Watson-Glaser™ III

Critical Thinking Appraisal

Goodwin Watson and Edward M. Glaser

User's Guide and Technical Manual

Pearson | TalentLens

For more information, contact us at **TalentLens.com.**

Table of Contents

List of Figures

List of Tables

Introduction

The Watson-Glaser is a test of critical thinking and reasoning. Critical thinking can be defined as the ability to identify and analyze problems, as well as seek and evaluate relevant information in order to reach an appropriate conclusion. The Watson-Glaser is used in organizations as an employee selection and development tool, and in academic settings as a selection tool and a measure of gains in critical thinking skills.

The third edition of the Watson-Glaser™ III (W-G III) is an online, item-banked assessment. Each administration of the test includes 40 items that are randomly selected from a large pool of scenarios and questions. This is a much more secure way of administering tests online in an unsupervised environment, because it is highly unlikely for test takers to have seen the items before or to complete the same items as someone else taking the test. The new item-banked Watson-Glaser III is scored using item-response theory (IRT). This scoring methodology adapts for minor differences in difficulty between test versions so that scores are equivalent.

Watson-Glaser III can be administered unsupervised or under supervision for greater control over test conditions, candidate identity, and behavior. Examinee data are immediately captured and automatically scored with user-selected norms, and computer-generated interpretive reports are immediately available to the administrator.

The Watson-Glaser III meets 21st century testing needs and retains its robustness as a measure of critical thinking ability. Watson-Glaser II continues to be available as a supervised, paper-and-pencil alternative. The Watson-Glaser II has good psychometric equivalence to the Watson-Glaser III. Though the Watson-Glaser II can be administered online, it is a fixed-form test and must always be supervised to maintain test security. Additional information about the development and administration is included in the Watson-Glaser II (2010) Technical Manual and User's Guide.

Versions of the Watson-Glaser are available in multiple languages including United Kingdom (UK) and United States (U.S.) English, French, French Canadian, Dutch, Spanish, and Latin American Spanish.

Measuring Critical Thinking Skills

The Watson-Glaser measures the fundamental cognitive ability of critical thinking. Critical thinking is an organized 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.

The Watson-Glaser is a multi-faceted measure of critical thinking. Three subscales comprise the test and five subtests comprise the subscales. The questions are of varying difficulty and format to measure all areas of critical thinking ability. The five subtests, listed below, require different, though interdependent, applications of analytical reasoning in a verbal context.

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

Recognize Assumptions: Identifying unstated assumptions or presuppositions underlying given statements.

Deduction: Determining whether conclusions follow logically from given information.

Interpretation: Weighing evidence and deciding if generalizations or conclusions based on data are warranted.

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

Each 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 school, and in newspaper or magazine articles. The scenarios and possible responses require critical evaluation and cannot be accepted without question. The scenarios are neutral or 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 with controversial content refer to political, economic, and social issues that frequently provoke emotional responses. Strong attitudes, opinions, and biases affect the ability of some people to think critically (Stanovich & West, 2008; West, Tolplak, & Stanovich, 2008).

Watson-Glaser 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 is designed to be administered with a generous time limit. It is appropriate for use with both general and high ability populations including those who have completed higher education. An untimed version is available in the U.S., but the timed version is most appropriate for use in high-stakes decision-making.

The RED Model: Keys to Critical Thinking

The Watson-Glaser second and third editions include one notable update to the original test. Factor analyses of previous versions of the instrument (Forms Short, A, and B) consistently revealed a structure in which three of the five subtests, Inference, Deduction, and Interpretation—all related to drawing conclusions—factored together. Recognize Assumptions and Evaluate Arguments remained independent factors. Watson-Glaser II and III scores reflect this factoring, as illustrated in the RED model of critical thinking (Watson & Glaser, 2009) in Figure 1.

Figure 1. The RED Model

The five Watson-Glaser subtests map directly to these areas. **R**ecognize Assumptions and **E**valuate Arguments are two of the five subtests that comprise the test. Inference, Interpretation, and Deduction subtests comprise **D**raw Conclusions. Scores are reported for the three subscales that the five subtests comprise.

Recognize 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 recognize 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 analyze such assertions objectively and accurately. Analyzing 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-generalization beyond the evidence.

Suggested Applications

The Watson-Glaser may be used in employment and educational contexts to help with selection, development, and career counseling. The use of the test is restricted to individuals with an appropriate qualification in test use (e.g., relevant training or credentialing).

Selection

The Watson-Glaser can be used to predict success in jobs that require critical thinking skills. Many organizations use 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 Watson-Glaser may be used to rank order applicants or in combination with other assessment methods to provide a full profile of an applicant.

Critical thinking is also essential in higher education settings, and critical thinking assessments can help identify students who are likely to be successful in programs where this skill is important. Research in a variety of settings, including nursing and psychology programs and law school, supports the potential usefulness of the Watson-Glaser for this purpose.

Development

Tests can be helpful in better understanding a person's strengths and weaknesses so that appropriate development goals and activities can be set. The Watson-Glaser 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 minimize 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 that need development. There are also training courses to help people develop critical thinking skills. Awareness of one's limitations allows a person to take actions to 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.

Outplacement and Career Guidance

The Watson-Glaser 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 his or her current role or profession and seeking an alternative. The purpose of the assessment process is to provide a broad perspective on suitable career paths and to help individuals choose options that best suit their own abilities, needs, and interests.

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

Training in Critical Thinking

Critical thinking is often taught in business and educational settings. The Watson-Glaser may be used to assess the extent to which students in these courses have mastered critical thinking skills.

Administering the Watson-Glaser

Only individuals who have been trained and certified in the administration and use of ability measures in an occupational or higher education setting may administer the Watson-Glaser and interpret the results.

Using Watson-Glaser III

Unsupervised, online administration of the Watson-Glaser III 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. This is often the most convenient approach in the early stages of recruitment for employment or career development. Candidates can take the test in their own location, saving time and cost. When you choose unsupervised testing, Pearson recommends that you retest candidates under secure conditions, later in the selection process. This can be done with a supervised, online test administration of the Watson-Glaser III or II, or the Watson-Glaser II paper-and pencil forms. When candidates know they will be retested, they are much less likely to try cheating during the first instance. If a candidate did try to cheat, the second testing would show whether his or her ability was at the required standard or not.

Using Watson-Glaser II

Watson-Glaser II can be administered online or with paper and pencil. This is a fixed-form test and is appropriate only when testing is supervised. Watson-Glaser II also can be used to verify scores from an unsupervised testing or as a stand-alone test.

Testing Considerations

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 online and unsupervised testing have been published.

- The International Test Commission provides several globally relevant and useful publications focused on test use, computer-based testing, and information for test takers:

  https://www.intestcom.org/files/guideline_computer_based_testing.pdf

  https://www.intestcom.org/files/guideline_test_use.pdf

  https://www.intestcom.org/files/test_taker_guide_brochure.pdf

- Many countries and professional organizations provide their own guidelines as well, such as:
  - The British Psychological Society's (BPS) Psychological Testing Centre: http://ptc.bps.org.uk/
  - The American Psychological Association (APA) commissioned a task force on issues in Internet testing (including unsupervised administration), and the results can be found here: https://www.apa.org/science/programs/testing/testing-on-internet.pdf
  - Government of Canada – Personnel Psychology Centre: https://www.canada.ca/en/public-service-commission/services/public-service-hiring-guides/best-practices-unsupervised-testing.html
  - The Australian Psychological Society: https://thereselanders.files.wordpress.com/2013/07/psychological-testing.pdf

Browser Requirements

The platform Pearson uses to deliver the Watson-Glaser III is reliable and stable. If test-takers lose connection during the test, their responses, and time taken are saved. The next time they log in, they can continue the test where they left off, and the timer resumes from this point.

Because different systems have different connection speeds, the timer stops while 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. The following Internet browsers are compatible with the assessment delivery platform:

- Internet Explorer® 8.0 or higher
- Edge®
- Firefox® (latest version, must use auto-update)
- Google Chrome™ (latest version, must use auto-update)
- Safari® (Mac 5.0+)

No additional hardware or software downloads are required to run the test. If you or your clients have difficulties with administration, please contact your local TalentLens office (see talentlens.com).

Equivalence

Online and paper-and-pencil responses to a large number of questions have been compared for equivalence. Because no significant differences were found, both administration formats are considered equivalent.

Security/Privacy

The Watson-Glaser online test is held on a secure server. Test-takers and administrators require a username and password to log into the platform. Qualified users must log in to administer the online test and view the results. For Watson-Glaser III unsupervised administrations, test-takers must agree to an “honesty contract” and confirm that they will take the test unassisted and not misuse the test. The pool of items is large enough to generate a very large number of tests (not taking into account the test constraints that are in place). Therefore, no two test-takers are likely to take exactly the same items.

Verifying Scores

When initial testing is unsupervised, administrators may wish to validate an examinee’s score by having him or her take the test again in a supervised session. If online testing is not available, the Watson-Glaser II paper-and-pencil test may be administered in a supervised setting.

Scores typically 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 and helps users identify pairs of scores that are very unlikely to be attained by the same candidate. When the second score is substantially lower than the initial score, administrators may wish to flag the score report for further consideration. Even when the probability of a candidate genuinely attaining a much lower score on the second test is low, it can happen occasionally, particularly if the examinee is experiencing high levels of anxiety.

Reading Ability Differences and English as a Second Language

Watson-Glaser directions and items are written at or below the ninth grade reading level. However, reasonable precautions must be taken when assessing candidates whose first language is not English. When possible, the Watson-Glaser should be administered in the examinee's first language. Contact TalentLens for information on language versions available. If a version is not available in the examinee's first language and he or she has difficulty with the language or the reading level of the items, note this and consider it when interpreting the scores.

Accommodating Examinees With Disabilities

The U.S. Americans with Disabilities Act (ADA) of 1990 requires an employer to reasonably accommodate the known disability of a qualified applicant, provided such accommodation would not cause an "undue hardship" to the operation of the employer's business. Reasonable accommodations enable candidates with special needs to comfortably take the test. Reasonable accommodations may include, but are not limited to, modification of the assessment format and procedure, such as live assistance, in which an intermediary reads the test content to a visually impaired candidate and marks the candidate's answers (Society for Industrial and Organizational Psychology, 2003). Consult with your qualified legal advisor or human resource professional for additional guidance on providing appropriate, reasonable accommodations.

Administration Directions

The Watson-Glaser 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.

The online version is administered through Pearson's online testing platform, an Internet-based testing system designed for the administration, scoring, and reporting of professional assessments. Candidates' responses are instantly processed and the system immediately generates an interpretive report upon completion of the test. The TalentLens online product catalog technical information section includes a variety of useful materials:

- This manual
- Sample reports
- Frequently asked questions
- Norm composition tables (descriptions of the samples used to generate the norms)

Before administering the test, make sure any test administrators are familiar with the administration instructions. If you are not familiar with the Watson-Glaser, a good way to prepare is to take the test prior to administration, being sure to comply with the directions and any time requirement. Doing so will help anticipate questions or issues that may arise during test administration.

Unsupervised Testing

Preparing for Testing

- Ensure you have an appropriate TalentLens account and that you have access to Watson-Glaser on the platform.
- Familiarize yourself with the test format and platform by taking the test and reading this User's Guide.
- Choose the relevant normative group for score reports.

Setting Up the Testing Session

- Ensure you have the correct email address for the test-taker.
- Email or speak to the candidate to provide all the information he or she needs (e.g., purpose of test, confidentiality, administered online, how feedback is provided).
- Ask the candidate if he or she needs any special accommodations to take the test (e.g., an untimed administration).

Conducting the Testing Session

- Add the test-taker to the platform. The platform generates the email invitation, which you can amend as needed.
- The platform administers the test according to the standardized procedure.

Generating Reports

- Responses are processed and score reports are immediately available to you via the platform.
- A Profile Report is automatically generated and available to the administrator and includes percentiles and standardized scores based on the norm group selected.

Verifying Scores

Where initial testing is unsupervised, users may wish to validate the score by administering the test again in a supervised 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, the Watson-Glaser II (paper-and-pencil, Forms D or E) may be used.

Scores always differ somewhat from one test administration to another. When the second score is substantially lower than the initial score, users may wish to flag the score for further consideration. Remember, even when the probability of a candidate obtaining a much lower score at second testing is low, it does happen; particularly when the test taker is experiencing high anxiety.

Supervised Online Administration

Preparing for Online Testing

Being thoroughly prepared before an examinee's arrival results in a more efficient supervised administration session. Examinees do not need pencils or scratch paper for the test. Do not allow examinees to access to any reference materials (e.g., dictionaries or calculators).

Administrators

- Ensure you are familiar with your organization's Testing Policy.
- Ensure you have an appropriate TalentLens account and that you have access to Watson-Glaser on the platform.
- Familiarize yourself with the test format and platform by taking the test and reading this User's Guide. Doing so enables you to anticipate and answer questions more efficiently.
- Choose the relevant normative group for score reports.

Setting Up the Testing Session

- Schedule the testing sessions. Consider how long the session takes, how many people will be tested, where the testing will take place, and the number of trained test administrators available. (It is recommend that test sessions include no more than 10 test takers per administrator.)

- In the email invitation to test takers:
  - Inform them about the nature of test, including how and why the test is being used.
  - Include the date, time, location.
  - Let them know if they need to bring anything with them (e.g., personal identification for check-in).
  - Ask the candidates to let you know if they require any special accommodations for taking the test.
- Do not change the standardized test administration procedures without seeking advice from an expert. Changing procedures voids the use of normative data. Contact TalentLens Customer Service if you are unsure about making accommodations.
- Prepare a test log. This can function as a register and detail any reasonable adjustments made for candidates with disabilities, as well as any unusual occurrences. See Appendix A for a sample test log.

Preparing the Testing Room

- Ensure all administrators have appropriate training and are familiar with the test.
- Post a “Testing in Progress” sign outside the testing room.
- Make sure the testing room is a suitable size and layout, has adequate lighting, is a comfortable temperature, and free of noise and possible distractions. Seat test-takers far enough apart, but not directly opposite each other, to avoid cheating and distraction. Ensure that potential disturbances are minimized.
- Check all computer equipment to ensure that it is in working order and that the candidates have been added to the system.
- All computer stations used to administer Watson-Glaser II online must be in locations that can be easily supervised.

Conducting the Testing Session

Test administrators must follow all standardized testing directions so that all candidates/test takers have the same opportunity to do well. Try to create a friendly, but purposeful, atmosphere to put test takers at ease and enable them to do their best. Start the testing session by introducing yourself.

- Tell the test takers:
  - Who you are
  - Your relationship to the organization
  - Purpose of testing
  - How the results will be used
  - Who will have access to the results
  - Storage of the results (data security)
  - What will happen after the testing
  - The logistics of the testing session: breaks, fire alarms expected, duration, toilets
- Ask examinees to turn off mobile phones and any other portable communication devices before starting the test.
- Ensure the computer is showing the initial test administration screen.

After the initial instruction screen for the Watson-Glaser displays and the examinee is seated, say,

The on-screen directions take you through the entire process. Do you have any questions before you begin? If you need to ask a question during the test, raise your hand. Please maintain silence during the test.

After answering any questions, say, **Please begin the test.**

Technical Issues

If an examinee's computer develops technical problems during testing, move the examinee to another suitable computer. If the technical problems cannot be solved by moving to another computer, please contact your local TalentLens office (see talentlens.com).

The contact information, including phone numbers, are included on the TalentLens.com website.

When Testing is Done

- Thank the test takers for attending and inform them of the next steps in the process.
- Complete the test log.
- Secure all test materials after the session. Avoid disclosure of test access information, such as usernames and passwords. All the computer stations used to administer Watson-Glaser online must be in locations that can be easily supervised.
- Ensure that data protection is followed. It is unethical and poor test practice to allow test score/report access to individuals who do not have a legitimate need for the information.

Generating Reports

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

Supervised Watson-Glaser II Paper-and-Pencil Testing

Materials Needed for Testing

- This manual
- 1 Test Booklet per examinee
- 1 Answer Sheet per examinee
- Two No. 2 pencils with erasers per examinee
- A clock or stopwatch if the test is timed
- Hand-Scoring Key if the test will be hand-scored

Intended as a test of critical thinking power, rather than speed, the Watson-Glaser II may be timed or untimed. A 30-minute time limit is recommended for timed administrations. Most examinees can complete the test, working at a reasonably comfortable pace, in 30 minutes. To facilitate accurate timing, record the starting time and the finishing time immediately after you have given the signal to begin or end. Allow 5–10 minutes to read the directions on the cover of the test booklet and answer questions.

Answering Questions

Examinees may ask questions about the test before you give the signal to begin. To maintain standard testing conditions, answer such questions by rereading the appropriate section of the directions. Do not volunteer new explanations or examples. It is the responsibility of the test administrator to ensure that examinees understand the correct way to indicate their answers on the Answer Sheet and what is required of them. The question period should never be rushed or omitted.

If any examinees have routine questions after the testing has started, try to answer them without disturbing the other examinees. However, explaining the meaning of words or items to candidates must be avoided, as this could lead to inappropriate prompting of candidate responses. If candidates have questions about the interpretation of an item, they should be encouraged to respond to the item as they best understand it.

Administering the Test

All directions that the test administrator reads aloud are in **bold type.** Read the directions exactly as they are written, using a natural tone and manner. Do not shorten the directions or change them in any way. If you make a mistake in reading a direction, say, **No that is wrong. Listen again.** Then read the direction correctly.

Scripted Instructions:

When all examinees are seated, give each examinee two pencils and an Answer Sheet.

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.

Scoring Watson-Glaser II Paper-and-Pencil

Appendix B provides instructions for hand scoring the Watson-Glaser II Paper-and-Pencil.

Test Security

Watson-Glaser II scores and reports are confidential and should be stored in a secure location accessible only to authorized individuals. It is unethical and poor test practice to allow test score/report access to individuals who do not have a legitimate need for the information.

The security of testing materials and protection of copyright must also be maintained by authorized individuals. Storing test scores and materials in a locked cabinet (or password-protected file in the case of scores maintained electronically) that can only be accessed by designated test administrators is an effective means to ensure their security.

Interpreting Watson-Glaser Results

Watson-Glaser results are generated as a Profile Report that typically is available within a minute after the test has been completed. A link to the report is available at your administrator's account in Pearson's online testing platform. Adobe® Acrobat Reader® is required to view and download the report. The administrator may view, print, or save an individual's report. A Development Report also is available for use in developmental interventions. This report shows how to leverage strengths in critical thinking and create a development plan. Report samples are available on the TalentLens website and in the online catalog—www.talentlens.com.

Choosing a Norm Group

A variety of different norm groups are available for the Watson-Glaser, some based on Occupational groups, some based on Position Type/Level, and some based on Educational Background. The U.S. norm groups are below. For UK norm groups see Appendix C.

Occupational Norm Groups

- Accountant
- Consultant
- Engineer
- Human Resource Professional
- Information Technology Professional
- Sales Representative

Position Type/Level Norm Groups

- Executive
- Director
- Manager
- Supervisor
- Professional/Individual Contributor
- Hourly/Entry Level
- Manger in Manufacturing/Production

Educational Background Norm Groups

- High school diploma (or GED)
- 1–2 years of college
- 3–4 years of college
- Bachelor's degree
- Graduate degree (Master's and Doctoral)

Further details for these norm groups are included in the online product catalog, in the technical information section. Norm groups available for other countries are included there, as well.

Using Local Norms

When an organization is testing many people, it can be more appropriate to create a local norm group reflecting the performance of applicants or incumbents in the organization. A local or in-house norm group must be based on a sufficiently large group of people (ideally, at least 200) who are representative of the people being assessed. If there is insufficient data to create a local norm group 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 his or her skills more broadly.

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. 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.

Interpreting Watson-Glaser III Scores

Test scores should be interpreted within the context for which the test is being used. An individual's scores should be compared to the appropriate norm group to obtain an accurate profile of his or her ability. It is important to integrate test results with other assessment information collected.

Interpreting the Watson-Glaser Profile Report

Following an online administration of the Watson-Glaser II or III, the administrator can access the individual's Profile Report. Results may be used to inform selection decisions, as well as build an examinee's awareness of his or her own strengths and development needs. To better understand the constructs measured, the report provides a detailed definition of each of the skill (i.e., Recognize Assumptions, Evaluate Arguments, and Draw Conclusions).

The examinee's results are reported as percentiles and as graphs. Brief interpretive summaries are provided at the overall and subscale level. For selection purposes, use only the overall score—it provides a more precise and complete picture of a candidate's critical thinking skills.

Percentile Scores

Both the Watson-Glaser II and III Profile Reports provide the overall and subscale scores in percentiles. Percentile scores often are used in feedback to test takers. Percentile scores are easily understood and enable test takers to understand how they have done in comparison to others. 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, however, are not equal units. They show the test taker's relative position or ranking in comparison to the norm group, but do not illustrate the amount of difference between scores. In a normal distribution, cases cluster more closely at the center of distribution than at the extremes. Differences are more exaggerated at mid-point while those at the extremes are relatively understated. For this reason, it is not appropriate to add or correlate percentiles with other scores. Watson-Glaser II Profile Reports include a raw score, which is the number correct out of a possible 40. However, the Watson-Glaser III raw score is *not* simply the number correct.

Watson-Glaser III Scores

The Watson-Glaser III Profile Report has an Additional Technical Information section that includes the Number Correct, *T*-, Sten, and Stanine scores. Item-banked tests produce a *theta* score that incorporates the difficulty level of each item. Therefore, the number of correct responses should not be used to make decisions. Use percentiles or one of the standardized scores that account for the minor differences in difficulty between items in different administrations.

***T*-scores.** *T*-scores are used most frequently 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.

T-scores 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 shared as feedback to untrained people or test takers. They can be difficult to comprehend without some understanding of statistics.

Sten and Stanine scores. The Sten score scale is a standardized, 10-point-scale with a mean of 5.5 and a standard deviation of 2. Stanine scores have a 9-point scale with a mean of 5 and a standard deviation of 2. Both Sten and Stanine scores are commonly used in feedback. Higher scores indicate better performance. Like *T*-scores, they are an even scale, but the smaller range often is easier to understand.

Theta scores. Unlike tests based on Classical Test Theory, where scores are based on the total number of correct items, the Watson-Glaser III uses advanced IRT scoring techniques. The output of the test is a theta score. This scale ranges from –3.00 to +3.00. This is an estimate of a candidate's ability level that takes into account, amongst other factors, the difficulty level of the items answered correctly. The implication of incorporating item difficulty into the scoring process is that two individuals answering the same number of correct items (traditional raw score), may have a different theta/Percentile/*T*-Score if the questions they answered varied in difficulty. This provides a much more sophisticated scoring system and gives credit to those that have been able to correctly answer the tougher questions in the test.

Theta scores have been omitted from the Watson-Glaser III computer-generated reports, because many users are unfamiliar with this type of score. Traditional metrics are reported for easier interpretation. However, users may extract theta score data, free of charge from the TalentLens test platform. The data extract provides the overall theta score and other test score information. Theta scores are included for the Recognizing Assumptions, Evaluating Arguments, and Drawing Conclusions subscales.

Interpreting the Watson-Glaser Development Report

Though the Watson-Glaser Development Report provides managers, coaches, or other development professionals with insights and specific guidance for strengthening an individual's critical thinking knowledge and skills, it is directed primarily toward the individual (e.g., "You scored higher than most of your peers.").

Best practices in the training and development literature suggest that the report is more effective when combined with interventions, such as coaching, classroom training, e-learning, and/or structured self-study (Goldstein & Ford, 2002). It is important to note that certain cognitive abilities that facilitate effective critical thinking (e.g., working memory; reading ability) are unlikely to change through developmental intervention. Still, critical thinking can be improved when efforts focus on improving knowledge and behavioral skills (Halpern, 1998; 2003).

- To promote an awareness of where the individual stands on each of the three RED dimensions, the report provides an in-depth review of the individual's assessment results for each of the three subscales, including interpretations of how his or her scores would translate into actual behaviors.

- To help individuals build their critical thinking skills, the report offers multiple, customized development suggestions grounded in the academic literature. The suggestions are based on the individual's subscale score ranges (as described previously), meaning he or she receives a different set of suggestions depending on whether his or her scores were in the "high range" (Strength to Leverage), in the "average range" (Further Exploration), or in the "low range" (Opportunity for Development).

- To enable individuals to translate the results into their day-to-day experiences, structured space is provided for them to reflect on the meaning of their results and the development suggestions that seem most useful to them. The report also provides guidance to help individuals apply knowledge of their critical thinking skills to key workplace competencies (e.g., decision making).

- To facilitate a strong development plan, the report offers guidance on how to create a realistic plan for building the individual's critical thinking skills based on best practices for development.

- The report concludes with suggestions for next steps that individuals should take to continue growing. In total, the Watson-Glaser Development Report offers individuals key insights, suggestions, and structured guidance to promote the growth of their critical thinking knowledge and skills.

Accuracy of Test Scores

Scores obtained on the Watson-Glaser and any other psychological test can be considered only an estimate of the test taker's true score. This is because no test is perfectly accurate (without error). The standard error of measurement *(SEM)* indicates the amount of error to be expected in a test taker's score. The amount of error can be expressed as range of raw score points (i.e., theta for Watson-Glaser III or number correct for Watson-Glaser II), standardized scale points, or percentile points. In any case, an individual's true score lies within that range. Standard error of measurement is described in more detail in the Reliability chapter of this manual.

Limitations of Test Scores

Though tests are carefully standardized, any changes to the administration process can result in unreliable test scores. Test scores should be interpreted carefully because they can be affected by many factors. Errors may arise from the administration of the testing session and scoring. Scores can also be affected by a test taker's state, for example anxiety or feeling unwell. Candidates with a disability or with English as a second language may be disadvantaged by the test format, even if accommodations are granted. For these reasons, explore and interpret scores carefully; especially test results from an unsupervised administration. Unsupervised results can and should be verified by retesting the final pool of applicants at the latter stages of an assessment process in a supervised session, or via information from other sources, such as a structured interview or assessment center exercise that measures the same abilities. On occasion, test scores may contradict alternative information about a test taker. When this happens, the test user should work with the test taker to explore the information and discover possible causes for these anomalies.

Using Watson-Glaser in the Selection Process

Reasoning ability tests have been shown to be the most effective single predictor of job performance and training success (e.g., Robertson & Smith, 2001; Salgado et al., 2003; Schmidt & Hunter, 1998; 2004). Using reasoning tests, such as the Watson-Glaser, enables employers to make more informed decisions about an applicant's ability and reduces poor recruitment decisions. The Watson-Glaser may be used as an initial screening of candidates, either unsupervised via the Internet or under supervision at the employer's premises. Watson-Glaser results can be used with other assessments (e.g., as part of an assessment center) to provide a full profile of an applicant. Before using the Watson-Glaser as part of a selection process, organizations need to ensure that the test is relevant and appropriate for the role. Using inappropriate tests can result in poor and unfair decisions.

When using testing (and assessment) in the selection process, the results provide information the employer can use to choose between job applicants. There are two key aspects to consider:

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

Performing a job analysis provides recruiters with a clear understanding of a job and of what it 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 other relevant characteristics. It often includes a competency profile.

The job description should be used to decide on the type of assessment that is most relevant. For example, the job description should indicate the level of difficulty at which critical thinking skills with verbal or numeric data, spatial ability, or mechanical aptitude need to be measured. The test user should also review this manual to ensure that the skills Watson-Glaser measures is relevant to the job and the individuals to be tested. This includes reviewing the norms, reliability, validity, and group comparisons. Standards of assessment should not be higher than what the job requires.

You should be satisfied that the norms provide suitable comparisons for your test takers. The norms should contain a representative sample with respect to the type of jobs applied for and background. Percentile scores can be used to understand the ability level of a particular candidate or to rank order candidates for short listing purposes. However, you should not create a short list based on a single measure, as this reflects only a single aspect of performance.

A good understanding of the role 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 organization could be required to show that any assessments used were carefully chosen and relevant if legally challenged.

Fairness in Selection Testing

Fair employment regulations differ across countries, and their interpretation is continuously subject to changes in the legal, social, and political environments. Consult with qualified legal advisors and human resource professionals if you have any questions.

Legal Considerations

The U.S. has governmental and professional regulations that cover all personnel selection procedures. Relevant source documents users may wish to consult include the Standards for Educational and Psychological Testing (AERA et al., 1999); the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, 2003); and the federal Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, 1978). For an overview of the statutes and types of legal proceedings which influence an organization's equal employment opportunity obligations, the user is referred to Cascio and Aguinis (2005) or the U.S. Department of Labor's Testing and Assessment: An Employer's Guide to Good Practices (2000).

Group Differences/Adverse Impact

According to the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, 1978), *adverse impact* typically is indicated when the selection rate for one group is less than 80% (or 4 out of 5) of another. Adverse impact is likely to occur with cognitive ability tests, such as the Watson-Glaser II and Watson-Glaser III. A test with adverse impact can be used for selection (Equal Employment Opportunity Commission, 1978), but the testing organization must demonstrate that the selection test is job-related, predicts performance, and is consistent with business necessity. A local validation study, in which Watson-Glaser scores are correlated with indicators of on-the-job performance, provides evidence to support the use of the test in a particular job context. A local study that demonstrates that the Watson-Glaser is equally predictive for protected subgroups, as outlined by the Equal Employment Opportunity Commission, helps to establish test fairness. Additional guidance on monitoring your selection system for fairness follows.

Monitoring the Organization's Selection System

To evaluate selection strategies and implement fair employment practices, an organization needs to know the demographic characteristics of applicants and incumbents. Monitoring these characteristics and accumulating test score data are necessary for establishing the legal defensibility of a selection system, including systems that incorporate the Watson-Glaser. Watson-Glaser is most effectively used where the following best practices are incorporated over time:

- At least once every 5 years, conduct a job analysis of the position for which you are administering the Watson-Glaser. A job analysis will help you determine if the job has changed in a way that requires adjustments to your assessment system.
- Periodically (e.g., once every 5 years) reassess the criterion-related validity of the selection system through local validation studies.
- Carefully monitor assessment scores for evidence of adverse impact. Evaluate adverse impact by comparing the selection rates for individuals from EEOC protected subgroups (e.g., gender or ethnicity) with selection rates of historically advantaged groups. Information needed to facilitate these analyses includes applicant demographics (e.g., voluntary information on gender, race/ethnicity, and age), assessment scores, and employment status (e.g., hired/not hired).
- Periodically re-examine cut scores in light of recent validity results, adverse impact, market data, and other factors (e.g., projected workload), and make adjustments as necessary.
- When sufficient samples of employees and candidates have been obtained (e.g., > 25 per demographic group), analyze the information collected to see if the selection procedure predicts equally for the majority group and EEOC protected groups. The Cleary model is a commonly-used approach to evaluate the fairness of selection tools (Guion, 1998). This model utilizes regression analysis to determine if a test demonstrates differential validity or prediction among subgroups of applicants. That is, it determines if a test over- or under-predicts job performance based on subgroup membership.

If you require assistance, please contact Pearson TalentLens.

Using the Watson-Glaser 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 that make the most of their strengths. Tests can be used to develop an awareness of potential, explore occupational awareness, and identify training needs. In choosing to use a test in this context, the user should evaluate what the test results yield and the information he or she is seeking.

In a development context, the Watson-Glaser can be helpful in better understanding a person's critical thinking skills. Consider an individual's score according to the three areas of the RED model. In the work setting, managers and employees can use the results to define appropriate development goals and activities. In a coaching or career exploration setting, individuals use this information to build on their strengths and minimize the impact of their weaknesses. This may be reflected in appropriate career choices and work strategies or in identifying opportunities to work on areas that require development. Other assessment (e.g., personality assessments and 360 feedback) results should be used with Watson-Glaser results to obtain a complete picture of an individual and his or her areas for development. The Watson-Glaser III generates a Development Report.

In an outplacement and career guidance context, the Watson-Glaser 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.

The Watson-Glaser produces two levels of career-relevant information:

- Career scope and potential for analytical work

  Performance on these 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 Watson-Glaser as those employed in roles that require critical thinking using language. Therefore, the level of scores on the Watson-Glaser 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 Watson-Glaser, then it can be inferred that he or she may be less suited to roles with a large verbal critical thinking component, such as, for example, marketing, legal advice, and human resource roles.

The Watson-Glaser will not directly help an individual realize his or her ambition, but can provide critical information for evaluating the possibilities available and an indication of potential for success.

Providing Feedback

If you or another administrator provide performance feedback to test takers, it is important to do it in language they can clearly understand. Feedback should be fair, accurate, and meaningful. Be aware that providing feedback can be a sensitive process, as some people have emotional reactions to information about their strengths and weaknesses.

You can provide feedback in writing, face to face, or over the telephone. Written feedback may be more appropriate when you have a large number of test takers and face-to-face or telephone feedback is not feasible.

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 standardization of test conditions, and so a fair and accurate assessment will rely on this.
- Ensure that scores have been converted to the appropriate percentiles, using a relevant norm group.
- Ensure that he or she has a clear understanding of the relevance of the Watson-Glaser results to the context in which it has been used.

An 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 are percentile scores.

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 Watson-Glaser as part of the recruitment process. The Watson-Glaser 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 individual's 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 a broad based representative group of managers who have completed this test, your score was at the 80th percentile. That is, you did better than 80% of the other managers who took it. Only 20% did better than you.

   or

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

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 incumbent will be required to use these skills to analyze 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 taker's knowledge of assessment and the Watson-Glaser test,
- provide relevant information about the test,
- describe the group against whom the test taker is being compared,
- describe performance in relation to that group,
- be stated in simple terms and 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, and
- be a positive experience for test takers, where information is related to their needs.

Feedback should always be meaningful to the test taker, and this means reporting how the applicant performed in comparison with the norm group used. For example, a candidate with a Watson-Glaser raw score that is classified as well-below average when compared to Graduate Degree norm group may be classified as average when compared to the High School norm group. Without this additional information, the test taker may be misinformed about his or her level of critical thinking.

History and Development of the Watson-Glaser

History

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 organizations 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.

The test has gone through a number of refinements and developments since its launch. These revisions incorporated enhancements requested by customers and advances in research and technology, while maintaining the qualities that have made the Watson-Glaser the leading critical thinking appraisal for nearly a century.

Figure 2. Watson-Glaser Development

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 US (Watson & Glaser, 1964). The forms were revised in 1980 to update the language, improve clarity, and eliminate racial and gender bias (Watson & Glaser, 1980). The new forms, A and B, were shorter at 80 items, but otherwise retained the basic test structure.

The first United Kingdom (UK) adaptation—Form C (Watson & Glaser, 1991)—was based on the U.S. Form B, which had been widely used in the UK with senior managers and in high-value occupational settings. American-English vocabulary and usage and content were adapted as appropriate for a UK test taker. In 2000, minor revisions were made to Form C and an extensive UK norming and standardization was completed. The result was the 80-item Watson-Glaser UK edition also called Watson-Glaser Form C (Watson, Glaser, & Rust, 2002).

A shorter 40-item form was developed in 1994 for use in employment-related training and career development contexts. A subset of Form A items were used and published as Form S (Short Form; Watson & Glaser, 1994; Watson & Glaser, 2006). Historical and test development information for the Short Form is available in the Watson-Glaser, Short Form Manual, 2006 edition. Historical and test development information for Forms A and B is available in the Watson-Glaser, Forms A and B Manual, 1980 edition. The more recent versions of the Watson-Glaser build on and redevelop these earlier forms.

With more demand for the test and for shorter forms around the world, two parallel short forms were developed. First, Form D was built on the foundation of the original Short Form. Items that were internationally appropriate and amenable to translation were chosen and updated. A parallel 40-item version, Form E, was developed from the original Form B for use in the U.S. Forms D and E comprise the Watson-Glaser II. Most recently, in order to support unsupervised Internet-based testing, a new item-banked test version has been added to the Watson-Glaser family of tests: the Watson-Glaser III.

Development of the Watson-Glaser, Third Edition

The current 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 precision for 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 a large number of equivalent forms.
- Enhanced computer generated reporting including a basic Profile Report and a Development Report.
- Improved subscale structure to enhance interpretability.

Development of the RED Model

To improve subscale structure and interpretability, a substantial review and revision of existing Watson-Glaser test forms was undertaken, leading to the publication of the Watson-Glaser II (Watson & Glaser, 2010). The clearer, more theoretically grounded RED model was introduced in the second edition.

Exploratory Factor Analysis

Development of the Watson-Glaser II RED Model began with investigation of the factor structure of Watson-Glaser Forms A, B, and the Short Form. A series of exploratory factor analyses were conducted with testlet scores from these forms. A *testlet* is one scenario and a set of two to six questions. Testlet scores are generated by summing the number of correct responses for items associated with each scenario. Evidence suggests that factor structures based on testlets are more robust than those based on individual items (Bernstein & Teng, 1989). A maximum likelihood extraction method with oblique rotation was used to analyze Form A (N = 2,844), Form B (N = 2,706), and Short Form (N = 8,508). 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 Recognize Assumptions, Evaluate Arguments, and Draw Conclusions (i.e., Inference, Deduction, and Interpretation loaded onto one factor). Given this evidence, logical appeal, and interpretive ease, the three-factor model was proposed for the Watson-Glaser II. An exploratory factor analysis with a UK sample of 714 who completed the Watson-Glaser Form C replicated the separation of Recognize 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 also can 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 Watson-Glaser II 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, presented in Table 1 and Figure 2 indicate that Model 1 did not fit the data as well as the other two models. Model 2 and Model 3 fit the data, and there was no substantive difference in model fit between the two. The phi coefficients in the five-factor model were problematic, suggesting that the constructs were not meaningfully separable. Given this evidence, the three-factor model was confirmed as the optimal model for the Watson-Glaser.

While assessing the reliability and construct validity of Watson-Glaser II forms and linking them to previous forms, confirmatory factor analyses run during the tryout stage were replicated. 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 subscales of the Watson-Glaser.

Table 1. Watson-Glaser II Confirmatory Factor Analyses (*N* = 306)

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

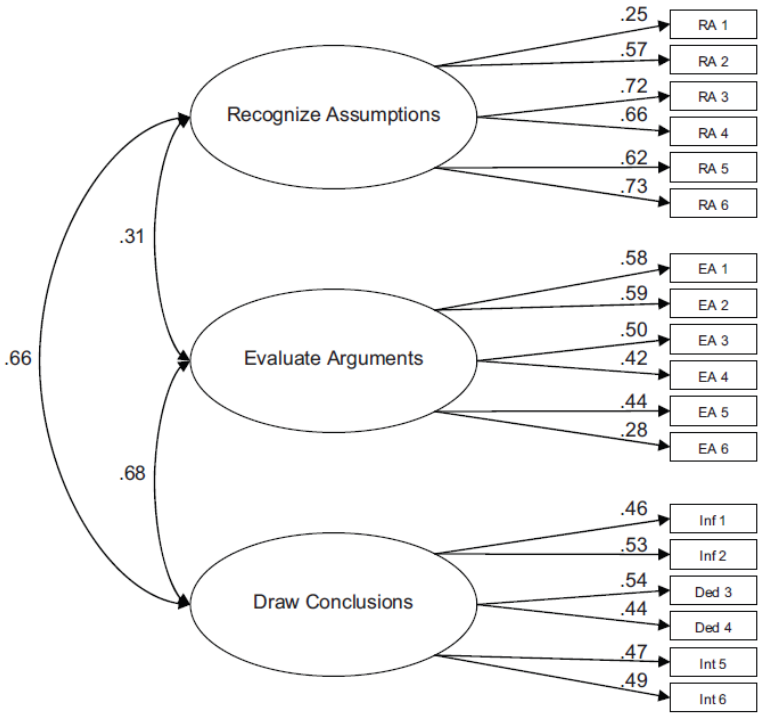

Figure 3. Three-Factor Model (Model 2) for Subtests and Testlets (*N* = 306)

In Figure 3, testlet scores were used as the unit of analysis; RA = Recognize Assumptions; EA = Evaluate Arguments; Inf = Infer; Ded = Deduce; and Int = Interpret

To summarize, the Watson-Glaser items and subscales were reorganized into a simplified, three-subscale structure. This included 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 (Recognize Assumptions and Evaluate Arguments) and a maximum of 16 items (Draw Conclusions).

Developing the Item Bank

Item development began after the three-factor test structure was finalized. A larger bank of items was required to support Internet delivery and generate parallel forms.

The item bank was developed in three main stages: item writing, item piloting, and data analyses. The following criteria were established for new items, content, and generating forms:

- New items must be equivalent to existing items.
- Forms generated from the item bank must reliably measure critical thinking skills.
- Content must be relevant for a global market.
- Content must be is appropriate for all demographic groups.
- Item-banked test forms must 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 of 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 that would be unequally familiar to different groups. A total of 349 items survived the review process and were included in the item trials.

Item Pilots

The new items were assigned to one of five 120-item forms. These forms included anchor items from the existing Watson-Glaser II and the Watson-Glaser Form C and were structured like all the Watson-Glaser tests with all five item types presented in the same order as the operational tests.

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.

A total of 1,563 people completed one or more of the 5 pilot forms. This included just over 1,200 postgraduate students in 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 Watson-Glaser tests, which are used predominantly in the recruitment and development of university graduates and other higher level staff.

Cases were removed from the data set where there was evidence that candidates had not completed the test seriously. These included people who 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 Watson-Glaser Form C as part of a recruitment process and 169 candidates who had completed the Watson-Glaser II as part of employment evaluation were included in the final data set for analysis.

Psychometric Analysis

Passages and items to be included in the item-bank were analyzed in three stages.

Stage 1: Classical Item Analysis. Each pilot form was analyzed separately. Item difficulty and discrimination were 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 that were not performing well were removed from the item pool. This included items that 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: Factor Analysis. In preparation for the IRT analysis, it was necessary to determine whether the item pool should be analyzed together, or whether the analysis should be carried out within the structure of the RED model with the three factors analyzed separately. The RED model factors are correlated and therefore a combined or separate analysis might be appropriate. Each pilot form was analyzed separately using principle components analysis. Scree plots were examined 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 unidimensional structure, rotation to simple structure was examined. This indicated a separation between the R items and the E and D items. 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. IRT Analysis: The BILOG-MG program (Zimowski, Muraki, Mislevy, & Bock, 1996) 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 unidimensional models were estimated for the full item pool and the R items separately from the E and 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 alternative trial forms based on different item parameter models. 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 .80 to .88 for different pairs of forms for the combined estimation. For separate estimation, values were substantially lower for the same test pairs ranging from .71 to .82. The model chosen was a three-parameter model treating all items as a unidimensional set.

Item Bank Configuration for Test Generation

The Watson-Glaser III item bank was first launched in the UK in 2011 and was updated in 2013 and 2017. The current English language bank is composed of 295 items. The item bank for the five subtests was configured 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 (i.e., any administration), as shown in Table 2. 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	**Watson-Glaser III**	**Watson-Glaser II**	**Watson-Glaser Form C**
Recognize Assumptions	12	12	16
Evaluate Arguments	12	12	16
Inference	5	5	16
Deduction	5	5	16
Interpretation	6	6	16
Total number of items	40	40	80

Evidence of Reliability

The reliability of a test refers to its accuracy, consistency, and stability of scores across situations (Anastasi & Urbina, 1997; Sattler, 2008). If the same test taker is repeatedly tested under identical conditions, you can expect a certain amount of consistency among his or her scores. The difference between a test taker's hypothetical true score and his or her obtained score is called *measurement error.* A reliable test has relatively little random and systematic error across administrations and provides consistent scores.

Reliability is expressed as a coefficient that ranges from zero to one. The closer the reliability coefficient is to 1.00, the more reliable the test and the less measurement error in test scores. When tests are used in employment contexts, reliabilities above .89 generally are considered excellent, .80–.89 good, and .70–.79 adequate. Values below .70 suggest the test may have limited applicability. For example, it may be used to provide developmental feedback, but would not be appropriate for making selection or promotion decisions.

Because 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.

Adding the *SEM* to and subtracting it from an examinee's test score creates a confidence interval or band of scores around the obtained score. The confidence interval likely includes the examinee's hypothetical "true" score, which represents the examinee's actual ability. The true score is a hypothetical value that can never be obtained. Testing always involves some measurement error, so 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 is within +1.96 and –1.96 *SEM* of the true score.

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

Internal Consistency Reliability

Internal consistency is an estimate of how consistently test items measure a construct. Internal consistency is a description of the homogeneity of the items in a test. Cronbach's alpha was used to measure the internal consistency reliability of the Watson-Glaser III. Reliability coefficients are sample dependent with tests showing higher reliability in more heterogeneous samples and lower reliability in groups that are all at similar levels of ability.

Table 3 shows reliability results for the Watson-Glaser for a number of samples. Data were gathered from a sample of 147 working adults in the U.S. to determine internal consistency reliability of the Watson-Glaser III. Analysis yielded a reliability value of .83, suggesting good internal consistency reliability for this new edition. Data also were gathered from UK participants who took the Watson-Glaser III, as well as from participants who took the Watson-Glaser II. The results show the reliability to be above .80 for all studies, aside from the smaller of the Watson-Glaser II UK groups. Table 3 also includes estimates of the T-score *SEM,* which has an average score of 50 and a standard deviation of 10.

Table 3. Internal Consistency Reliability Statistics

Test/Form	Sample	Mean	*SD*	*r*	*SEM*	*T*-score *SEM*
Watson-Glaser™ III	U.S. occupational, online completion *n* = 147	–0.35	1.01	.83	.42	4.1
	UK pilot *n* = 355	–0.17	.80	.84	.32	4.0
	UK pilot *n* = 355	–0.12	.82	.82	.35	4.2
	UK pilot *n* = 318	–0.06	.89	.84	.36	4.0
	UK pilot *n* = 318	–0.07	.92	.86	.35	3.7
Watson-Glaser II Form D	1011	—	—	.83	2.63	—
Form E	1043	—	—	.81	2.64	—
Watson-Glaser II Form D	UK occupational, online completion *n* = 169	30.4	5.0	.75	2.5	5.0
	UK standardization *n* = 1546	57.2	8.3	.81	3.6	4.4

Test–Retest Reliability

Prior Evidence of Test–Retest Reliability

The Watson-Glaser measures the cognitive ability of deductive reasoning, 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. Table 4 summarizes some of these results. In 1980, Form B was administered twice to a group of 96 students with a 3-month testing interval and correlated at 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 Watson-Glaser two weeks apart. The test–retest correlation was .81 and the difference in mean scores between the first testing and the second testing was not statistically significant (d = 0.16).

In 2006, test–retest reliability was evaluated using a sample of 57 job incumbents drawn from various organizational 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 not statistically significant (d = 0.17).

The last two 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 the Watson-Glaser II Technical Manual and User's Guide (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 good 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

	First testing		Second testing			
W-G Form	**Mean**	***SD***	**Mean**	***SD***	***r***	***SEM***
Form B	57.4	8.1	56.8	8.4	.73	0.07
Short	30.5	5.6	31.4	5.9	.81	0.16
Short	29.5	7.0	30.7	7.0	.89	0.17
Short paper-pencil	28.8	5.7	29.5	5.5	.88	0.13
Short online	30.1	5.7	30.6	5.5	.86	0.09

Alternate Form Reliability

Alternate form reliability considers whether similar scores are returned from different test forms. This is particularly important for the Watson-Glaser III, where candidates each receive a randomly assembled test form with different items. To evaluate this, pairs of parallel test forms were constructed from within the different item pilot pools so that scores could be compared. In addition, two pairs of nonequivalent forms 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 tests had alternate form reliabilities above .80 and that, even for the nonequivalent forms, alternate form reliability is around .80 with no statistically significant differences in mean scores between forms.

Table 5. Watson-Glaser III Alternate Form Reliability

	Sample	Mean	*SD*	*r*	Cohen's *d*
Equivalent Pair A	UK pilot sample *n* = 355			.82	0.06
Test 1		–.17	.80		
Test 2		–.12	.82		
Equivalent Pair B	UK pilot sample *n* = 318			.88	0.01
Test 3		–.06	.89		
Test 4		–.07	.92		
Non-Equivalent Pair C	UK pilot sample *n* = 308			.78	0.04
Easy test		–.17	.89		
Difficult test		–.21	.90		
Non-Equivalent Pair D	UK pilot sample *n* = 282			.80	0
Easy test		–.24	.91		
Difficult test		–.24	.89		

A variety of studies have been conducted assessing the correlation between various Watson-Glaser forms (see Table 6 for a summary). These reliabilities for samples of adults are good, and those based on samples of students are adequate. Studies of forms preceding the Watson-Glaser III are reported in more detail in Watson and Glaser (1980), Watson and Glaser (1991) and Watson and Glaser (2010).

Table 6. Alternate Forms Reliability, Watson-Glaser Multiple Form Comparisons

Study	Mean	*SD*	*r*
n = 147 U.S. working adults			.80
W-G II	26.18	5.7	
W-G III	–0.35	1.0	
n = 209 U.S. working adults			.82
W-G II Form D	22.3	6.3	
W-G II Form E	22.9	5.6	
n = 636 U.S. trial participants			.85
W-G II Form D	27.1	6.5	
W-G Short	29.2	5.7	
n = 288 12th grade U.S. students			.75
Watson-Glaser			
Form A	46.8	9.8	
Form B	46.6	9.3	
n = 53 UK 6th form (high school) students			.71
Watson-Glaser			
Form B	56.8	8.3	
Form C	57.4	9.5	

Reliability of the Subscale Scores

The Watson-Glaser III provides an overall test score and three subscale scores based on the RED model. (These scores are also included on the Watson-Glaser II reports, but described as high, medium, or low.) In developmental contexts, users may wish to compare scores for different parts of the model to understand an individual's relative strengths and weaknesses. The reliability of the subscale scores is lower than for the full test, as they are based on fewer items. This should be considered when interpreting profiles. Table 7 shows the reliability coefficients and *SEMs* in brackets (in *T*-score units) of the subscale scores for the Watson-Glaser II and Watson-Glaser UK, and Watson-Glaser III (unsupervised trials). The reliability of the subscale scores is good to excellent for the longer Watson-Glaser UK, but is mostly only adequate-to-good for the shorter 40-item Watson-Glaser II UK sample. Some coefficients are lower than would be desirable, because there are as few as 12 items for some of the subscale scores. This means that a longer test form is needed if interpretation at the subscale level is important. The last column provides the reliabilities and *SEMs* (in *T*-score unit) of the subscales for the item-banked test, based on the long 120-item pilot forms. The reliability for the 40-item Watson-Glaser III will be somewhat lower than those shown in Table 7.

Table 7. Subscale Internal Consistency Reliability

	Watson-Glaser II (40 Items)	**Watson-Glaser II (40 Items)**	**Watson-Glaser UK (80 Items)**	**Watson-Glaser III unsupervised pilot (120 Items)**
Sample	U.S. standardization *n* = 1,011	UK occupational *n* = 169	UK occupational *n* = 714	UK student *n* = 2,446
Total/Subscale		**Reliability *(SEM)***		
Total	.83 (2.6)	.75 (2.5)	.92 (3.4)	.90 (4.3)
Recognize Assumptions	.80 (1.3)	.66 (1.4)	.83 (1.4)	.81 (5.8)
Evaluate Arguments	.57 (1.5)	.43 (1.4)	.75 (1.6)	.66 (4.4)
Draw Conclusions	.70 (1.7)	.60 (1.5)	.86 (2.7)	.81 (3.0)

Tables 8 and 9 show the standard error of difference (SEDiff) in raw score, Sten, and *T*-score units between pairs of subscale scores. These show the smallest difference between pairs of scores required for 68% confidence of a real difference. Doubling these values provides 95% confidence interval.

For example, if a person has *T*-scores of 55, 60, and 67 on Watson-Glaser Form C for Recognize Assumptions, Evaluate Arguments, and Draw Conclusions, respectively, then only the difference between Recognize Assumptions and Draw Conclusions is greater than twice the standard error of difference (5.6 x 2). Therefore, we have 95% confidence that the person is better at drawing conclusions than at recognizing assumptions. The difference between Evaluate Arguments and Draw Conclusions is 7 *T*-score points, which is greater than the 6.2 standard error of difference. Therefore, we can say with 68% confidence that the person is better at drawing conclusions than evaluating arguments. However the difference between Recognize Assumptions and Evaluate Arguments is only 5 points, which is less than the relevant standard error of difference (6.5) therefore we cannot conclude that there is a difference in the person's ability in these two areas.

Table 8. Standard Error of Difference Between Watson-Glaser II Subscale Scores

	Evaluate Arguments			Draw Conclusions		
Watson-Glaser II	**Raw score**	**Sten**	***T*-score**	**Raw score**	**Sten**	***T*-score**
Recognize 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 Watson-Glaser Form C Subscale Scores

	Evaluate Arguments			Draw Conclusions		
Watson-Glaser UK	**Raw score**	**Sten**	***T*-score**	**Raw score**	**Sten**	***T*-score**
Recognize Assumptions	2.1	1.3	6.5	3.0	1.1	5.6
Evaluate Arguments				3.1	1.2	6.2

Lines of Evidence Supporting Validity

Validity is the degree to which specific data, research, or theory supports that a test measures what it is intended to measure and applies to its intended population. (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). “Validity is high if a test gives the information the decision maker needs” (Cronbach, 1970). To establish the usefulness of the Watson-Glaser, evidence of content validity and construct validity—including internal factor structure and convergent and discriminant validity—are presented.

Evidence of Content Validity

Evidence of content validity exists when the content of a test includes a representative sample of tasks, behaviors, knowledge, skills, or abilities of the identified construct. Watson and Glaser articulated the critical thinking skills to be measured many years ago by (Glaser, 1937; Watson & Glaser, 1952), and still correspond to critical thinking skills articulated in current models (Facione, 1990; Fisher & Spiker, 2004; Halpern, 2003; Paul & Elder, 2002).

Watson and Glaser (Glaser, 1937; Watson & Glaser, 1994) considered critical thinking to include

- attitudes of inquiry that involve an ability to recognize 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 generalizations 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.

Watson-Glaser 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.

The Watson-Glaser 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 subject 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 content validation 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 standardized 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 Watson-Glaser. In classroom and instructional settings the course content and objectives of such instructional programs should correspond to the constructs measured by the Watson-Glaser.

Evidence of Construct Validity

Evidence of construct validity is the extent to which the test measures the theoretical construct or trait it is designed to measure. 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. 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 Watson-Glaser more specifically. Exploratory and confirmatory factor analysis results were described in an earlier section and these explored and confirmed the internal structure of the test.

Subscale Intercorrelations

Correlations among the Watson-Glaser subscales are shown in Tables 10, 11, and 12. The correlations are attenuated by the unreliability of the subscales and, therefore, are lower for the shorter Watson-Glaser II, which is less reliable than the longer Watson-Glaser UK. The Watson-Glaser III 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. Corrected correlations are generally moderate to high between the subscales, which is consistent with representing three facets of the critical thinking construct.

Table 10. Intercorrelations Among Watson-Glaser II Subscale Scores (*n* = 169)

Scale	Mean	*SD*	Recognize Assumptions	Evaluate Arguments	Draw Conclusions
Total	30.4	5.0			
Recognize Assumptions	8.9	2.3	.76		
Evaluate Arguments	8.6	1.8	.73	.32 (.60)	
Draw Conclusions	8.1	1.5	.67	.28 (.44)	.40 (.79)

Table 11. Intercorrelations Among Watson-Glaser Form C Subscale Scores (*n* = 714)

Scale	Mean	*SD*	Recognize Assumptions	Evaluate Arguments	Draw Conclusions
Total	60.1	11.8			
Recognize Assumptions	12.8	3.3	.74		
Evaluate Arguments	11.5	3.2	.78	.40 (.51)	
Draw Conclusions	35.8	7.2	.95	.58 (.69)	.66 (.82)

Table 12. Intercorrelations Among Watson-Glaser III Subscale Scores (*n* = 2446)

Scale	Mean	*SD*	Recognize Assumptions	Evaluate Arguments	Draw Conclusions
Total	–0.01	0.95			
Recognize Assumptions	0.00	0.94	.86		
Evaluate Arguments	0.00	0.84	.73	.47 (.64)	
Draw Conclusions	0.02	0.95	.86	.58 (.71)	.63 (.86)

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 correlates with other cognitive ability measures, including nonverbal, verbal, and numerical reasoning; achievement; and critical thinking. A summary of U.S. studies is included in the Watson-Glaser II Technical Manual and User's Guide (2010). Correlations with achievement tests ranged from .39 to .51. Correlations with other reasoning tests ranged from .48 to .70. A study with 62 university students who completed the Wechsler Adult Intelligence Scale®, fourth edition (WAIS®–IV; Wechsler, 2008), and the Watson-Glaser II found a correlation of .52 between the Watson-Glaser total score and the WAIS–IV Full Scale IQ. The strongest correlation between WAIS–IV FSIQ and Watson-Glaser subscale was with Draw Conclusions. Three WAIS–IV index scores (Verbal Comprehension, Perceptual Reasoning, and Working Memory) correlated moderately with the Watson-Glaser, but the Processing Speed Index did not reach statistical significance. The correlation with the Fluid Reasoning Index was .60. An earlier study of Form Am with the original WAIS (1955) yielded a correlation of .41 with Full Scale IQ and .55 with Verbal IQ in a sample of 49 managers and executives. Overall, these results suggest that the Watson-Glaser is related to IQ, but it is more strongly related to reasoning power than processing speed.

A study of 41 participants in varied occupations and the Watson-Glaser Short form had a correlation of .53 with Raven's Advanced Progressive Matrices (Watson & Glaser, 2006). Further, in another sample of 307 participants in varied occupations, the Watson-Glaser Short Form correlated at .37 with Raven's (Pearson, 2015).

In the UK, the correlation between a measure of numerical reasoning (the Numerical Data Interpretation Test [NDIT]; Pearson, 2017) and the Watson-Glaser III for a sample of 6,734 applicants who work for a large UK-based company in the finance sector was .17. When corrected for restriction in range for the relatively high ability sample, the estimate of the true correlation is .29. In another sample of 91 working adults in the US, the correlation between Watson-Glaser III and NDIT was .47. Because these two tests both measure reasoning ability, they would not be expected to be related, but because NDIT uses mainly numerical content and Watson-Glaser is made up of verbal content, only a moderate correlation would be expected; therefore, these two studies support the construct validity of the Watson-Glaser.

Measures of Personality

In terms of the Big Five model of personality (e.g., McCrae & John, 1992), the Watson-Glaser as a cognitive ability measure is expected to correlate the strongest with the Openness Factor. Achievement motivation from the Conscientiousness factor may have low, but significant, correlations with the Watson-Glaser, and Neuroticism or Emotionality may 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 center 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 U.S. university graduates.

Evidence of 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 usefulness of the test.

Cronbach (1970) characterized 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. The selection ratio—percentage of applicants to be selected—also affects the usefulness of a test.

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 & Glaser, 2006) and Forms A and B (Watson & Glaser, 1980), respectively. A few selected findings are summarized in Table 13.

Table 13. Previous Studies Showing Evidence of Criterion-Related Validity

		Watson-Glaser			Criterion			
Group	***n***	**Form**	**Mean**	***SD***	**Description**	**Mean**	***SD***	***R***
Law firm applicants (Pearson TalentLens, 2013)	250	W-G Form C	67.55	3.94	Supervisory Ratings:	2.97	0.56	.44**
Legal training course students (Bar Standards Board, 2013)	123	W-G Form C	57.1	13.2	Final course grade	74.6	5.0	.62**
Job incumbents across multiple industries (Watson and Glaser, 2006)	142	W-G Short	30.0	6.0	Supervisory Ratings:			
					Analysis and Problem Solving	37.5	6.7	.33**
						31.9	6.1	.23**
					Judgement and Decision Making	101.8	16.9	.28**
					Total Performance	3.1	1.2	.24**
					Potential			
Job applicants and incumbents across multiple industries (Watson & Glaser, 2006)	2,303	W-G Short	31.0	5.6	Organisation Level	3.1	1.2	.33**

Table 13 *continued*

Group	*n*	Form	Mean	*SD*	Criterion Description	Mean	*SD*	*R*
Analysts from a government agency (Watson & Glaser, 2006)	64	W-G Short	32.9	4.4	Supervisory Ratings:			
					Analysis and Problem Solving	38.3	6.6	.40**
					Judgement and Decision Making	32.8	5.8	.40**
					Professional/Technical	17.1	2.4	.37**
					Knowledge and Expertise	100.4	14.3	.39**
					Total Performance	3.2	1.2	.25*
					Potential			
Leadership assessment centre participants from a national retail chain and a utility service (Kudish & Hoffman, 2002)	71	W-G Forms A/B	—	—	Assessor Ratings:			
					Analysis	—	—	.58*
					Judgment	—	—	.43*
Middle-management assessment centre participants (Spector, Schneider, Vance, & Hezlett, 2000)	189–407	W-G Forms A/B	66.5	7.3	Assessor Ratings:			
					In-basket	2.9	.7	.26*
					In-basket Coaching	3.1	.7	.16*
					Leaderless Group	3.0	.6	.19*
					Project Presentation	3.0	.7	.25*
					Project Discussion	2.9	.6	.16*
					Team Presentation	3.1	.6	.28*
					Openness to Experience (CPI Personality Trait)	41.8	6.4	26*
First year students on a Pennsylvania (US) nursing program (Behrens, 1996)	41	W-G Forms A/B	50.5	—	Semester 1 GPA	2.5	—	.59**
	31		52.1		Semester 1 GPA	2.5	—	.53**
	37		52.1		Semester 1 GPA	2.5	—	.51**
Education students (Gadzella, Baloglu, & Stephens, 2002)	114	W-G Forms A/B	51.4	9.8	GPA	3.1	.51	.41**
Educational psychology students (Williams, 2003)	158–164	W-G Short	—	—	Exam 1 Score	—	—	.42**
					Exam 2 Score	—	—	.57**
Education students (Taube, 1995)	147–194	W-G Forms A/B	54.9	8.1	GPA	2.8	.51	.30*
Educational psychology students (Gadzella, Stephens, & Stacks, 2004)	139	W-G Forms A & B	—	—	Course Grades	—	—	.42**
					GPA	—	—	.28**

As shown in Table 13, a study of applicants for legal roles at a global law firm examined the correlation between average performance measured on four occasions over 2 years and Watson-Glaser test scores gathered during the recruitment phase. Performance was measured via supervisor ratings, with each participant given an overall rating on a four point scale. The results showed a moderate correlation between average performance and the Watson-Glaser.

Research on job incumbents highlighted in Watson and Glaser (2006) found that Watson-Glaser scores (Short Form) had a moderate correlation with supervisory ratings of Analysis and Problem Solving behaviors, and a small correlation with supervisory ratings on a dimension made up of Judgment and Decision Making behaviors.

The Watson-Glaser is correlated with organizational success. For example, one study (see Watson and Glaser, 2006) found that for a large sample job incumbents across 9 industry categories, Watson-Glaser Short Form scores correlated moderately with job success as indicated by organizational 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.

Analysts from a U.S. government agency (discussed in Watson and Glaser, 2006) had Watson-Glaser Short Form scores that correlated moderately with supervisory ratings on each of two dimensions composed of (a) Analysis and Problem Solving behaviors and (b) Judgment and Decision Making behaviors, and also correlated moderately with supervisory ratings on a dimension composed of behaviors dealing with Professional/Technical Knowledge and Expertise as well as with "Total Performance" and Overall Potential.

Using a sample of leadership assessment center participants, Kudish and Hoffman (2002) reported that Watson-Glaser 80-Item (U.S. form) scores had a large correlation with ratings on Analysis and a moderate correlation 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 U.S. Ratings on Analysis and Judgment were based on participants' performance across assessment center 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 center exercise performance for managerial and executive level assessment center participants. They found that Watson-Glaser scores significantly correlated with six of eight assessment center exercises, and related more strongly to exercises involving cognitive problem-solving skills (e.g., in-basket scores) than exercises involving interpersonal skills (e.g., in-basket coaching exercise). Scores also had small but significant correlations with "Total Performance," a sum of ratings on 19 job performance behaviors, and with ratings on a single-item measure of Overall Potential.

In the educational domain, Behrens (1996) found that Watson-Glaser scores were correlated highly with semester GPA for three freshmen classes in a Pennsylvania nursing program. Similarly, Gadzella, Baloglu, and Stephens (2002) found Watson-Glaser subscale scores explained 17% of the total variance in GPA (equivalent to a multiple correlation of .41) for a sample of Education students. Williams (2003), in a study of educational psychology students, found Watson-Glaser total scores correlated moderate to high with mid-term and final exam scores. Taube (1995) found Watson-Glaser scores to have a significant, moderate correlation with GPA in a sample of U.S. students. Finally, Gadzella, Stephens, and Stacks (2004) reported a significant correlation between Watson-Glaser scores and both GPA (small) and course grades (moderate) for a group of educational psychology students.

The relationship between the Watson-Glaser II and job performance was also examined using a sample of 65 managers and their supervisors from the claims division of a national U.S. 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. 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.

Results showed that the Watson-Glaser II total score correlated .44 with supervisory ratings on a scale of core critical thinking behaviors 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 Behaviors (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 (n = 65)

	Watson-Glaser II Form D			
Supervisory rating scales	**Total score**	**Recognize Assumptions**	**Evaluate Arguments**	**Draw Conclusions**
Core Critical Thinking Behaviors	.44 (.30)*	.33 (.23)*	.17 (.11)	.48 (.33)**
Evaluating Quality of Reasoning and Evidence	.43 (.29)**	.32 (.22)*	.17 (.12)	.46 (.32)**
Bias Avoidance	.36 (.25)*	.31 (.22)*	.20 (.14)	.30 (.21)*
Creativity	.38 (.26)*	.25 (.17)	.15 (.10)	.45 (31)*
Job Knowledge	.34 (.24)*	.14 (.10)	.34 (.24)*	.30 (.21)*
Overall Performance	.17 (.12)	.03 (.02)	.04 (–.03)	.37 (.26)*
Overall Potential	.39 (.27)*	.13 (.09)	.21 (.14)	.53 (37)**
W-G II Form D descriptive statistics				
Mean	28.4	8.8	7.8	11.8
SD	5.4	2.3	2.3	2.6

Note. *p < .05 (one-tailed); ** p < .01 (one-tailed)

Watson-Glaser III Criterion Validity

A study of the validity of the Watson-Glaser Form C predicting outcomes on the Bar Professional Training Course (BPTC) was undertaken in 2010 and published in 2013 (Bar Standards Board, 2013). This is a post-graduate vocational course for those wishing to practice as a barrister in England and Wales. A total of 123 students on the course completed the W-G Form C for the purposes of the study during a 1-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 Watson-Glaser was 0.62. This is a very strong correlation and indicates that the Watson-Glaser 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 Watson-Glaser III in 2011. These were calibrated as described in the development section of this manual to provide scores on a standardized 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 Watson-Glaser and test results are shown in Table 15. 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 Watson-Glaser to predict performance in a vocational training context.

An independent review of the assessment's predictive quality was later commissioned by the Bar Standards Board (2015). Data from 998 students that had completed the test as part of the entry process onto the training course was compared with subsequent final course grades following completion of the training. This data once again confirmed the strong correlation between the test and course performance (r = .55).

Data were gathered from a sample of business school students who had completed the W-G III as part of the recruitment process, alongside an essay exercise, interview, and the CORE abilities test, which is a measure of general ability (Pearson, 2016). As shown in Table 15, scores on W-G III correlated with the essay task and the CORE ability test scores. The students' subsequent business degree final course grades were then correlated with the test scores. The results of this component of the study yielded a moderate correlation (r = .38) between W-G III scores and final course grade,

suggesting the test is a useful predictor of likely course success. Table 15 also shows the results of similar research with legal students. For these latter two sample the correlation between W-G III and final course grades is large.

Table 15. Watson-Glaser III Correlations With Final Course Grades and Several Recruitment Measures

		Watson-Glaser III			Exam Results
Student group	***n***	**Mean**	***SD***	**Criterion**	***r***
Business school (Pearson, 2016)	47	—	—	Final course grade	.38**
	47			Course admission–essay	.37**
	60			Course admission–interview	.15
	46			CORE Abilities	.40**
Legal training course (Bar Standards Board, 2015)	998	52.26 (T-score)	5.4 (T-score)	Final course grade	.55**
Legal training course (Bar Standards Board, 2013)	988	−.13	.9	Final course grade	.51**

Power Versus Speed

The Watson-Glaser is intended as a test of critical thinking power rather than speed. A pure power test is a test in which all items are attempted and performance is judged by the correctness of responses. A pure speed test, on the other hand, is composed of items that are so easy that examinees would almost never give a wrong answer to an item; however, the test is of sufficient length that no examinee would complete the entire test in the allotted time. Most tests are mixtures of speed and power components (Rindler, 1979). They fall along a continuum of speededness, where some tests are highly speeded and others are designed to minimize the speed component. Even tests focusing on power typically use time limits for practical administration purposes.

To ensure that Watson-Glaser measures power, rather than speed, a generous time limit has been established. (W-G III is also available with no time limit in the U.S., but this version is not recommended for high-stakes testing.) Even so, no matter how much time is allowed on a test, a small percentage of candidates will not attempt all items. For tests like Watson-Glaser that are designed to minimize the speed component, it is important to demonstrate that scores are not overly affected by speed, since this can threaten test validity by altering the construct that is measured (e.g., Lu & Sireci, 2007).

Several widely accepted “rules of thumb” are available to assess the degree to which a test is speeded (e.g., Swineford, 1974). For example, a test is considered *unspeeded* if:

- all test takers finish at least 75% of the items (or respond to at least one question beyond three-fourths of the way through the test), and
- at least 80% of test takers respond to all items (or reach the last question).

Another approach to assessing speededness is to take a fairness-oriented perspective, that is, speededness can be defined as “the extent to which some examinees are disadvantaged by the time limit on a test, relative to other examinees” (Schnipke, 1995, p. 4). This can be assessed by looking at the relative ranks of examinees assessed under both power and speed conditions. A perfect or near perfect correlation indicates that a test is unspeeded.

To ensure that the current time limit for Watson-Glaser III (30 minutes) is not introducing an unintended speed component, and to assess the degree to which scores from timed and untimed administrations are similar, a research study was conducted. The Watson-Glaser III was administered twice to a group of 137 online survey participants, once with a 30-minute time limit and once with no time limit (the design was counterbalanced with half taking the speeded version first). Results showed that the 30-minute time limit had little effect on scores.

- Average scores did not differ significantly across the timed and untimed condition.
- Theta scores correlated .73 across the two conditions, which is very close in magnitude to the test-retest reliability.
- All except one of the participants completed all 40 items under both conditions. That person completed 39 of 40 items in the timed condition. Based on the guidelines provided, the Watson-Glaser III can be viewed as an unspeeded test.

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 artifact of the test and whether the ability measured is relevant to job performance.

A difference can be an artifact of the test where the content or the way questions are asked particularly favors one group over another. For example, if the Watson-Glaser contained passages that were of more interest and familiar to men than women, this could create an unfair difference. If the Watson-Glaser test were to be used to select people for a role where high-level critical thinking is not relevant, it would result in unfair decisions for those who did not perform well on the test. For this reason, users must ensure that the test is relevant to the context of its use.

If the test proves to be relevant, but there are group differences, then a higher standard of test performance cannot be required than is justifiable based on the job requirements. If there are mean differences in test performance across groups, there will be greater adverse impact against the lower scoring group when there are higher cut scores than with lower scores. In other words the percentage of those selected who are from the higher scoring group will be greater the higher the cut score used.

In the US and 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 job requirements.

Tables 16–20 show the results of comparing performance on the Watson-Glaser for different demographic subgroups. They provide an indication of whether group differences are likely to be found when the test is used. However users should remember that all samples 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 that 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 Watson-Glaser 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 differences 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. Values of Cohen's d above 0.8 are considered large, above 0.5 moderate, and above 0.2 small. Below this level, values can be considered negligible. Remember that very small differences can reach statistical significance in large samples, but they have little impact on relative success rates. Cohen's *d* should be used as the main statistic for evaluating this type of group difference.

The data come from a number of different UK samples. The Watson-Glaser II data come from a mixed occupational group of 169 people who completed the test online for selection or development purposes. The UK standardization sample for Watson-Glaser Form C is a large mixed occupational group of 1,546 respondents and includes people from commercial, industrial, and public organizations. Job levels range from clerical workers and security staff through to senior managers

and professionals. Just over half had obtained a university degree. Another large sample consisted of 2,321 job applicants to a department of the government. Over 80% were external applicants but there were some internal applicants and the majority had obtained a university degree. There also are three groups of law students. The first is a group attending a particular college who completed the Watson-Glaser UK. The second group (*n* = 182) also completed the Watson-Glaser Form C as part of a pilot validation of the test for use in selection to the Bar Professional Training Course (BPTC). The third group of 1,501 BPTC students completed one of the five new item pilots that were used to develop the item bank for the Watson-Glaser III, under both supervised and unsupervised conditions. Their scores are reported on a raw theta scale, which has an approximate mean of zero and a standard deviation of one. The results presented in Tables 16–20 are based only on those who answered the relevant demographic questions and, therefore, sample sizes in the tables vary.

Male–Female Comparisons

Table 16. Watson-Glaser Score Male–Female Comparisons

	Male			Female			
Test/Group	**Mean**	***SD***	***n***	**Mean**	***SD***	***n***	**Cohen's *d***
Watson-Glaser II Mixed occupational	31.5	4.5	69	29.7	5.2	68	0.37
Watson-Glaser Form C Standardization	50.6	10.3	1019	50.0	9.3	520	0.01
Watson-Glaser Form C Government dept.	60.2	9.3	997	58.4	8.9	1269	0.20
Watson-Glaser Form C Law students 1	60.2	10.1	311	60.8	10.9	48	–0.06
Watson-Glaser Form C Law students 2	57.8	14.8	65	57.2	10.5	82	0.05
Watson-Glaser III Law students 3	–0.05	0.90	449	–0.20	0.82	545	0.18

Except two small effect sizes, most of the comparison showed 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 Watson-Glaser II, but this remains only a small effect, with men scoring higher than women on average.

Ethnic Origin Comparisons

Table 17. Watson-Glaser Ethnic Origin Comparisons

	White			Black			Asian			
Test/Group	**Mean**	***SD***	***n***	**Mean**	***SD***	***n***	**Mean**	***SD***	***n***	**Cohen's *d***
Watson-Glaser Form C Standardization	58.0	8.1	1332	50.1	6.9	60				0.99
	58.0	8.1	1332				52.1	8.1		0.73
Watson-Glaser Form C Government dept.	61.8	7.7	1484	51.3	9.6	245				1.31
	61.8	7.7	1484				52.1	8.1	313	1.25
Watson-Glaser Form C Law students 1	63.4	7.7	370	54.5	13.7	21				1.09
	63.4	7.7	370				55.2	12.4	71	0.95
Watson-Glaser Form C Law students 2	61.7	6.7	46	55.1	7.1	11				0.98
	61.7	6.7	46				53.7	6.6	21	1.20
Watson-Glaser III Law students 3	0.14	0.85	533	–0.49	0.78	82				0.75
	0.14	0.85	533				–0.52	0.72	265	0.82

Large effect sizes were found for all the ethnic group comparisons. Black groups tended to have the lowest scores, though this was not true for some of the law student groups. Differences ranged from .75 of one standard deviation to 1.25 standard deviations. These are large differences, but similar to those found on other cognitive ability tests (e.g., Hough, Oswald, & Ployhart, 2001). There may be some confounding between the ethnic groups used in these comparisons and people who have English as a second language, but most of those from minority ethnic background had English as their primary language.

Validation results for the third law sample (see following section) showed that the Watson-Glaser was valid for each ethnic group. Therefore, there is no evidence that these score differences reflect unfair bias in the test itself. However, users should be aware that there will likely 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

Table 18 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 Watson-Glaser II sample where the youngest test takers have the lowest average test score. The Cohen's *d* statistic and significance test is a comparison of the youngest and oldest groups and the two largest groups in each sample where these are different, because 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 they apply for higher level positions in new jobs. 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 18. Watson-Glaser Score Age Comparisons

	16–24			25–44			45+			Cohen's *d*
Test/Group	**Mean**	***SD***	***n***	**Mean**	***SD***	***n***	**Mean**	***SD***	***n***	**16–24 vs. 45+/ largest groups**
Watson-Glaser II Mixed occupational	28.6	5.9	15	30.9	4.8	105	30.7	5.5	16	–0.37 / 0.04
Watson-Glaser Form C Standardization	62.6	9.3	380	58.1	11.8	182	61.9	6.9	25	0.08 / 0.44*
Watson-Glaser Form C Government dept.	60.6	8.5	671	58.7	9.1	1354	57.7	10.4	269	0.21*/ 0.32*
Watson-Glaser Form C Law students 2	59.5	10.0	98	54.1	16.7	46	59	7.8	3	0.05 / 0.43*
Watson-Glaser III Law students 3	–0.12	0.84	647	–.011	0.95	280	-0.39	0.76	59	0.32* / –0.01

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 aggregated. These results may mask greater differences for specific disabilities. The most common disability noted for the law students group was dyslexia. The 31 students who described themselves as having dyslexia 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 Watson-Glaser is accessible to people with a range of disabilities.

Table 19. Watson-Glaser Score Disability Comparisons

	No Disability			Disability			
Test/Group	**Mean**	***SD***	***n***	**Mean**	***SD***	***n***	**Cohen's *d***
Watson-Glaser II Mixed occupational	30.6	4.9	126	31.5	5.5	6	–0.18
Watson-Glaser Form C Government dept.	59.1	9.1	2220	58.3	10.1	89	0.09
Watson-Glaser III Law students 3	–.014	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 Watson-Glaser test items are written to have a reading age of 15 or older, 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 also available in French, French Canadian, Dutch, Spanish, and Latin American Spanish.

Table 20. Watson-Glaser Score Primary Language Comparisons

	Primary Language English			Primary Language Other			
Test/Group	**Mean**	***SD***	***n***	**Mean**	***SD***	***n***	**Cohen's *d***
Watson-Glaser UK Law students 1	62.3	8.7	437	56.1	14.2	44	0.66*
Watson-Glaser UK Law students 2	59.2	11.7	124	50.0	14.6	20	0.75*
Watson-Glaser III Law students 3	–0.09	0.86	867	–0.33	0.79	80	0.28*

Differential Validity

The group difference analyses presented in Tables 16–20 showed small differences for most comparisons, but larger differences with respect to ethnicity. Differences in test scores lead to adverse impact when the test is used in selection. Such impact can only be justified if it can be shown that the differences are related to course outcomes for that group. The statistical approach to looking at this is a hierarchical regression. If including the group membership variable in the regression equation significantly improves prediction it suggests some bias in the results. Both the group membership variable and its interaction with test score are tested. If there is no improvement in prediction then the score differences on the test are also reflected in the outcome.

Differential validity analyses were performed on the sample of students in a vocational law course (also included in Tables 16–20). Hierarchical analyses were conducted for those variables for which the sample was large enough. 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. Analyses were conducted for those variables for which samples were sufficiently large, and are summarized in Table 21.

Table 21. Differential Validity Results Summary

Group	Results
Male-Female	Regression analysis showed no evidence of any bias in the test predicting course outcomes for females versus males.
Ethnic origin	This regression was run once for the Asian group compared with the majority and once for all remaining minority group members. There was no evidence of bias against these two groups, which scored lower on the test on average. The test predicted exam performance well for both groups, and in fact slightly over-predicted the performance for these two groups.
Age	A statistically significant difference was found with the test marginally favoring younger candidates but the effect was very small, accounting for less than 3% of the prediction. This is unlikely to have a meaningful impact on decision making.
Primary Language	The difference in course results was larger than that predicted by the test scores. Therefore, predictions based on test scores of those with English as a second language did not underestimate their level of performance in the course.
Disability	Correlation of .50 for group with disabilities and .52 for remaining group suggests test predicts well for both groups.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Americans With Disabilities Act of 1990, Titles I & V (Pub. L. 101-336). United States Code, Volume 42, Sections 12101–12213.

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

The Bar Standards Board. (2015). *BCAT Impact and Performance Evaluation.* London: The Bar Standards Board.

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.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43* (166), 38295–38309.

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.

Glaser, E. M. (1937). An experiment in the development of critical thinking. *Contributions to Education, No. 843.* New York: Bureau of Publications, Teachers College, Columbia University.

Goldstein, I. L., & Ford, J. K. (2002*). Training in organizations* (4th ed.). Belmont, CA: Wadsworth.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449-455.

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

Hough, L. M., Oswald, F. L., & 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(1/2*), 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.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide.* Thousand Oaks, CA: Sage Publications.

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.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement Issues and Practice, Winter,* 29–37.

McCrae, R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality 60.2:* 175–215.

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.

Pearson (2017). *NDIT™ numerical data interpretation test:* User's guide and technical manual. London: Author.

Pearson (2015). *Raven's™ advanced progressive matrices* (APM–III): User's guide and technical manual. London: Author.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*, 261–270.

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 generalization 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.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

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.

Swineford, F. (1974). *The test analysis manual* (ETS SR 74-06). Princeton, NJ: Educational Testing Service.

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.

Viswesvaron, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557–574.

Watson, G. (1925). *The measurement of fairmindedness*. *Contributions to Education, No. 176*. New York: Bureau 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 Jovanovich.

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 for 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. (2010). *Watson-Glaser critical thinking appraisal* (2nd ed.): Technical manual and user's guide. San Antonio, TX: Pearson.

Watson, G., & Glaser, E. M. (2012). *Watson-Glaser critical thinking appraisal* (3rd ed.): User's guide and technical manual and user's guide. London: Pearson Education Limited.

Watson, G., Glaser, E. M., & Rust, J. (2002). *Watson-Glaser critical thinking appraisal* (UK edition). London: Pearson Education Limited.

Wechsler, D. (1955). *Wechsler adult intelligence scale*. New York: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler adult intelligence scale* (4th ed.). Bloomington, MN: NCS Pearson.

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.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (1996). *BILOG-MG.* Skokie, IL: Scientific Software International.

Appendix A

Watson-Glaser Test Log

Maintaining a test log is best practice and should always be maintained.
The following Test Log may be photocopied for use in your organization.

Watson-Glaser Test Log

Organization	
Purpose of Testing	Selection / Development / Appraisal
Test(s) Used **(Circle all that apply.)**	Watson-Glaser II / Watson-Glaser III
Test Administrator(s)	
Date	
Start time	
Finish time	

Candidate List

	Candidate Name	Test	Retest		Candidate Name	Test	Retest
1.				6.			
2.				7.			
3.				8.			
4.				9.			
5.				10.			

Disturbances / Unusual Occurrences

Appendix B

Scoring Watson-Glaser II Paper-and-Pencil

Please contact your local TalentLens office to obtain the hand-scoring key and Form D or E norms tables for use in scoring.

Scoring With the Hand-Scoring Key

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 key over the Record Form.
2. For each subtest count the correctly marked circles.
3. Add the subtest scores to create a total raw score.
4. Record the total raw score in the box provided on the front of the record form.

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.ed

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

Step 3 – Converting raw scores into standardised scores using published norm tables which are availiable from your local TalentLens office.

Appendix C

Watson-Glaser III UK Norm Groups for Online Testing

Education Level Norm Groups	Gender and Ethnicity Breakdown
A Level N. 1212 The A Level norm group consists of 1212 people who have taken the W-G III who have been educated up to A Level. Their age ranges from 16 – 64 years. They completed the W-G III between June 2011 and March 2016.	59% Male, 38% Female, 3% not specified. 83% White, 8% Asian, 4% Not Specified, 3% Mixed, 2% BAC.
Bachelor Degree N. 7548 The Bachelor Degree norm group contains 7548 graduates between the ages of 20-64, who have all been educated up to the Bachelor Degree level. They completed the W-G III between June 2011 and March 2016.	53% Male, 44% Female, 3% Not Specified. 77% White, 11% Asian, 6% Not Specified, 3% BAC, 3% Mixed.
Doctoral Degree N. 681 The Doctoral Degree norm contains 681 people who have been educated up to Doctoral Degree. They are between the ages of 25 to 64 years. They completed the W-G III between June 2011 and March 2016.	58% Male, 38% Female, 4% Not Specified. 74% White, 12% Asian, 9% Not Specified, 3% BAC, 2% Mixed.

GCSE N. 830 This norm group contains 830 16-64 year olds who have been educated up to GCSE level. They completed the W-G III between June 2011 and March 2016.	58% Male, 40% Female, 2% Not Specified. 91% White, 3% Not Specified, 3% BAC, 2% Asian, 1% Mixed.
Higher Education Diploma N. 1876 The Higher Education Diploma norm group consists of 1876 people aged between 16-64 years. The group is comprised of people who have been educated up to Higher Education Diploma level. They completed the W-G III between June 2011 and March 2016.	53% Female, 45% Male, 2% Not Specified. 84% White, 6% Asian, 4% BAC, 4% Not Specified, 2% Mixed.
Master Degree N. 10169 This group, made up of 10169 people between the ages of 20-64, have all been educated up to Masters Degree level. They completed the W-G III between June 2011 and March 2016.	50% Female, 46% Male, 4% Not specified 71% White, 14% Asian, 6% BAC, 6% Not specified, 3% Mixed.

Graduate Norm Groups	Gender and Ethnicity Breakdown
Law N. 2425 This group is made up of 2425 people applying for graduate positions in Law. They are aged between 20-39 years. The majority of the group (92%) are aged between 20-24 years. They completed the W-G III between June 2011 and March 2016.	55% Female, 45% Male. 71% White, 17% Asian, 5% Mixed, 4% BAC, 3% Not specified.
Professional Services N. 1559 This group is made up of 1559 people applying for graduate positions in Professional services They are aged between aged 20 – 39 years. 82% of the people are aged between 20-24 years, 12% are aged between 25-29 years, 2% are aged between 30-34 years and 1% are aged between 35-39 years. They completed the W-G III between June 2011 and March 2016.	64% Male, 35% Female, 1% Not Specified. 60% White, 28% Asian, 5% BAC, 4% Not Specified, 3% Mixed.
Occupation Norm Groups	**Gender and Ethnicity Breakdown**
Accountants N. 1574 This group is made up of 1574 Accountants aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	60% Male, 37% Female, 3% Not Specified. 76% White, 13% Asian, 6% Not Specified, 3% BAC, 2% Mixed.

Admin Clerical N. 1314 This group is made up of 1314 people in Admin Clerical roles. They are aged between 16-64 years. They completed the W-G III between June 2011 and March 2016.	59% Female, 38% Male, 3% Not Specified. 77% White, 9% Asian, 6% BAC, 5% Not Specified, 3% Mixed.
Consultants N. 1254 This group is made up of 1254 people who are working in Consultancy roles. They are aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	62% Male, 34% Female, 4% Not Specified. 76% White, 11% Asian, 7% Not Specified, 3% BAC, 3% Mixed.
Customer Service N. 737 This group is made up of 737 people working in Customer Services roles. They are aged between 16-64 years. They completed the W-G III between June 2011 and March 2016.	49% Female, 49% Male, 2% Not Specified. 70% White, 19% Asian, 5% BAC, 4% Not Specified, 2% Mixed.
Financial Analysts N. 516 This group is made up of 516 Financial Analysts. They are aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	66% Male, 29% Female, 5% Not Specified. 61% White, 23% Asian, 8% Not Specified, 4% BAC, 4% Mixed.

Firefighters and Police officers N. 333 This group is made up of 333 people who are in Firefighter or Police Officer roles. They are aged between 25-64 years. They completed the W-G III between June 2011 and March 2016.	86% Male, 11% Female, 3% Not Specified. 93% White, 3% Not Specified, 2% BAC, 1% Asian, 1% Mixed.
HR professionals N. 1149 This group is made up of 1149 people who are in HR professional roles. They are aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	66% Female, 32% Male, 2% Not Specified. 83% White, 8% Asian, 4% BAC, 3% Not Specified, 2% Mixed.
IT professionals N. 474 This group is made up of 474 IT Professionals aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	76% Male, 19% Female, 5% Not specified. 66% White, 17% Asian, 8% BAC, 7% Not Specified, 2% Mixed.
Legal professionals N. 4337 This group is made up of 4337 Legal Professionals. They are all aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	58% Female, 39% Male, 3% Not Specified. 69% White, 14% Asian, 7% BAC, 6% Not Specified, 4% Mixed.

Marketing professionals N. 483 This group is made up of 483 people in Marketing roles. They are aged between 16-64 years. They completed the W-G III between June 2011 and March 2016.	52% Male, 45% Female, 3% Not Specified. 80% White, 8% Asian, 6% Not Specified, 3% BAC, 3% Mixed.
Medical professionals N. 589 This group is made up of 589 people in Medical positions. They are aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	58% Female, 40% Male, 2% Not Specified. 82% White, 11% Asian, 4% Not Specified, 2% BAC, 1% Mixed.
Retail Sales N. 437 This group is made up of 437 people in Retail Sales. They are aged between 16-64 years. They completed the W-G III between June 2011 and March 2016.	59% Male, 38% Female, 3% Not Specified. 75% White, 13% Asian, 5% BAC, 5% Not Specified, 2% Mixed.
Self Employed N. 506 This group is made up of 506 people who are Self Employed. They are aged between 16-64 years. They completed the W-G III between June 2011 and March 2016.	58% Male, 39% Female, 3% Not Specified. 78% White, 13% Asian, 4% Not Specified, 3% BAC, 2% Mixed.

Teachers N. 367 This group is made up of 367 Teachers. They are aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	52% Male, 45% Female, 3% Not Specified. 76% White, 8% Asian, 7% BAC, 7% Not Specified, 2% Mixed.
Transportation N. 209 This group is made up of 209 people in Transportation roles. They are aged between 16-64 years. They completed the W-G III between June 2011 and March 2016.	72% Male, 26% Female, 2% Not Specified. 83% White, 9% Asian, 3% Mixed, 3% Not Specified, 2% BAC.
Position Norm Groups	**Gender and Ethnicity Breakdown**
Directors N. 1198 This group is made up of 1198 people in Director roles. They are aged between 20-64 years, the majority (76%) being between the ages of 40-59 years. They completed the W-G III between June 2011 and March 2016.	64% Male, 32% Female, 4% Not Specified. 86% White, 6% Not Specified, 5% Asian, 2% BAC, 1% Mixed.
Senior Managers N. 933 This group is made up of 933 people in senior manager roles. They are aged between 20-64 years. They completed the W-G III between June 2011 and March 2016.	56% Male, 40% Female, 4% Not Specified. 82% White, 8% Asian, 6% Not Specified, 2% BAC, 2% Mixed.

Managers N. 1476 This group is comprised of 1476 people in managerial roles who took the W-G III. They are aged between 20-64 years. They completed the test between June 2011 and March 2016.	53% Male, 43% Female, 4% Not Specified. 81% White, 8% Asian, 6% Not Specified, 4% BAC, 1% Mixed.
Professional contributors N. 7274 This group is made up of 7274 professional contributors. They are aged 16-64 years. They completed the W-G III between June 2011 and March 2016.	50% Male, 46% Female, 4% Not Specified. 73% White, 13% Asian, 8% Not Specified, 4% BAC, 2% Mixed.
UK Population N. 5689 This group is representative of the UK general working population. It includes 5689 people who have taken the Watson Glaser. They are aged between 16-64 years. They completed the Watson Glaser between June 2011 and March 2016.	Female 51%, Male 49%. 86% White, 8% Asian, 3% BAC, 2% Mixed, 1% Other. 17% 16-19 years 14% 20-24 years 12% 25-29 years 9% 30-34 years 8% 35-39 years 10% 40-44 years 12% 45-49 years 11% 50-54 years 5% 55-59 years 2% 60-64 years