3 | 10.1 | 89 | 0.09 |
| **W-GCTA Unsupervised, law students 3** | -.14 | 0.87 | 783 | -0.19 | 0.91 | 75 | 0.06 |

### Primary Language Comparisons

Primary language information was available for the three samples of law students. Students whose primary language is English performed better than those for whom English is a second language. The W-GCTA test items are written to have a reading age of 15 or below, but this may still be difficult for someone who does not have a good grasp of English. The difference is smallest for the third law student sample and can generally be expected to vary according to the language level of test takers. Users should consider whether use of the test is appropriate for test takers who do not have a good grasp of written English. The test is available in Chinese, Dutch, German, French, Japanese, Korean, Spanish and Swedish.

**Table 21:** W-GCTA score primary language comparisons

| Test Form and Group | Primary Language English | | | Other Primary Language | | | Cohen's d |
|---|---|---|---|---|---|---|---|
| | Mean | Sd | n | Mean | sd | n | |
| **WGCTA$^{UK}$, law students 1** | 62.3 | 8.7 | 437 | 56.1 | 14.2 | 44 | 0.66* |
| **WGCTA$^{UK}$, law students 2** | 59.2 | 11.7 | 124 | 50.0 | 14.6 | 20 | 0.75* |
| **W-GCTA Unsupervised, law students 3** | -0.09 | 0.86 | 867 | -.33 | 0.79 | 80 | 0.28* |

**Sexual Orientation Comparisons**

Reporting levels for sexual orientation were lower than for other demographic variables and therefore samples are a little small. For most groups those gay, lesbian and bisexual respondents performed better on the test but none of the effects reached statistical significance for these groups.

**Table 22:** W-GCTA score sexual orientation comparisons

| Test Form and Group | Heterosexual or Straight | | | Gay Lesbian or Bisexual | | | Cohen's d |
|---|---|---|---|---|---|---|---|
| | Mean | Sd | n | Mean | sd | n | |
| **W-GCTA Supervised, mixed occupational group** | 30.3 | 4.9 | 127 | 35.0 | 4.0 | 6 | -0.97 |
| **WGCTA$^{UK}$, standardisation sample** | 61.0 | 10.8 | 93 | 59.1 | 12.7 | 24 | 0.17 |
| **WGCTA$^{UK}$, government department** | 59.2 | 9.1 | 2070 | 61.3 | 8.3 | 64 | -0.23 |
| **WGCTA$^{UK}$, law students 2** | 56.8 | 13.0 | 127 | 62.9 | 6.8 | 9 | -0.48 |
| **W-GCTA Unsupervised, law students 3** | -0.14 | 0.85 | 875 | -.04 | .95 | 31 | -0.12 |

**Comparison by Religion**

There is a mixture of negligible, small, medium and large differences when comparing groups defined by religious belief. Those with no religious belief tend to perform the best followed by minority religions included in the 'other' category. It is likely that religion is confounded with ethnic background and therefore these differences are predominantly reflecting the differences seen in the ethnic origin comparisons.

**Table 23:** W-GCTA score religion comparisons

| | 0. None | | | 1. Christian | | | 2. Other | | | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Sd | n | Mean | Sd | n | Mean | Sd | n | 0-1/0-2/1-2 |
| **W-GCTA Supervised, mixed occupational group** | 31.2 | 5.0 | 63 | 29.9 | 5.0 | 63 | 30.4 | 2.6 | 5 | 0.26<br>0.16<br>0.10 |
| **WGCTA$^{UK}$, law students 1** | 60.8 | 10.1 | 257 | 62.9 | 6.8 | 52 | 58.1 | 12.3 | 108 | 0.22<br>0.25*<br>0.44* |
| **WGCTA$^{UK}$, law students 2** | 64.2 | 6.8 | 41 | 58.1 | 12.0 | 60 | 48.3 | 14.3 | 35 | 0.60*<br>1.46*<br>0.76* |

**Table 24:** W-GCTA score religion comparisons for employees at a government department

| WGCTA$^{UK}$, government department | | | | Cohen's d | | | |
|---|---|---|---|---|---|---|---|
| | Mean | Sd | n | 2 | 3 | 4 | 5 |
| **1. Christian** | 58.5 | 8.9 | 1087 | 0.74* | 0.78* | 0.69* | 0.01 |
| **2. Muslim** | 52.0 | 8.4 | 163 | | 0.06 | 0.05 | 0.74* |
| **3. Hindu** | 51.5 | 9.2 | 66 | | | 0.10 | 0.70* |
| **4. Sikh** | 52.4 | 8.1 | 59 | | | | 0.66* |
| **5. Other** | 58.6 | 12.8 | 19 | | | | |

**Table 25:** W-GCTA score religion comparisons for law students completing the W-GCTA Unsupervised

| W-GCTA Unsupervised, law students 3 | | | | Cohen's d | | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Sd | n | 1 | 2 | 3 | 4 | 5 |
| **0. None** | 0.18 | 0.87 | 261 | 0.35* | 0.99* | 1.08* | 0.44* | 0.10 |
| **1. Christian** | -0.11 | 0.82 | 390 | | 0.65* | 0.77* | 0.11 | 0.24 |
| **2. Muslim** | -0.62 | 0.69 | 151 | | | 0.16 | 0.60* | 1.00* |
| **3. Hindu** | -.73 | 0.66 | 38 | | | | 0.77* | 1.14* |
| **4. Buddhist** | -.20 | 0.72 | 36 | | | | | 0.39 |
| **5. Other** | 0.09 | 0.77 | 40 | | | | | |

# Using the W-GCTA

The W-GCTA measures high level reasoning and critical thinking skills, relevant to problem solving and decision-making in a variety of graduate and managerial roles. Critical thinking can be defined as the ability to identify and analyse problems, and seek and evaluate relevant information in order to reach appropriate conclusions.

The tests are suitable for use in a variety of organisational contexts, including selection, development and career counselling across commercial, industrial and public sector organisations. Reference groups are available for UK population, different levels of management roles, and workers from different employment sectors.

### Who can use the W-GCTA

The W-GCTA is a restricted psychometric instrument and should only be used by people who are appropriately trained in the use of ability measures in an occupational context. Holders of the *BPS Test User, Occupational, Ability* qualification can access the test. In addition people who have received specific training in the use of the W-GCTA or other relevant test training may be able to purchase the test. Please contact Pearson TalentLens to verify whether you have appropriate training to use the W-GCTA. A trained test user has the knowledge and skills to decide whether a test is appropriate for use and to administer and interpret scores appropriately. Where tests are regularly administered under supervision, a trained test administrator may undertake the administration of the test but is not qualified to interpret the results or provide feedback to candidates.

### When to use the W-GCTA

In choosing to use the one of the W-GCTA test forms, test users should be satisfied that these tests are relevant and appropriate to each situation. This will depend on the purpose of testing and the group being assessed. The section below provides guidance on using the test in different contexts and choosing an appropriate version and norm group by which to compare the test takers.

### Using the W-GCTA for Selection

Tests of reasoning ability have been shown to be the most effective single predictor of job performance and training success (e.g. Salgado et al., 2003; Robertson & Smith, 2001; Schmidt & Hunter, 1998; 2004). This means that by using reasoning tests, such as the W-GCTA, more informed decisions on an applicant's ability can be achieved, reducing poor recruitment decisions. The W-GCTA may be used as an initial screen to sift out poor candidates either unsupervised via the internet or under supervision at a Pearson Vue centre or the employer's premises. It can be used in combination with other assessments (as part of an assessment centre) to provide a full profile of an applicant. Before using the W-GCTA as part of a selection process, organisations

should ensure that the test is relevant and appropriate for the role. Using inappropriate tests can result in poor and unfair decisions.

In selection, the purpose of testing (and assessment) is to provide information needed to choose between job applicants. The information is collected for the employer's use.
There are two key aspects to consider:

1. Is an assessment of critical thinking skills relevant?
2. If so is the W-GCTA relevant in terms of difficulty level and the group to be tested.

Job analysis provides recruiters with a clear understanding of a job and of what that job entails. Job analysis is the process of breaking down a job to its tasks, requirements, and performance criteria. There are formal methods of job analysis which are most effective (e.g. questionnaires, critical incident analysis), but as a minimum there should be a discussion with people who know the job well. It is advantageous to talk to both managers and job incumbents as they may have different perspectives on the role. Other informants may also be helpful (e.g. customers, trainers, reports). The information gathered is used to write a job description and person specification. A job description lists components of the job, duties or tasks, responsibilities and the required standards of performance. A person specification lays out the personal characteristics necessary to do the job; these include specific skills and abilities, interests, and disposition. They often include a competency profile.

The information in the job description and person specification should be used to decide on the type of assessment that will be relevant to the role, for example the level of difficulty at which critical thinking skills with verbal or numeric data, spatial ability, mechanical aptitude should be measured.

The test user should also look at research findings to ensure that the test is relevant in terms of its level and the group to be tested. This will include norms, reliability, validity, and group comparisons. Standards of assessment should not be higher than that which the job requires.

You should be satisfied that the test(s) has norms that provide suitable comparators for your test takers. These should contain an appropriate representative sample both with respect to the type of jobs applied for and background of the sample. Normed scores can be used to understand the ability level of a particular candidate or to rank order candidates for short listing purposes. However it is not advised that short listing should be undertaken using a single measure as this will only reflect a single aspect of performance. Further information on using norms is provided at the end of this section.

For a discussion of reliability, validity and fairness of the W-GCTA, refer to previous chapters.
A good understanding of the role together with careful selection of tests and norm groups ensures sound evidence for the decision to use a test and this process should be documented. The organisation could be required to show that any assessments used were carefully chosen and relevant if legally challenged.

If you require assistance to decide on the appropriateness of W-GCTA for your selection context, please contact Pearson TalentLens. A specific interview report is available to assist interviewers with questions following administration of the W-GCTA for selection.

**Case Study 1**

A financial organisation wanted to identify employees for their High Flyers scheme. The programme was open to all employees. Successful employees would be given placements designed to help them develop their skills and gain a broader perspective of the organisation. In addition, special training opportunities would be provided. The recruitment manager reviewed the job descriptions for all posts in which candidates would be placed. Critical reasoning was clearly required, as roles involved analysis of financial markets, but the recruitment manager was not sure whether this was mainly numerical reasoning or whether more general critical thinking was required. As this was a new scheme there was little information available about the details of the requirements. The recruitment manager spoke with trainers of relevant courses to get more information on the requirements, and discovered trainees need to use and review technical written documents, for example legal and policy documents. Based on this evidence, the recruitment manager decided there was a clear need for reasoning in the verbal domain and included the Watson-Glaser in the selection process.

**Case Study 2**

A job analysis carried out for Business Analysts in a retail organisation two years ago, identified key tasks to be financial reporting, forecasting sales and report writing. The organisation had been using a numerical test to assess the critical thinking skills required for these roles.

Recently the organisation has introduced some new technology which automatically reports and forecasts sales. The recruitment manager felt that numerical critical thinking skills may be less important to this role than they once were. To gather more information, she spoke with the designers of the technology and two Sales Managers, who had been using the new technology for 6 months. She found that Sales Managers no longer needed to perform complex analysis of data. However they did need to critically assess the implications of the reports that they received and integrate this with information about current market trends. The recruitment manager decided that critical thinking skills were still necessary, but that it would be better to measure these in a verbal context given the changes to the role.

## Using the W-GCTA for Development, Outplacement and Career Guidance

In guidance, the purpose of testing is to provide individuals with information they need to make realistic occupational decisions. Tests can be used to develop an awareness of potential, explore occupational awareness and identify special training needs.

Tests should be used when they will help the individual to make decisions and explore broader options. The test must be used in agreement, and the test taker must be aware of the kind of information that can be gained, and the decisions that are relevant to this, as well as the limitations of the test.

In choosing to use a test in this context, the test user should evaluate what the test can do and the information being sought. The W-GCTA allows individuals to develop an awareness of their potential; their ability to think critically and work with information. The test allows a broad assessment for those who are unclear about what they want to do; for those who have a clear direction, the tests provide a more specific assessment of their potential to succeed in relevant roles or training.

**In a development context**, the W-GCTA can be helpful in better understanding a person's critical thinking skills. Their score can be broken down into the three areas of the RED model to allow exploration of their strengths and areas for development. In the work setting, this can help managers and employees to define appropriate development goals and activities. In a coaching or career exploration setting, this can help individuals' awareness of their critical thinking ability, allowing them to consider ways to build on their strengths and minimise the impact of their weaknesses. This might be through appropriate career choices and work strategies or identifying development opportunities where it is possible to work on areas which require development. There are some training courses which assist people to develop critical thinking skills. It is recommended that other assessments are used in combination with W-GCTA in order to have a complete picture of an individual and their areas for development (e.g. personality assessments and 360 feedback). A specific W-GCTA development report is available.

**In an outplacement and career guidance context**, the W-GCTA might be appropriate for someone facing redundancy, a change of circumstances, or seeking an alternative role or profession. The purpose of the assessment process is to provide a wide perspective on suitable career paths and to help individuals to choose options which best suit their abilities, needs and interests. This can help people develop an awareness of their own potential.

### Selecting the Appropriate Test Version

Once it has been determined that the W-GCTA is an appropriate test to use for a particular scenario it is necessary to decide which version to use. There are both Supervised (fixed form paper and pencil test) and Unsupervised (online test) versions of the test available, both consisting of 40 items, administered over a 30 minute period.

***When to use W-GCTA Supervised***

The fixed supervised version is appropriate where testing will take place at the tester's premises under full supervision. The Supervised version can be used to verify scores from an unsupervised testing, or as a stand alone test.

***When to use W-GCTA Unsupervised (online)***

This unsupervised, online administration of the W-GCTA is suitable for selection or development contexts where administration is required and an administrator is not available or required. Time and therefore cost of administration is significantly reduced. The system delivers a (almost) randomly generated test consisting of 40 questions from a pool (or 'item bank') of 310 questions.

Unsupervised online testing can be the most convenient approach in the early stages of recruitment. Candidates can take the test in their own location without the need for travel which adds time and cost to the process. The system can generate scores and feedback reports as soon as the candidate has completed the test. Because different tests are generated for each candidate, it is not possible for the answer key for a test to be known in advance. However, when unsupervised testing is used it is strongly recommended that candidates are retested under secure conditions later in the selection process. This can be done using another online test but administered under supervision, or using the W-GCTA Supervised. A verification process has been designed to help with this process.
When candidates know they will be retested they are much less likely to attempt to cheat when they take the test in the first instance. In addition, if a candidate did try to cheat, the second testing would show whether their ability was at the required standard or not.

***Issues in Unsupervised Testing***

There are a number of issues with the use of unsupervised tests, particularly when used in high stakes settings i.e. for selection purposes. These have been raised and discussed by various experts in the field and a number of guidelines for the use of unsupervised testing have been published:

- International Test Commission:

http://www.intestcom.org/_guidelines/guidelines/index.html

- The British Psychological Society's (BPS) Psychological Testing Centre:

http://www.psychtesting.org.uk/blog$.cfm/2010/11/4/New-technologybased-testing-guidelines-produced

Some topics are discussed below in reference to the W-GCTA:

**Technology Issues**

Our platform is reliable and stable to deliver online assessments. If test-takers lose connection while taking a test, their responses and time taken are saved. The next time they login they can therefore continue the test where they left it and the timer will also continue from this point.

Different connection speeds are also accounted for. The timer stops whilst the next question page is downloading and restarts again once test-takers can see the question. The platform supports most internet browsers including those most commonly used - Internet Explorer 7, Firefox, Chrome and Safari. No additional hardware of software downloads are required to run the test. An advice team are available to provide technical (IT) support to test-takers and administrators.

***Quality Issues***

Response equivalence for a large number of questions has been compared when taken online and paper-and-pencil. No significant differences were found indicating that both administration formats are equivalent. Equivalence has also been found between the unsupervised, online pool of items and the paper-and-pencil version of the test.

***Control Issues***

We recommend a "controlled mode" of administration for the unsupervised, online version. This means that no direct human supervision is required but that the test is only made available to known test-takers (i.e. that the test link is sent to their e-mail address). Each link sent by the platform can only be entered once. We strongly insist on a "managed mode" of administration for the supervised (online and paper-and-pencil) version. This means that there is a high level of supervision and control over the testing conditions (i.e. that the test is taken in a testing or assessment centre). The pool of questions will be updated regularly and exposure of questions will be monitored. Any questions that have been over exposed will be removed and replaced.

***Security / Privacy Issues***

The test is held on a secure server. Test-takers and administrators will require a username and password to login to the platform. Only those with the BPS Level A certificate will be able to login to administer the online test and view the results. Test-takers are asked to agree to an 'honesty contract' where they are asked to confirm that they will take the test unassisted and without misusing the test. The pool of items is large enough for over one trillion different tests to be generated. Therefore, no two test-takers are likely to take exactly the same items. This enhances the security of the test and questions in the pool.

***Verification of Scores***

Where initial testing is unsupervised, users may wish to validate the score later. This can be done by administering the test again at a supervised testing session. The random nature of the online test means that candidates will receive a different form of the test. If online testing is not available, paper and pencil administration of the W-GCTA Supervised, a secure test version, may be used.

Scores will always differ somewhat from one test administration to another. The standard error of measurement provides an estimate of how much scores can be expected to differ. For the purposes of verification we wish to identify pairs of scores which are very unlikely to be attained by the same

candidate. In particular, when the second score is substantially lower than the initial score users may wish to flag the score for further consideration. However, it should be remembered that even when the probability of a candidate genuinely attaining a much lower score at second testing is low, this does happen, particularly when the test taker is experiencing high levels of anxiety.

For the purposes of verification, a lower score on second testing with a difference magnitude greater than that attained by 95% of other candidates is flagged. Only 2.5% of genuine candidates would score in this range. Such an unusual score merits further consideration.

Users can ask candidates to take a second test online under supervision, in which case the system will indicate whether a verification flag has been generated for a large difference. Alternatively, where a paper and pencil verification test is used, the verification score can be entered into the system to determine if a verification flag has been generated.

***Procedure for developing flagging criteria***

The average score difference will depend on the standard error of measurement which will be larger at the extremes of the ability distribution than in the middle. This is because only a minority of test items will have high information at these ability levels. In order to take this into account in the procedure, the score difference for flagging was set separately for 10 tranches of ability, based on trial participants scores on all items they attempted.

From each of 2 pilot versions, two equivalent 40 item test forms were generated with no overlap. Pilot participants were scored on each form and the score difference between the scores for the 2 forms for each candidate was calculated. For each ability tranche the 97.5$^{th}$ percentile difference was calculated for the test forms. The average value across the two trial versions was calculated and these values were smoothed to provide the criterion difference to initiate flagging.

This provides a criterion score for each initial score below which second testing scores are flagged. This theta score is translated into the equivalent W-GCTA Supervised raw score using the standard look up table to provide equivalent criteria when the paper and pencil administration is used.

# Administering the W-GCTA

The W-GCTA can be administered in a supervised or unsupervised setting. It is also available online or in paper-and-pencil format. The differences in administration are described below.

## Supervised Testing Session

### *Preparation for a Supervised Testing Session*

- Ensure you are familiar with your organisation's Testing Policy.
- Schedule the testing sessions. Consider the duration of the session, the number of people to be tested, book an appropriate room, trained test administrators and additional invigilators (we recommend a ratio of at least 20:1 test takers to invigilators).
- Ensure the organisation has sufficient test materials in stock. To order materials contact Pearson TalentLens customer services. If the test is to be administered online under supervision check that you have an appropriate account with Pearson TalentLens and that you have access to W-GCTA on the platform.
- Invite test takers to the testing session. Inform them about the nature of tests, including how and why they are being used, the date, time, location and whether they are required to bring anything with them (e.g. some testing centres require personal identification to be checked).
- Distribute W-GCTA familiarisation materials. It is good practice and enhances fairness to ensure that all test takers are provided with familiarisation materials, either in booklet form or via an internet link.
- Find out about disabilities and special requirements that any test takers have. Arrangements should be made to accommodate these. You should not change the standardised test administration procedure without taking advice from an expert as this can change the meaning of the scores. Contact Pearson TalentLens for advice if you are unsure about making accommodations.
- Prepare the test log, this can act as a register and detail any reasonable adjustments to be made for candidates with disabilities as well as any unusual occurrences.

### *Setting up the Testing Session*

- Ensure all Administrators and Invigilators have appropriate training and are familiar with the W-GCTA.
- Ensure a suitable room for testing, considering size, space, layout, lighting, temperature, noise and possible distractions. Test takers should be seated apart, but not directly opposite each other to avoid cheating and distraction. Ensure that potential disturbances are be minimised, e.g. phones are unplugged, 'Testing in Progress' signs are used.
- If online administration will be used, check that all equipment is in working order and that the candidates have been added to the system.

***Conducting the Testing Sessions***

The testing session must be standardised to provide test takers with the same opportunity for doing well. It is advised that the instructions are closely followed. Try to engender a friendly but purposeful atmosphere to put test takers at ease and enable them to work at their best. Start with an informal introduction to the testing session.

- An introduction should include:
  - Who you are.
  - Your relationship to the organisation.
  - Purpose of testing.
  - How the results will be used.
  - Who will have access to the results?
  - Storage of the results (data protection).
  - What will happen after the testing?
  - The logistics of the testing session: breaks, fire alarms expected, duration, toilets.
  - Give candidates the opportunity to ask questions.
- Ensure all mobile phones and electrical equipment are turned off and all candidates are ready to start the session.
- On starting the session, ask them to maintain silence from this point on. They should raise their hand if they have any questions.

| Online Administration | Paper and Pencil Administration |
|---|---|
| • Ensure the computer is showing the initial test administration screen.<br>• See Appendix B<br>• Appropriate instructions and test administration will be automatically delivered by the system | • Follow standardised instructions (see appendix D)<br>• Ensure the testing session is precisely timed. The start time should always be written on the test log.<br>• Ensure that test takers are completing the Record Forms appropriately<br>• Ensure that all test materials are collected before anyone leaves the room. |

- Thank the test takers for attending and inform them of the next steps of assessment or process.
- Complete a test log.
- Securely store test materials following session.
- Ensure that data protection is followed.

***Scoring the test***

- Ensure that you know which norm group is to be used and what type of scale scores are to be reported on (percentiles, sten scores etc).

| Online Administration | Paper and Pencil Administration |
|---|---|
| • Standardised scoring is managed by the platform | • Follow BPS good practice guidelines for test scoring<br>• Follow the scoring instructions from Pearson TalentLens (these can be found in Appendix E)<br>• Bureau scoring service is available from TalentLens |

**Unsupervised Testing Session**

***Preparation for an unsupervised Testing Session***

***Preparing for the Testing Session***

- Familiarise yourself with the test format and platform.
- E-mail or speak to the candidate to provide all information they require (e.g. purpose of test, confidentiality, administered online, how feedback is provided).
- Find out about any disabilities or special requirements that test takers may have (the online timer can be turned off if required).

***Setting up the Testing Session***

- Ensure you have an appropriate account with TalentLens and that you have access to
- W-GCTA on the platform.
- Check your understanding of how the platform works.
- Ensure you have the correct e-mail address for the test-taker.

***Conducting the Testing Session***

- Add the test-taker to the platform and send them the automatic e-mail invite from the platform, which can be amended to suit. A sample email invitation text can be found in Appendix A
- Standardised administration is managed by the platform.

***Scoring the test***

- Standardised scoring is managed by the platform.

***Interpreting scores***

- A profile report is automatically generated from the platform containing a t-score and percentile for each norm group available.
- Norm tables allowing conversion from raw score to t-score and percentiles are available.

***Verification of Scores***

Where initial testing is unsupervised, users may wish to validate the score later. This can be done by administering the test again at a supervised testing session. The random nature of the online test means that candidates will receive a different form of the test. If online testing is not available, paper and pencil administration of the W-GCTA Supervised, a secure test version, may be used.

For the purposes of verification, a lower score on second testing with a difference magnitude greater than that attained by 95% of other candidates is flagged. Only 2.5% of genuine candidates would score in this range. Such an unusual score merits further consideration.

Users can ask candidates to take a second test online under supervision, in which case the system will indicate whether a verification flag has been generated for a large difference. Alternatively where a paper and pencil verification test is used, the verification score can be entered into the system to determine if a verification flag has been generated.

## Interpretation of W-GCTA scores

Test scores should be interpreted in the given context, against appropriate norm groups and related to additional information. It is important to relate the scores to the purpose of testing, making appropriate connections between the test score and what this actually means in terms of a career or a specific job. Test scores provide an indication of an individual's performance against that of a group of others who have taken the test. To allow ease in the comparison of this individual against others, test scores are standardised.

Following online administration (supervised or unsupervised) of the W-GCTA you will receive an automatic profile report that contains the test-taker's percentile score and t-score. Because there are many different forms of the test, each with minor variations in difficulty, a raw score is not available. Instead the scoring algorithms, based on Item Response Theory (IRT) take into account the exact difficulty level of the items each person completed. The test-taker's score will automatically be compared to the UK General Population norm group. You may wish to compare the score to an alternative norm group.

Computer generated reports are available to aid the interpretation of test results. These are:

- A profile report containing the test-taker's score against all norm groups
- An interview report with questions based on the test-taker's score at the subscale level. Interview questions will change depending on these subscores.
- A development report for use in development interventions. The report shows how leverage strengths in critical thinking and create a development plan.

Report Samples are available on the TalentLens website – www.talentlens.co.uk.

### Choosing a Norm Group

Test scores are compared with a norm group to aid interpretation. A norm group provides a meaningful standard for interpreting test scores. Based on a frequency distribution of scores from those who have previously taken the test, the norm group shows how the individual has performed relative to an appropriate comparison group and translates the score onto a standardised scale with known properties. Organisations can use published norms or create their own norm group.

#### *Published Norms*

Scores can be interpreted using the most appropriate pre-existing norm groups given. Published norms allow an organisation to benchmark the performance of their employees or job applicants to that of others for a particular role (e.g. Senior Managers).

***What norm groups are available?***

A variety of different norm groups are available for the W-GCTA. Many groups are available across the different test versions so that scores can be compared. The current norm groups recommended for ***selection*** purposes are:

- UK General Population
- Graduates
- Graduates from Law, Business, Economics or Finance
- Public Sector Graduates
- Private Sector Graduates
- Managers
- Public Sector Managers
- Private Sector Managers
- Senior Management

The norm groups available for ***development*** purposes are:

- Development and Outplacement
- Public Sector High Flyer Development Programmes

Further details of these norm groups are provided in Appendix F.

***Local Norms***

When an organisation is testing many people it can be more appropriate to create a local norm group reflecting the performance of applicants or incumbents in the organisation. A local or in-house norm group must be based on a sufficiently large group of people (ideally, at least 150) who are representative of the people being assessed. If there is insufficient data to create a local norm or the group is unrepresentative in some way (e.g. available scores come from graduate recruits and you wish to assess more experienced managers) it may be preferable to use a more general published norm group.

It is important that the comparison group used should be appropriate for the test use. Where possible the comparison group should be taking the test for a similar purpose (e.g. selection, development). For career guidance the most general norm groups may be most appropriate as these allow the person to benchmark their skills more broadly. Alternatively the score could be compared with a number of different norm groups to show how the person's ability matches up to different types of job and industry sectors.

For selection the norm group should as far as possible reflect the selection context. In particular the comparison group should be applying for roles at a similar level although industry sector and job type are also important. Where the job level is not clear, typical educational background may provide an indication of this. Lastly, where possible the norms should reflect the diversity of the applicant sample with respect to gender, age, ethnicity etc.

If you would like to create your own norm group for comparison purposes, please get in touch with Pearson TalentLens.

**Case Study**
A law firm has decided to use the W-GCTA in selecting professional staff. In particular they wish to use it for selecting for entry level positions. They have between 4 and 7 vacancies each year and receive around 250 applications. Because of the large number of applications they have decided to use the Unsupervised version for an initial sifting process before inviting successful candidates to an assessment centre. They have not used the test before so do not have any data to create their own norm group so in the first year they choose the *Graduates from Law, Business, Economics or Finance* norm as most similar to their organisation and role. After the first year of use, they contact Pearson TalentLens to create their own norm group from the data collected during

**Interpreting scores**

Test scores should be interpreted within the context for which the test is being used. Scores should be compared to appropriate norm groups which will show how the individual's ability measures up to that of others in a relevant comparison group. It is important to integrate test results with other assessment information that is collected. Inferences from the tests scores should be related to the purpose of testing, and the implications for job performance or career choice clarified.

Standardised scales are used to compare an individual's performance on a test against a comparison group. The advantage of a standardised scale is that a standardised score has a consistent meaning across different tests and comparison groups. Below the standardised scales used for the W-GCTA are described.

***T scores***

The T scores are most frequently used with ability measures. The T score scale has an average score of 50 and a standard deviation of 10. Higher scores indicate better performance. When scores are normally distributed 67% of test takers will score between T scores of 40 and 60. The advantage of T scores is that they represent an even scale – that is, the difference between scores of 70 and 80 is the same as the difference between scores of 45 and 55. In addition it is possible to apply the standard error of measurement to a T score to allow for a band of error around a score. It is possible to add and subtract T scores and to correlate them with other measures.

They provide a good level of differentiation for ability tests with enough points on the scale to represent all the different score levels.
Generally, T scores should not be used in feedback to untrained people, including the test taker, as they can be difficult to comprehend without some understanding of statistics.

***Percentiles***

Percentile equivalent scores are often used when giving feedback to test takers. These have the advantage of being readily understood and allow test takers to understand how they have done in comparison to others. It is important not to confuse percentiles with percentages. A percentile is the percentage of test takers who score lower than a given score. This means that a test taker who scores at the 70th percentile has scored higher than 70 percent of the comparison group. A score at the 30th percentile is better than 30% of the comparison group.

Percentiles are not equal units. They show the relative position or ranking of each test taker in comparison to the norm group, but do not illustrate the amount of difference between scores. In a normal distribution, cases will be clustered more closely at the centre of the distribution than at the extremes. Differences at this mid-point are therefore more exaggerated while those at the extremes are relatively understated. For this reason it is not appropriate to sum or correlate percentiles with other scores.

***Banded or graded scores***

To simplify scores further they may be banded or graded into the following categories:

1 – Well above average performance, 90$^{th}$ percentile and above
2 – Above average performance, 70$^{th}$ – 90$^{th}$ percentiles,
3 – Average, 30$^{th}$ – 70$^{th}$ percentiles,
4 – Below average performance, 10$^{th}$ – 30$^{th}$ percentiles,
5 – Well below average performance, below 10$^{th}$ percentile.

***Stens and Stanines***

The Sten scores scale is a standardised 10-point-scale ranging from 1 to 10, with a mean of 5.5 and a standard deviation of 2. Stanine scores are a slight variation on stens, ranging from 1 to 9, with a mean of 5 and a standard deviation of 2. Both stens and stanines are commonly used in feedback. Higher scores indicate better performance. Like T scores they are an even scale but the smaller range is often easier to understand.

**Accuracy of test scores**

Scores obtained on the W-GCTA and any other psychological test can only be considered an estimate of the test takers true score. This occurs as no test is perfectly accurate (without error). The standard error of measurement (SEM) indicates the amount of error to be expected in a test takers score. It can be expressed in raw score points or in standardised scale points but not in percentile points.

The standard error of measurement for the W-GCTA test forms is provided in the chapter on reliability. Using the SEM allows a band or confidence interval to be constructed around a score. Scores should be interpreted as bands, rather than precise points taking into account the potential for error.

Banding of scores serves to check against over-emphasising small differences between scores. The standard error of measurement (SEM) can be used to create a band for a score when expressed as a raw score or a T score. A score should be considered to fall within a band from one SEM's below the score to one SEM above the score. This can be useful when providing feedback to test takers.

**Example**
Femi has a T score of 45, John has T score of 52 and Juliet has a T score of 55 on the W-GCTA Unsupervised. The T-score SEM for this version is 4.1. Therefore Femi's score can be considered to lie between 40.9 and 49.1. John's ranges between 47.9 and 56.1 and Juliette's should be considered to be between 50.9 and 59.1. The SEDiff is 1.414 x 4.1 = 5.8. Therefore T-score differences of more than 6 points can generally be considered to indicate a real difference. We can conclude that Femi has a lower score than both Juliet and John but that John's score is too close to Juliet's to say that it is lower than hers.

**Example**
An organisation recruiting a senior manager held an assessment centre which consisted of the W-GCTA (UK – 80 item), a personality questionnaire and feedback interview, and case study exercise. Two candidates scored 75 and 71 on the W-GCTA (UK – 80 item). The SEDiff of the W-GCTA (UK – 80 item) is 4.8. The recruitment manager decided that these scores were not significantly different. The recruitment manager was confident that applicants did not differ in terms of their critical thinking skills but was reassured to know that both applicants met the minimum level.

***Comparing scores between test takers***

The standard error of difference (SEDiff) can be used when comparing scores from different test takers or the same test taker on different tests. When comparing the performance of two individuals on one test the standard error of difference is used to judge whether the scores are significantly different.

The standard error of difference is equivalent to the 1.414 x the standard error of measurement for comparing two scores on the same test.

The standard error of difference can also be used when comparing scores from different tests. In this case it is calculated as the square root of the sum of the squares of the SEM of the tests being compared. This would be the appropriate error band to use when comparing scores from the different subtests of the W-GCTA.

**Example**
Tim has a W-GCTA(UK) T score of 60 and T scores of 65, 60 and 54 on Recognising Assumptions, Evaluating Arguments and Draw Conclusions respectively. The standard error of difference for comparing the test scores are 6.5 for RA and EA, 5.6 for RA and DC and 6.2 for EA and DC. In this case only the difference between RA and DC exceeds the appropriate SEDiff and therefore we can conclude that Tim is better at Recognising Assumptions than Drawing Conclusions but that there is no difference between his ability to Evaluate Arguments and either his ability to Recognise Assumptions or Draw Correct conclusions.

***Limitations of test scores***

Test scores should be interpreted carefully. Errors may arise from the administration of the testing session and scoring. Scores can also be affected by a test takers state, for example anxiety or felling unwell. Candidates with a disability or with English as a second language may be disadvantaged due to the test format. For these reasons scores should be explored carefully and interpreted with caution.

On occasion test scores may contradict alternative information on a test taker. In this case, the test user should work with the test taker to explore the information and discover possible causes for these anomalies.

Tests have been carefully standardised. Any changes to this process can result in unreliable test scores. Used correctly psychometric tests are a powerful tool that can provide important information on the test taker, but for these reasons tests are designed to be used alongside alternative assessment methods.

# Feedback guide

It is best practice to provide test takers with appropriate feedback on their performance in a psychometric test or assessment process and feedback increases the perceived acceptability of the test.

Feedback is an essential process to inform test takers of their test performance and the implication of this in a language they can understand. Information given should be fair, accurate and understandable and any questions the test taker has should be clearly answered. Providing feedback to the test taker can be a sensitive process as some people have emotional reactions to information about their strengths and weaknesses.

Feedback can be written, face to face or over the telephone. Written feedback may be appropriate where there is a large number of test takers and face to face or telephone feedback will not be feasible or cost effective.
Pearson TalentLens are able to provide computer generated reports for the W-GCTA.

In preparation for feedback, the qualified person providing feedback should:

- Consult the test log to establish if there were any problems or interruptions that may have affected test performance. Interpretation of test scores assumes standardisation of test conditions, and so a fair and accurate assessment will rely on this.
- Ensure that scores are translated into the appropriate standard scores or percentiles using a relevant norm group.
- Ensure they have a clear understanding of the relevance of the W-GCTA in the context it has been used.

### Using appropriate scores in feedback

The approach to feedback should be adopted that produces the most helpful outcome from the test taker's point of view. Test takers should not be made to feel uncomfortable during the feedback, but it is important that they receive a realistic understanding of their performance.

Scores should be used that are accessible to the test taker. The most commonly used scores for feedback include percentiles, stens and stanines.

Feedback should allow test takers to understand the purpose of the assessment, the relevance of assessment and how well they performed in the test. This should follow three steps:

1. Describe the tests used and the purpose of the assessment, for example:

   *"Your recently completed the W-GCTA as part of the recruitment process. The W-GCTA consists of five components or subtests that are designed to assess verbal reasoning that is your ability to critically evaluate written information";*

2. Describe the individuals results in context to the comparison group, say what the group is, why it has been selected, why it is relevant and how the individual's performance compares to the group,

   *For example, you may say to a test taker*

   *"Compared with the broad based representative group from the working population who have completed this test, your score was at the 80th percentile. That is you did better than 80% of the other test takers who took it. Only 20% of test takers did better than you."*

   *Or,*

   *"Compared with others who have applied for the position of underwriter in this organisation, your score was at the 45th percentile. That is you did better than 45% of other test takers who taken the test for this purpose"*

   Or

   *"Your score fell in Band 3 or Grade C. This means that your performance can be classified as average in the group of firemen with which you were compared."*

3. Describe relevance of the scores for the purpose in which they are being used.

   *For example,*

   *"Critical thinking skills are crucial to the role of XXX. The role holder will be required to use these skills to analyse written information from time to time in the job."*

Good feedback should:

- put the test taker at ease,
- be pitched at the appropriate level for the Test Takers knowledge of psychometric assessment, and the W-GCTA test,
- provide relevant information about the test,
- describe the group against which the test taker is being compared,
- describe performance in relation to that group,
- avoid technical terms or jargon,

- place test results in context of other information gained during assessment,
- provide test takers with the opportunity to ask questions,
- be a positive experience for test takers, where information is related to their needs.

Feedback should always be meaningful to the test taker. In selection, this is reporting how the applicant performed in comparison with the norm group used. In development or guidance, in some cases it can be useful to compare the test takers score against a range of norms to ensure that it is place in context and fully understood.

For example, a candidate with a WGCTA$^{UK}$ raw score of 53 is classified as well-below average in comparison with UK Banking applicants (5th percentile), but average compared to the UK general population (34th percentile). Without this additional information the test taker may be misinformed about their level of critical thinking.

A specific interview report is available to assist interviewers with questions following administration of the W-GCTA for selection.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A.. & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, N.J.: Prentice Hall.

Bernstein, I. H., Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105,* 467-477.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Deary, I. J., Whalley, L. J., Lemmon, H., Crawford, J. R., & Starr, J. M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish mental survey. *Intelligence, 28*, 49–55.

Facione, P. A. (1990). *Executive summary: The Delphi report*. Millbrae, CA: California Academic Press.

Fischer, S. C., & Spiker, V. A. (2000). *A model of critical thinking*. Report prepared for the U.S. Army Research Institute.

Gadzella, B. M., Baloglu, M., & Stephens, R. (2002). Prediction of GPA with educational psychology grades and critical thinking scores. *Education, 122(3*), 618–623.

Gadzella, B. M., Stephens, R., & Stacks, J (2004). *Assessment of critical thinking scores in relation with psychology and GPA for education majors*. Paper presented at the Texas A & M University Assessment Conference, College Station, TX.

Geisinger, K. F. (1998). Review of Watson-Glaser Critical Thinking Appraisal. In J. C. Impara & B. S.

Plake (Eds.), *The thirteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking*. Mahwah, N.J. Lawrence Erlbaum.

Hough, L.M., Oswald, F.L. and Ployhart, R.E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9(112*), 152 – 194.

Impelman, K., & Graham, H. (2009). *Interactive effects of openness to experience and cognitive ability*. Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Klaczynski, P. A., Gordon, D. H., & Fauth, J. (1997). Goal oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology, 89*, 470–485.

Kudish, J. D., & Hoffman, B. J. (2002, October). *Examining the relationship between assessment center final dimension ratings and external measures of cognitive ability and personality*. Paper presented at the 30th International Congress on Assessment Center Methods, Pittsburgh, PA.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220.

Paul, R. W., & Elder, L. (2002). *Critical thinking: Tools for taking charge of your professional and personal life*. Upper Saddle River, NJ: Financial Times Prentice Hall.

Robertson, I.T., & Smith, M. (2001). Personnel Selection. *Journal of Occupational and Organisational Psychology, 74 (4)*, 441-472.

Salgado, J., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalisation of GMA and cognitive abilities. *Personnel Psychology, 56*, 573-605.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, F., & Hunter, J. (2004). General mental ability in the world of work. *Journal of Personality and Social Psychology, 86*, 163-173.

Sa, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*, 497–510.

Stanovich, K.E., & West, R.F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342–357.

Stanovich, K.E., & West, R.F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672–695.

Spector, P. A., Schneider, J. R., Vance, C. A., & Hezlett, S. A. (2000). The relation of cognitive ability and personality traits to assessment center performance. *Journal of Applied Social Psychology, 30(7)*, 1474– 1491.

Taube, K. T. (1995, April). *Critical thinking ability and disposition as factors of performance on a written critical thinking test*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

U.S. Department of Labor. (1999). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author.

Watson, G. (1925). *The measurement of fairmindedness*. *Contributions to Education, No. 176*. New York: Burea of Publications, Teachers College, Columbia University.

Watson, G., & Glaser, E. M. (1952). Watson-Glaser Critical Thinking Appraisal manual. New York: Harcourt, Brace, & World.

Watson, G. & Glaser, E.M. (1964). *Watson Glaser Critical Thinking Appraisal Manual for Forms Ym and Zm*. New York: Harcourt Brace Jovanovitch.

Watson, G., & Glaser, E. M. (1980). Watson-Glaser Critical Thinking Appraisal, Forms A and B manual. San Antonio, TX: The Psychological Corporation.

Watson, G., & Glaser, E. M. (1991). *Watson Glaser Critical Thinking Appraisal. British Manual Forms A, B and C*. London: The Psychological Corporation.

Watson, G., & Glaser, E. M. (1994). Watson-Glaser Critical Thinking Appraisal, Form S manual. San Antonio, TX: The Psychological Corporation.

Watson, G., & Glaser, E. M. (2006). Watson-Glaser Critical Thinking Appraisal, Short Form manual. San Antonio, TX: Pearson.

Watson, G., & Glaser, E. M. (2009). Watson-Glaser II Critical Thinking Appraisal, Technical Manual and User's Guide. San Antonio, TX: Pearson.

Watson, G., & Glaser, E. M. (2010). Watson-Glaser II Critical Thinking Appraisal, Technical Manual and User's Guide. San Antonio, TX: Pearson.

Watson, G., Glaser, E.M., & Rust, J. (2002). Watson-Glaser Critical Thinking Appraisal UK Edition. London: Pearson Education Limited.

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100, 930–941*.

# Appendix A

## Example Unsupervised Administration Email Invitation

Dear XXX,

You have been invited by xxxxxxx to complete an online psychometric instrument, the Watson-Glaser Critical Thinking Appraisal test as part of the recruitment process for xxxx roles.

**Before clicking on the test link at the end of this email, please read this information carefully**

You will have **up to 30 minutes** to complete the test once you start. Make sure you are in a quiet room free from distractions and that you give yourself time to complete the questionnaire without being disturbed. The test has 40 questions.

As soon as you have clicked on the **Start button** the time will begin. **You will not be able to pause the test and come back to it once you have started.**

When you have completed question 40, click on **next** then click on the **finish** button. You will then be taken to the demographic questions (a 'prefer not to say' option is provided). You do need to **proceed to the end of these** and click the **submit** button.

You should complete the test on your own without assistance from others. You will be retested with another equivalent version of the test if asked to the later stages of the recruitment process. **(Optional)**

You will be given full instructions on how to complete the test once you have accessed the online testing website by clicking on the link below.

**Why have I been asked to take the test?**

You have been invited to complete the Watson Glaser Critical Thinking Appraisal test. Critical thinking is the ability to identify and analyse problems and seek and evaluate relevant information in order to reach appropriate conclusions. It is important for decision making and career success in general. xxxxx uses the test as part of their selection process to assess your suitability for the role.

Your results will remain strictly confidential and will only be seen by xxxxxx recruitment team. All results will be stored in accordance with the Data Protection Act. It is our policy to store all test results for six months, after which they will be made anonymous and used for monitoring and research purposes only *(as per your test user policy)* .

Please contact xxxxxxxx if you have any special requirements for completing the test. It is good practice to offer feedback on your test results, should you require this, and xxxxxx will be able to advise you on how you can obtain feedback.

**When you're ready to complete the test, click on the link below**:

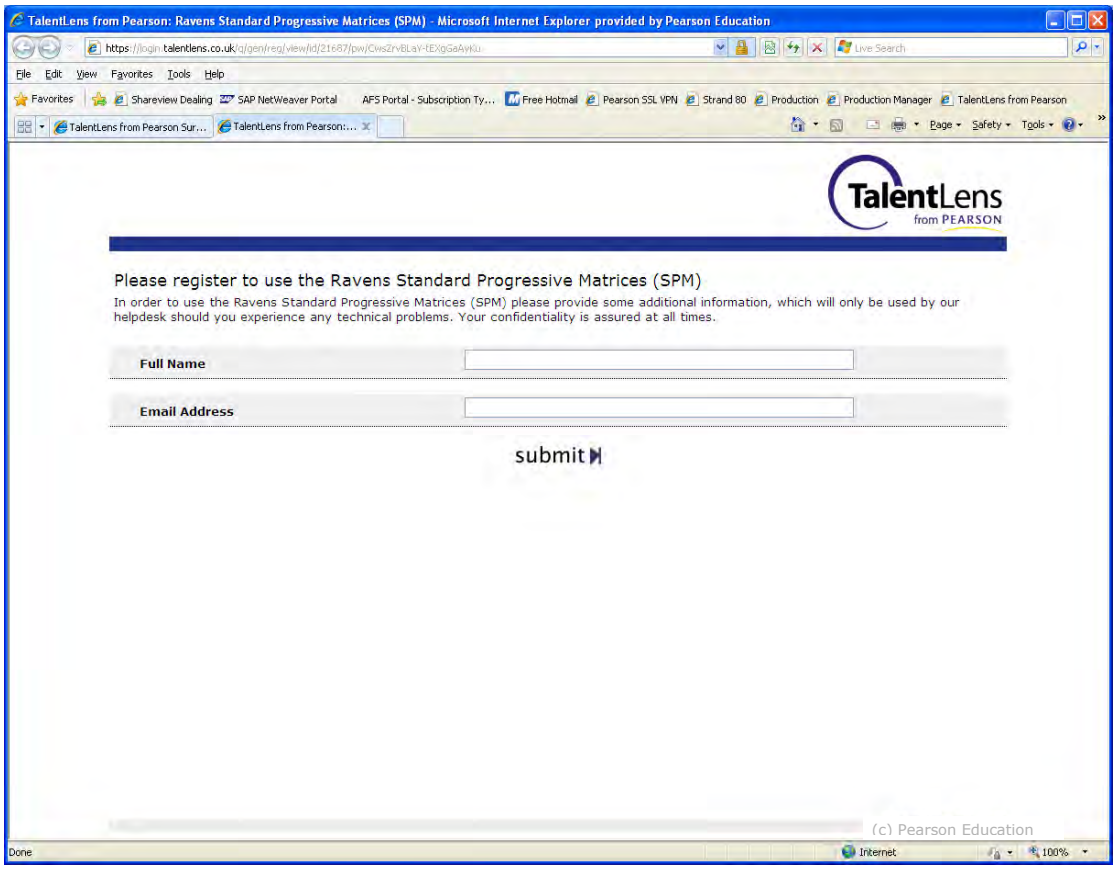
 

# Appendix B

## Supervised Online Administration Instructions

Once your PPU survey/assessment is created, copy the PPU URL to the browser of the computer(s) to be used for supervised testing.

If testing is to be anonymous, test takers will be taken to the beginning of the survey/assessment. If test takers are required to self-register for the survey/assessment, they will see the following page.




**Supervised Test Administrator Introduction:**

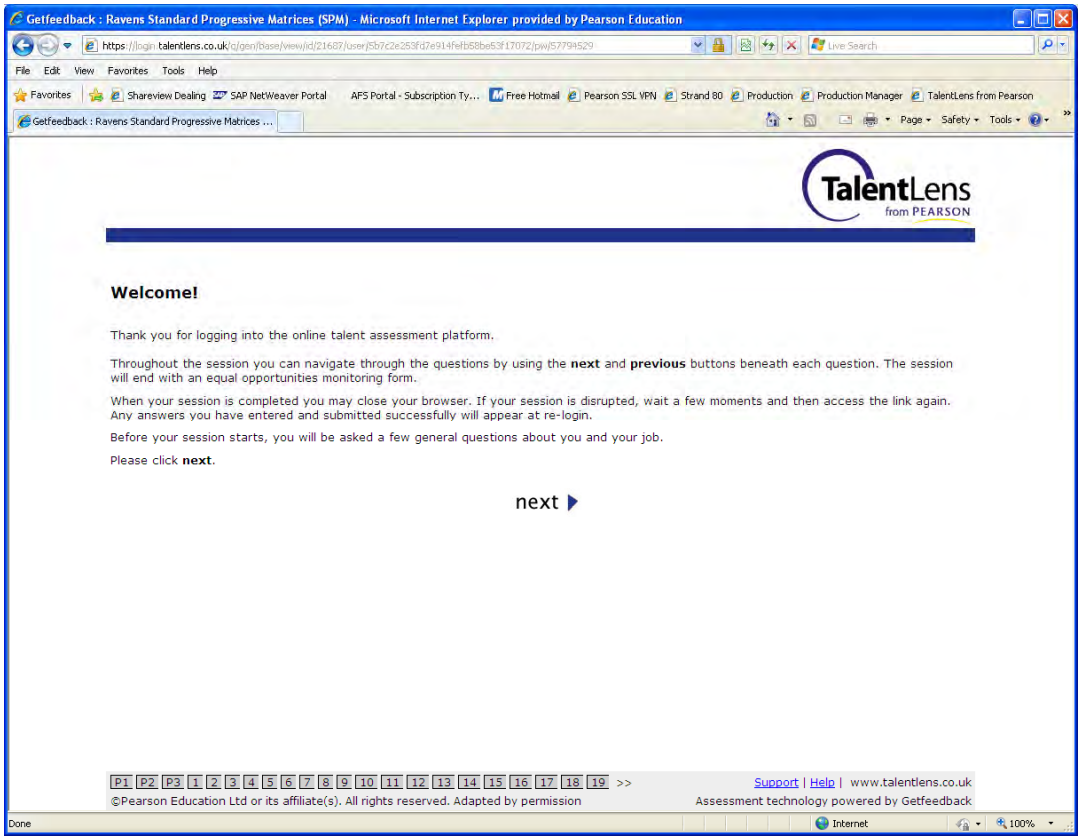
After a short welcome introduction say:
**To sign on please enter you name and your email address in the boxes provided, and then click submit.**

When all candidates have signed on say:
**The onscreen directions will take you through the entire process, which begins with a welcome page and some general questions about you. At the end of the test there are a few more general questions.**

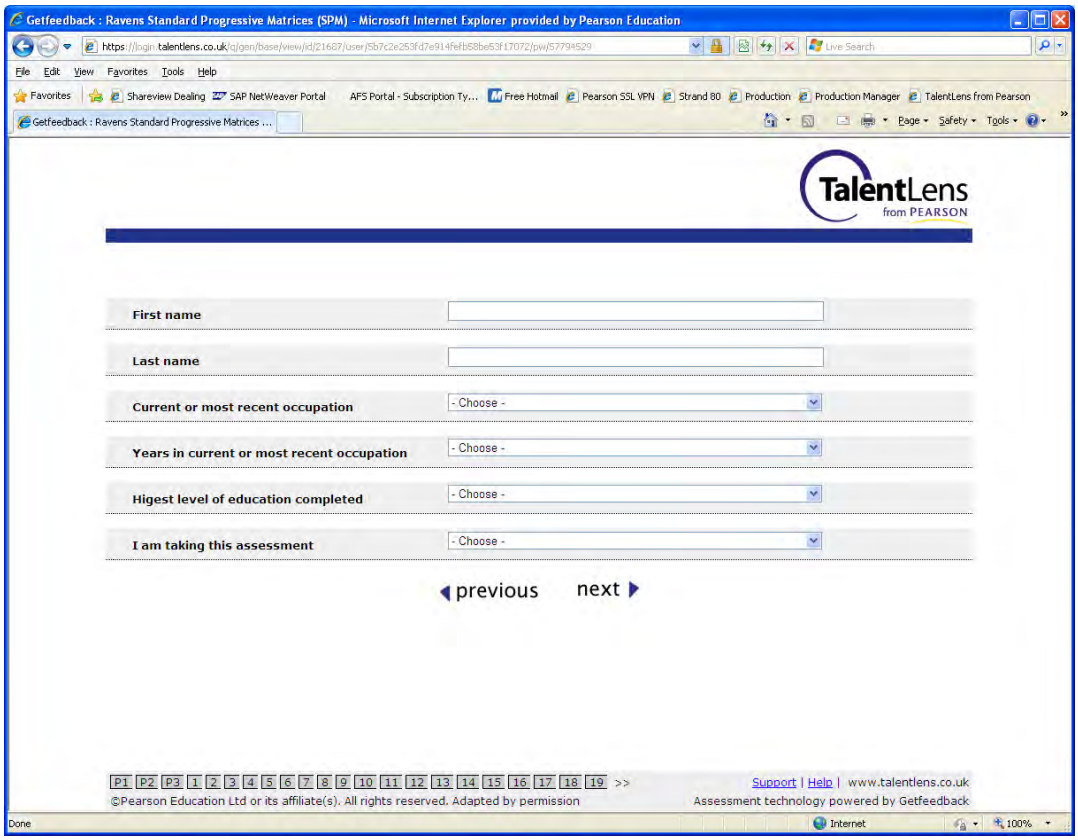
While all candidates are completing the general questions ask:

**Do you have any questions before you click next to start the assessment?** (Ensure all candidates have completed the onscreen general questions and are ready to begin the test)

All surveys/assessments begin with a generic welcome instruction

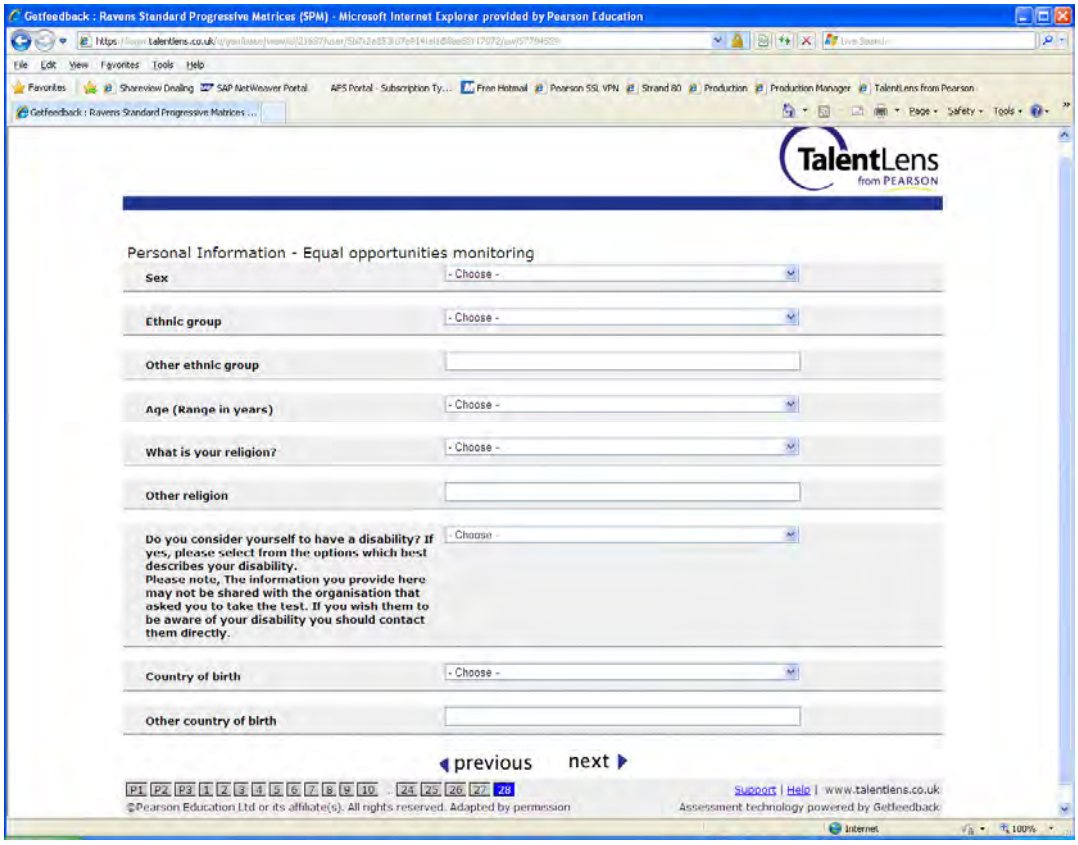



Pre assessment general questions – all fields are required.



Back end general questions – are all optional. All data is anonymous and strictly confidential.



# Appendix C

## Test Log

A test log should always be maintained. An example is provided below. This page may be photocopied for use of the W-GCTA in your organisation.

***Test Log***

| | |
|---|---|
| **Organisation** | |
| **Purpose of Testing** | Selection / Development / Appraisal |
| **Test(s) Used** | WGCTA$^{UK}$/ W-GCTA Supervised |
| **Administration** | Paper and Pencil / On-line |
| **Test Administrator** | |
| **Test Invigilators** | |
| **Date** | |
| **Start time** | |
| **Finish time** | |

**Candidate List**

| | **Candidate Name** | | **Candidate Name** |
|---|---|---|---|
| 1. | | 6. | |
| 2. | | 7. | |
| 3. | | 8. | |
| 4. | | 9. | |
| 5. | | 10. | |

**Materials Checklist**

| **Materials** | **Number checked out** | **Number checked in** |
|---|---|---|
| WGCTA$^{UK}$ Test Booklets | | |
| WGCTA$^{UK}$ Record Forms | | |
| W-GCTA Supervised Test Booklets | | |
| W-GCTA Supervised Record Forms | | |
| Test Instructions | | |
| Stopwatches | | |
| Pencils | | |
| Erasers | | |
| Pencil Sharpeners | | |

**Disturbances / Unusual Occurrences**

# Appendix D

## W-GCTA Supervised Paper and Pencil Test Administration Instructions

### Test Administration Instructions

Following an introduction to the testing session, say:

**From now on, please do not talk among yourselves, but ask if anything is not clear.**

Distribute the Test Booklets and say:

**Do not open these booklets until you are told to do so.**

Then distribute the Record Forms and say:

**Please complete the candidate details on the first side of this form.**

If you are collecting equal opportunities data, say:

**Equal opportunities data are collected to monitor fairness in testing. Completion of this section is optional.**

Otherwise say:

**You do not need to complete the equal opportunities section.**

Allow the test takers time to complete the details on the front of the Record Form. Then say:

**In this test, all the questions are in the Test Booklet. There are five (5) separate parts to the test in the booklet and each one is preceded by its own directions and examples, which should be read carefully.**

**For each question, decide what you think is the best answer. As your score will be derived from the number of items you answer correctly, try to answer each question even if you are not sure if the answer is correct. Record your choice by putting a cross in the appropriate place on the Record Form. Always be sure that the answer space has the same number as the question in the Test Booklet. Do not make any other marks on the Record Form. If you change your mind about an answer make sure that you rub out the first mark completely. Do not spend too much time on any one question. When you finish a page, go straight on to the next one, working through each of the tests in turn.**

**If you finish all of the tests before the time is up, you may go back and check your answers.**

Say:

**You will have 30 minutes to work on the test. Now open your Test Booklet and read the directions on the first page.**

After allowing time for test takers to read the directions, say:

**Are there any questions about what you are to do?**

Answer any questions, preferably by re-reading the appropriate sections of the directions, then say:

**Ready? ... Begin.**

Immediately start your timing procedure. If any of the test takers finish before the end of the test period, either tell them to sit quietly until everybody has finished or collect their materials and dismiss them quietly.

While the group is taking the test, move about the room making sure that each test taker is marking the Record Form properly.

At the end of 30 minutes, say:

**Stop! Put your pencils down. This is the end of the test.**

### Concluding Administration

At the end of the testing session, collect the Test Booklets, Record Forms and pencils and thank everyone for attending.

The W-G Supervised is a demanding test to take. The style of the items in some of the subtests makes it difficult for test takers to achieve a confident appreciation of their performance in the test. From this point of view it can be an uncomfortable experience and some words of reassurance at this point may be appropriate. It may be constructive to clarify the contribution of the test within the context of other aspects of selection or appraisal procedures. It would also be constructive to reassure test takers regarding the confidentiality of test scores.

# Appendix E

## Scoring Instructions

### *Scoring the W-GCTA*

If you require the bureau service please contact Pearson TalentLens.

### *Step 1 – Check Record Forms*

1. Check each Record Form to ensure there are no multiple responses to the same item, missed items, or partly erased answers (where test takers have changed their response).
2. Where any multiple responses or missed items are found, these should be crossed out with a coloured line that will show through the acetate key.

### *Step 2 – Obtaining the raw score*

1. Place the acetate scoring keys over the Record Form.
2. For each subtest count the correctly marked spaces.
3. Record each subtest total in Box 1 of the 'Test Score Summary' table.
4. Add the subtest scores to create a total raw score.
5. Record the total raw score in the 'Test Score Summary' table.
6. Transfer this total raw score to Box A on the Record Form cover.

Note:

*The W-GCTA awards one point for every correct answer.*

*Do not count items as correct when more than one response has been selected, even though the right answer is one of those marked.*

*The maximum W-GCTA Supervised total raw score is 40.*

### *Step 3 – Converting raw scores into standardised scores using published norms*

Norm Group conversion tables are downloadable free of charge to TalentLens registered users from the restricted resource area at www.talentlens.co.uk

# Appendix F

## Norm Groups

The current norm groups recommended for ***selection*** purposes and general use are:

- UK General Population
- Graduates
- Graduates from Law, Business, Economics or Finance
- Public Sector Graduates
- Private Sector Graduates
- Managers
- Public Sector Managers
- Private Sector Managers
- Senior Management

The norm groups available specifically for ***development*** purposes are:

- Development and Outplacement
- Public Sector High Flyer Development Programmes

Norm tables are downloadable free of charge to TalentLens registered users from the restricted resource area at www.talentlens.co.uk

| Norm Group | Gender and Ethnicity Breakdown |
|---|---|
| **UK General Population**<br>N. 1546<br><br>This group consists of a broad sample of the general population including over 50 different occupations from a range of organisations within the UK.<br><br>Age ranged from 17 to 72 years.<br><br>84% reported their educational level; 53% held a degree or higher professional qualification, 26% A-Levels and 20% GCSE's.<br><br>The data were collected throughout 2001 and 2002 for the W-GCTA UK and RANRA standardisation. | One third of the group were female.<br><br>87% were White, 7% Asian, 4% Black and 2% Other. |

| | |
|---|---|
| **Graduates**<br>N. 2900<br><br>Over 90% of this group were job applicants, predominantly applying for graduate, managerial and professional roles. The group includes a mix of public and private sector applicants, however, a large proportion were applicants to a central government department. This group were screened prior to this testing with a competency-based assessment, and had at least one year's managerial experience. Similar information is not available for the rest of the group.<br><br>Age ranged from 20 to 72 years; 75% of the group were under 32.<br><br>55% of the group reported their religion, approximately half described themselves as<br>Christian, one third reported no religion, just under 10% described themselves as Muslim.<br><br>55% of the group reported their sexual orientation; 97% described themselves as heterosexual 1% of the group described themselves as intersex, transgender or transsexual.<br><br>The majority of this data were collected between 2000 and 2008, and the remainder accumulated from 1993 onward.<br><br>W-GCTA Test Format: predominantly the W-GCTA UK edition with some data from Form C. | Half of the group were female.<br><br>Over 90% reported their ethnicity; 65% were White, 14% Asian, 10% Black. |
| **Managers**<br>N. 3562<br><br>This group consists of managers from both the public and private sector. A large proportion of the group were applicants to a central government department. These applicants were screened prior to testing with a competency-based assessment, and had at least one-year's managerial experience. Similar information is not available for the rest of the group.<br><br>Background data were available for three-quarters of the group.<br><br>Age ranged from 19 to 66 years. Two-thirds were between 25 and 50.<br>Almost 80% of the group reported their religion; approximately half described themselves as Christian, one-third reported no religion, just under 10% described themselves as Muslim.<br><br>76% reported their sexual orientation; 97% were heterosexual, 1.7% reported themselves to be intersex, transgender, or transsexual.<br><br>Data on education was available for 70% of the group; 95% held a degree level qualification or higher.<br><br>Data were collected predominantly between 2000 and 2008, with some data collected since 1993.<br><br>W-CGTA Test Format: predominantly the W-GCTA UK edition with some data from Form C. | Half the group were female.<br><br>71% were White, 12% Asian and 10% Black. |

| | |
|---|---|
| **Senior Management**<br>N. 256<br><br>This group consists of senior managers, senior executives and directors from various organisations and industries, both public and private sector.<br><br>Age ranged between 25 and 65 years. Around three-quarters of the group were<br>32 years or above.<br><br>Data were collected between 1993 and 2007.<br><br>W-GCTA test format: W-GCTA UK and Form C. | Data on gender were available for 97% of the group. 89% were male.<br><br>Data on ethnicity were available for 35% of the group. 99% were White. |
| **UK Private Sector Graduates**<br>N. 189<br><br>This group consists of graduate job applicants. They were either already employed in, or applying for roles in, the private sector.<br><br>Age ranged from 21 to 56 years; two-thirds of the group were under 30.<br><br>The majority of this data were collected between 2000 and 2008, and the remainder accumulated from 1993 onward.<br><br>W-GCTA Test Format: predominantly the W-GCTAUK edition with some data from Form C. | One-third of the group were female.<br><br>88% were White, 8% were Asian. |
| **UK Public Sector Graduates**<br>N.1926.<br><br>Over 90% of this group were job applicants, predominantly applying for graduate, managerial and professional roles within the public sector and a large proportion were applicants to a central government department. This group were screened prior to this testing with a competency-based assessment, and had at least one year's managerial experience. Similar information is not available for the rest of the group.<br><br>Age ranged from 20 to 72 years; 75% were under 32.<br><br>84% reported their religion; approximately half described themselves as Christian, one-third reported no religion, and just under 10% described themselves as Muslim.<br><br>Over 80% reported their sexual orientation; 97% were heterosexual, 1.5% of the group described themselves as intersex, transgender, or transsexual.<br><br>The majority of this data were collected between 2000 and 2008, and the remainder accumulated from 1993 onward.<br>W-GCTA Test Format: predominantly the W-GCTAUK edition with some data from Form C. | Just over half of the group were female.<br><br>88% reported their ethnicity; 66% White, 14% Asian, 11% Black |

| | |
|---|---|
| **Graduates in Law, Business, Economic or Finance**<br>N. 530<br><br>This group consists of graduate job applicants to the public sector achieving a 2.1 or higher in law, business, economics or finance.<br><br>Aged ranged from 21 to 60 years. 70% were below 30.<br><br>93% reported their religion; just over half were Christian, almost one quarter<br>were of no religion.<br><br>92% reported their sexual orientation, 98% were heterosexual. Just over 2% of the group described themselves as Intersex, Transgender or<br>Transsexual.<br><br>Data were collected 2006.<br><br>W-CGTA Test Format: W-GCTA UK edition. | Just over half were female.<br><br>98% reported their ethnicity.<br>Just over half were White, one-fifth Asian and almost one-fifth Black. |
| **Public Sector Managers**<br>N. 2694<br><br>This group consists of managers in the public sector. A large proportion of applicants were to a central government department. These applicants were screened prior to this testing with a competency-based assessment, and had at least one-year's managerial experience. Similar information is not available for the rest of this group.<br><br>Age ranged from 20 to 65 years; three quarters of the group were aged between 20 and 40.<br><br>Almost 80% reported their religion; approximately half described themselves as<br>Christian, one-third reported no religion, and just under 10% described themselves as Muslim.<br><br>77% reported their sexual orientation; 97% were heterosexual, 2% of the group reported themselves to be intersex, transgender or transsexual.<br><br>Data were collected predominantly between 2000 and 2008, with some data collected since 1993.<br><br>W-CGTA Test Format: predominantly the W-GCTA UK edition with some data from Form C. | Just over half were female.<br><br>Ethnicity data were available for 90% of the group; 70% were White, 12% Asian and 10% Black. |

| | |
|---|---|
| **Private Sector Managers**<br>N. 725<br><br>This group consists of managers working for a range of organisations in the private sector.<br><br>Background data were available for 14% of the sample.<br><br>Age ranged between 21 and 56; 80% were aged between 20 and 40 years.<br><br>Data was collected 1993 and 2007.<br><br>W-CGTA Test Format: predominantly Form C with some data from the WGCTA UK Ed. | 88% were male.<br><br>99% were White. |

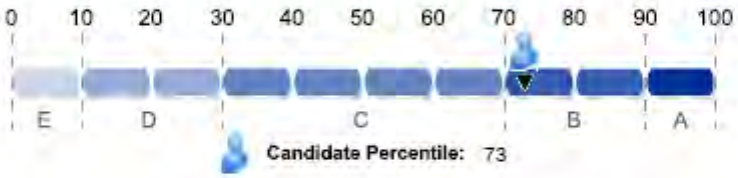
# Appendix G

## Example W-GCTA Assessment Report

### Overall Performance on First Test

TalentLens from PEARSON

**Candidate:** sample@pearson.com Report
**Norm Group:** UK General Population



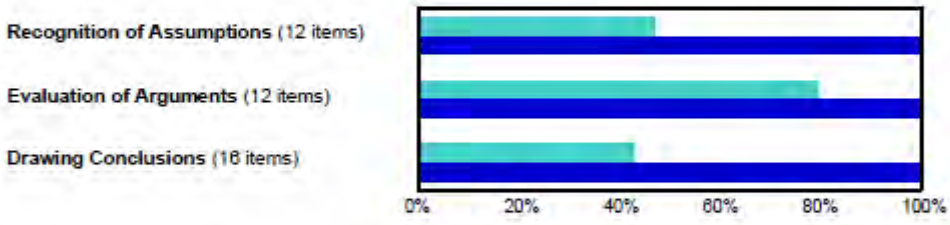

**A - Well above average** (91st percentile and above)
**B - Above average** (71st – 90th percentiles)
**C - Average** (31st – 70th percentiles)
**D - Below average** (11th – 30th percentiles)
**E - Well below average** (10th percentile and below)

The bar above shows overall performance on the W-GCTA UK in comparison to UK General Population who have previously completed the test.

**Number of Questions Answered: 40**
**T-Score: 56**
**Time Taken: 27 mins**

### Subtest Performance on First Test

The subtests of the W-GCTA UK can be considered in 3 clusters of critical thinking. In the graph below, the light coloured bar shows the candidate percentile on each cluster **compared to the UK General Population** . The dark coloured bar shows the percentage number of questions answered.




**Subtest scores can provide useful information in a development or guidance context, but should not be used for recruitment and selection. It is recommended that W-GCTA UK should be used in combination with other assessment techniques.**

name sample@pearson.com Report page 2



## First Test Performance against other norm groups

Norm groups available for ***selection*** purposes :

| Comparison Group | Percentile |
|---|---|
| • UK General Population | 73 |
| • Graduates | 58 |
| • Managers | 60 |
| • Senior Management | 48 |
| • UK Private Sector Graduates | 56 |
| • UK Public Sector Graduates | 56 |
| • Graduates in Law, Business, Economic or Finance | 64 |
| • Public Sector Managers | 58 |
| • Private Sector Managers | 67 |
| • Executive Search candidates (UK) | 42 |

Norm groups available for ***development*** purposes :

| Comparison Group | Percentile |
|---|---|
| • Development and Outplacement | 55 |
| • Public Sector High Flyer Development Programmes | 55 |

## Re-test Results

Following completion of the online, unsupervised version of the Watson-Glaser test, this candidate has taken the test again but under supervised conditions.

The second test score indicates that there is a statistical difference between the first and second test score. Therefore, you should speak to the candidate to understand why this discrepancy has occurred. Please remember that cheating is only one possible reason and there are many other potential reasons e.g. stress, illness or disruption during the second test.

A development report is also available. You may wish to provide this for the test-taker as it contains more detailed feedback on their test score and tips on how to further develop their critical thinking skills.



