



Watson-Glaser Critical Thinking Appraisal III (WG-III)

Contents

Introduction ▶

**About efficacy
reporting at Pearson ▶**

Executive summary ▶

**How did evidence inform
the design of Watson-Glaser? ▶**

**What does the evidence say
about Watson-Glaser? ▶**

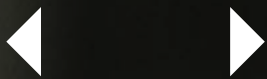
Watson-Glaser in action ▶

Discussion ▶

References ▶



Introduction



In 2013, we were the first company to make a commitment to measure our impact on some of the outcomes that matter most such as academic achievement. But there was no rule book and no model to follow. We've had to carve our own path to define what efficacy looks like in education.

While our approach is rigorous, the concept underlying it is simple: we use evidence and research to design products and solutions to help individuals achieve the outcomes that matter to them. Then, we measure the impact of using our products, report that impact in a transparent way, and use what we learn to help individuals – and ourselves – continuously improve.

Today, we are taking what we have learned and evolving our approach. We are focusing more on designing products to have a measurable impact, not just during education, but on employability and lifelong learning as well.

We want our commitment to efficacy to be a reason for individuals to believe in Pearson, to see us as their trusted guide to lifelong learning, as they navigate a changing world of work. Skills that are hard to automate, like communication and critical thinking, are in more demand than ever. And now that the idea of a job for life is gone, people need to continuously grow, demonstrate their skills and adapt their talent.

People need a lifetime of learning and so we must refocus and redesign learning. The way we learn needs to support the development of the key skills people need to thrive today and in the future.



Efficacy in 2020

A critical segment of Pearson's portfolio is its Assessment business. This report on the Watson-Glaser Critical Thinking Appraisal is part of our ongoing commitment to communicate about our impact in a transparent way for our assessment offerings. Watson-Glaser measures a person's critical thinking skills -- their ability to question assumptions, objectively evaluate information and arguments, and make logical and well-informed decisions.

Our commitment to efficacy is on-going and all our 2020 efficacy reports are available on pearson.com/news-and-research/efficacy

Special thanks



We want to thank all the customers, test-takers, research institutions and organizations we have collaborated with to date. If you are interested in partnering with us on future efficacy research, have feedback or suggestions for how we can improve, or want to discuss your approach to using or researching our assessments, we would love to hear from you at efficacy@pearson.com.

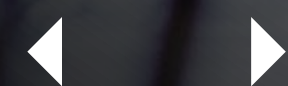
Kate Edwards, PhD

*SVP Efficacy & Learning
Pearson*

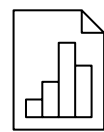


Learn more about the
processes and principles
for efficacy [here](#)

About efficacy reporting at Pearson

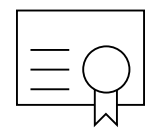


As part of our commitment to being open and transparent about how we design, develop, and evaluate the impact of use of our products on learning, we produce a range of efficacy publications, including reports and guides. This report is one of our *Assessment Reports*.



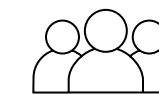
Technical Research Reports ¹⁻²⁻⁴

These describe a single piece of impact evaluation research into the use of a product, undertaken to meet the standards expected for publication in a peer-reviewed academic journal. Selected statements in our Technical Research Reports are independently assured by Pricewaterhouse Coopers (PwC).



Product Guides ³⁻⁴

These explain what the evidence about a single product means for users of that product. Product Guides combine research findings with stories from real users to help you replicate best practice with the product and achieve the best outcomes for learners.



Assessment Reports ³⁻⁵

These summarize the evidence about a single assessment's capability to measure a trait or ability in a valid, reliable and fair manner. These reports are not independently assured, because we do not expect assessments to have a direct effect on outcomes for individuals.



Product Efficacy Reports ¹⁻³⁻⁴

These summarise all the relevant impact evaluation research related to the use of a single product. This includes research described in Technical Research Reports and learning research that informed the product's design and use. Selected statements in our Product Efficacy Reports are independently assured by PwC.



Qualification & Certification Reports ³⁻⁵

These reports include information about how the design of the qualification or certification was informed by research. They bring in evidence about how the qualification is delivered, and how it supports experience and progression. It summarizes relevant Technical Research Reports associated with the assessment of the qualification and impact evaluation research related to learner outcomes.

Key

- ¹ – Independently assured by PwC
- ² – Details a single study
- ³ – Summarizes all relevant evidence
- ⁴ – Evaluates impact on learner outcomes
- ⁵ – Evaluates assessment quality indicators: validity, reliability, and fairness



Efficacy and assessments

Usually when we talk about the efficacy of a product, we mean the impact of its use on outcomes for individuals, like achievement and progression.

Assessments are a little different. Taking a test may not be a learning experience in itself, but test results can be used to make decisions about an individual, such as readiness for an educational program or suitability for a job. So the efficacy of an assessment is the extent to which it provides an accurate snapshot of what an individual knows and can do.

We judge the efficacy of assessments like Watson-Glaser against three Assessment Quality Indicators (AQIs): validity, reliability, and fairness. These factors help us evaluate if the assessment gives an accurate picture of the individual's knowledge and capabilities.



Validity

Validity depends on evidence that the assessment is suitable for a **specific intended purpose**, and that we can interpret the results as intended. Validity is always context-sensitive; we cannot say that an assessment is or is not valid, period, only that it is or is not valid *for a particular purpose*.



Reliability

Reliability depends on evidence that the **results stay consistent** over time, over multiple forms of the assessment, and/or over multiple scorers.



Fairness

Fairness depends on evidence that the **results mean the same thing for all intended test-takers**. This means it is not systematically biased against any group of test-takers and the way it is administered does not hinder any test-takers in demonstrating their ability in the area being assessed.

Our AQIs are based on attributes defined in the Standards for Educational and Psychological Testing, developed in 2014 by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). These standards have long been recognized as best practice for both developing and evaluating assessments, and play a role in legal defences of assessment.



Executive Summary



Executive summary

Critical thinking skills are essential for success in education and the workplace. Mentions of critical thinking as a requirement in job postings in the United States more than doubled between 2009 and 2014 (Korn, 2014). For many, the purpose of earning a post-secondary degree is to learn the knowledge and skills, like critical thinking, to demonstrate they are qualified for employment. Recognizing this, 92% of teachers in an international survey identified critical thinking as one of the most important skills needed for success in higher education (Stewart, 2014).

The Watson-Glaser Critical Thinking Appraisal is a verbal ability test measuring critical thinking ability. It assesses an individual's ability to analyze, interpret, and draw logical conclusions from written information — critical thinking skills. It addresses the challenges of higher education professionals, HR professionals and employers by giving them the means to measure critical thinking ability as a first step toward supporting its development.

The structure of the Watson-Glaser assessment and its scoring was informed by the RED model, which breaks critical thinking down into the ability to recognize assumptions, evaluate arguments, and draw conclusions. This report summarizes the studies that demonstrate the validity, reliability and fairness of the test in measuring critical thinking.

In addition, the way an assessment is implemented has as much influence on its efficacy as its design, so we also interviewed customers to investigate how they are using Watson-Glaser in real situations.

“We have made it part of the campus culture ... this is what we do here”.

[How the University of South Florida uses Watson-Glaser](#)

"Watson Glaser proves to them that they can become better thinkers".

[How the US Air Force uses Watson-Glaser](#)

Visit the product website: talentlens.com/watson-glaser-critical-thinking-test



How did evidence inform the design of Watson-Glaser?



The Watson-Glaser Critical Thinking Appraisal is a verbal ability test measuring critical thinking ability. It assesses an individual's ability to analyze, interpret, and draw logical conclusions from written information.

In this section

- 1** The importance of critical thinking
- 2** The RED model
- 3** Over-exposure
- 4** Biases
- 5** Norms
- 6** History and reach of Watson-Glaser

This ability is a crucial stepping stone to logical thinking, decision making, and problem-solving, all of which are essential not just for academic and career success, but for becoming a positive participant in our society. Thousands of employers, colleges and schools around the world use Watson-Glaser to identify great employees to hire and high potential employees to develop, and to choose which students to admit into challenging programs.

Watson-Glaser presents test-takers with a series of passages or scenarios, each of which is accompanied by a number of items for them to respond to. The test is completed online, and is suitable for both supervised and unsupervised administration. Most versions of the test are timed, though the recommended time limit is intentionally generous. There is also an optional follow-up interview component, which administrators can use to gain more insight into the raw test scores. Pearson also offers professional development support for the people responsible for interpreting and taking action based on the assessment results.

Organizations use Watson-Glaser scores to select and develop employees for roles that require careful analysis, logical decision making and problem solving, and to predict employees' performance in these roles. In academic settings, Watson-Glaser is used to select students for particular courses.

Our own studies show that Watson-Glaser scores correlate highly with course grades in a number of undergraduate courses (Watson & Glaser, 2019). However, the main application of Watson-Glaser is as a tool for assessing potential employees before making hiring decisions.



The importance of critical thinking

In this section

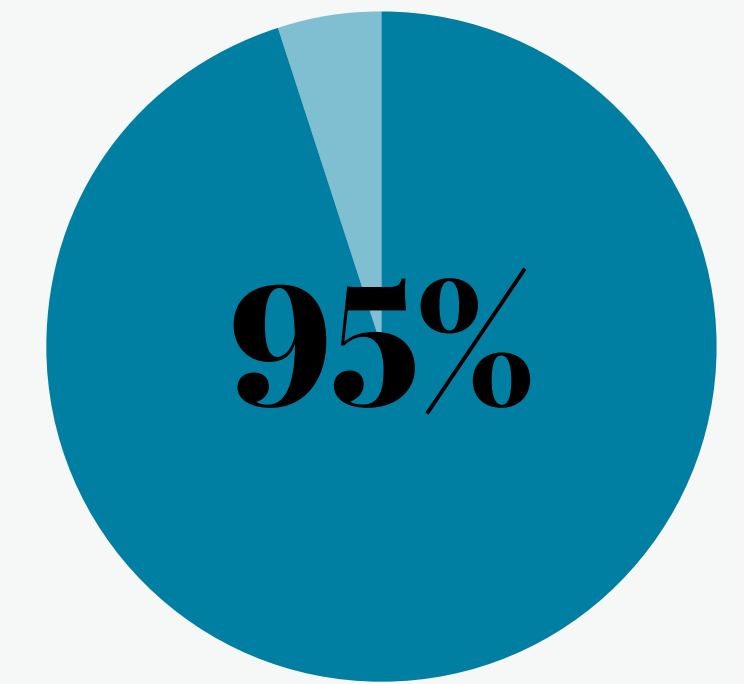
- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 Norms
- 6 History and reach of Watson-Glaser

Critical thinking is what allows us to tell reliable sources from disinformation and allows us to form our own opinions based on the things we read. It's what allows us to make and justify good, fully informed decisions. It is the essential skill for the information age.

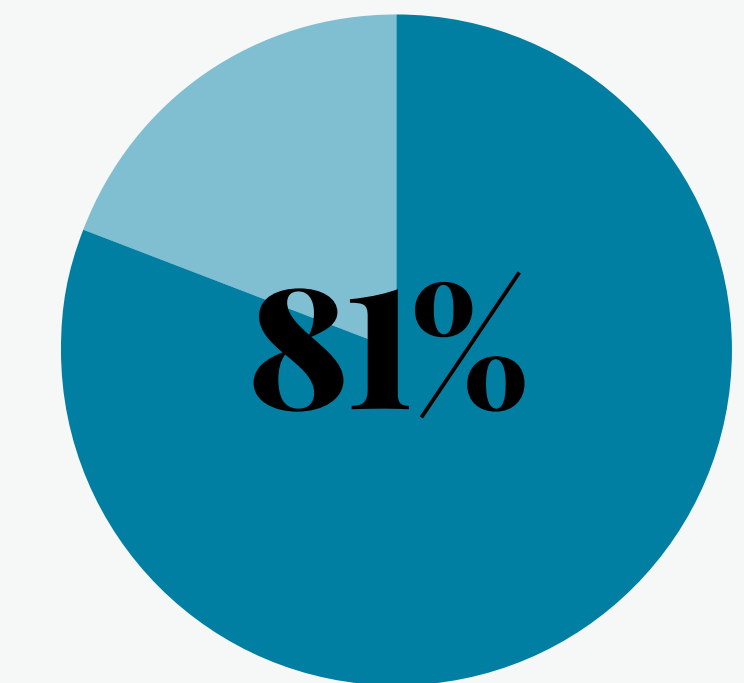
There is a wealth of research and literature on critical thinking and how to define it. According to this body of research, at its core, critical thinking entails questioning assumptions, objectively evaluating information and arguments, and making logical and rational decisions.

Critical thinking is believed to play a central role in everyday workplace and academic skills like logical thinking, decision-making, argumentation, and problem-solving (Butler et al., 2012; Ennis, 1985; Facione, 1990; Halpern, 2003).

Employers recognize these benefits; they look for them in graduates and invest in developing them in existing employees. In surveys, 95% of 128 Australian STEM employers identified critical thinking as “very important” for STEM graduates (Rayner & Papakonstantinou, 2015) and 81% of US employers believed colleges should emphasize critical thinking more strongly (Association of American Colleges and Universities, 2011). Mentions of critical thinking as a requirement in job postings in the USA more than doubled between 2009 and 2014 (Korn, 2014).



of 128 Australian STEM employers identified critical thinking as “very important” for STEM graduates



of US employers believed colleges should emphasize critical thinking more strongly



The importance of critical thinking

The recognition of critical thinking goes beyond HR, through all levels of the organization. In interviews with leaders at 200 companies, critical thinking was among the skills mentioned most frequently as being essential for both academic and career success (Educational Testing Service, 2013); and in a survey by the American Management Association, employers rated critical thinking as the most important of four key managerial competency requirements (American Management Association, 2019).

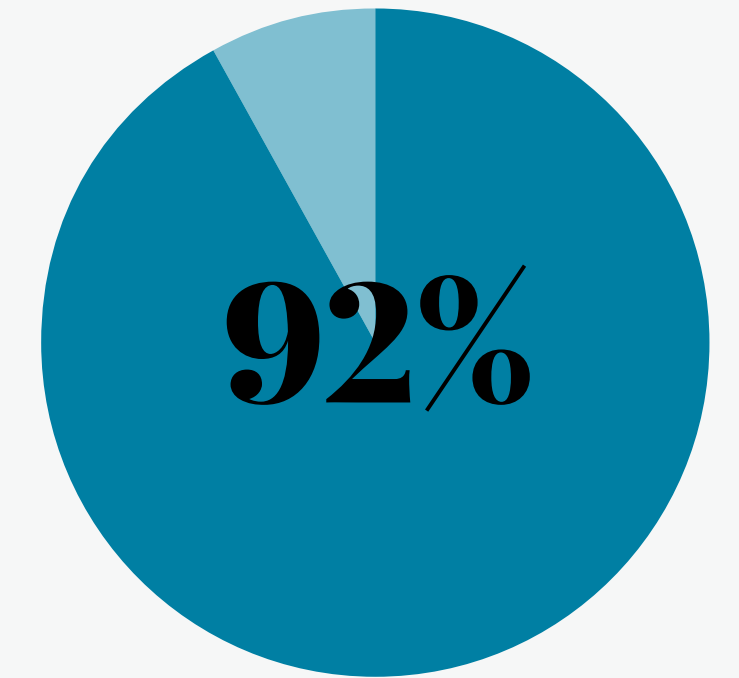
Academic institutions also recognize the importance of critical thinking. In an international survey by the Times Educational Supplement, 92% of teachers identified critical thinking as one of the most important skills needed for success in higher education (Stewart, 2014). The Organization for Economic Co-operation and Development has named critical thinking a core skill for college students all around the world (OECD, 2012), and 95% of chief academic officers from 433 higher education institutions share this priority, rating critical thinking as one of the most important skills for students to acquire (Association of American Colleges and Universities, 2011).

There is evidence that improving their critical thinking ability helps students perform better academically. Students who receive critical thinking instruction have been shown to be more willing to accept scientifically based theories (Rowe et al., 2015), and studies have shown that critical thinking ability predicts college grade point average (ACT, no date).

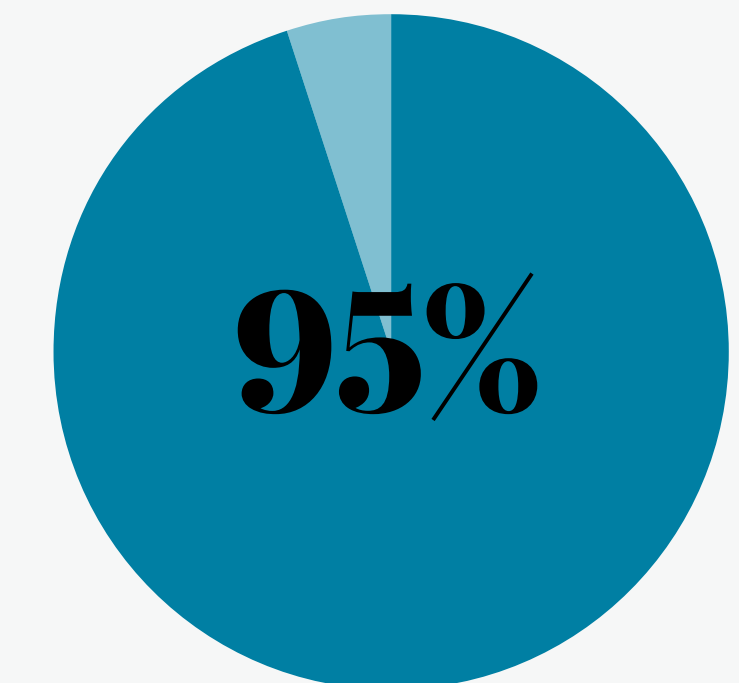
Watson-Glaser supports organizations and institutions that recognize the value of critical thinking, by providing a method of assessing individuals' critical thinking ability relative to their peers.

In this section

- 1 The importance of critical thinking**
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 Norms
- 6 History and reach of Watson-Glaser



of teachers identified critical thinking as one of the most important skills needed for success in higher education



of chief academic officers from 433 higher education institutions rated critical thinking as one of the most important skills for students to acquire

The RED model

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 Norms
- 6 History and reach of Watson-Glaser

For the purposes of designing and developing the Watson-Glaser assessment, critical thinking is defined as the ability to question assumptions, objectively evaluate information and arguments, and make logical and well-informed decisions.

As part of the ongoing development of Watson-Glaser ([see History and reach](#)), Pearson has defined the RED model of critical thinking. This states that critical thinking requires three aptitudes: recognize assumptions, evaluate arguments, and draw conclusions.



Recognize assumptions

It is deceptively easy to listen to a comment or presentation and assume the information presented is true even with no evidence to back it up. Noticing and questioning assumptions helps to reveal information gaps or unfounded logic. We also need to examine assumptions from different viewpoints.

Evaluate arguments

The art of evaluating arguments involves analyzing information objectively and accurately, questioning the quality of supporting evidence, and understanding how emotion influences the situation. Common barriers include confirmation bias, or allowing emotions to get in the way of objective evaluation.

Draw conclusions

People who can arrive at conclusions that logically follow from the range of available evidence are often characterized as having “good judgment”. They are careful not to generalize inappropriately beyond the evidence and they can change their position when the evidence calls for it.

The RED model

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 Norms
- 6 History and reach of Watson-Glaser

Watson-Glaser scores test-takers on five subscales: Recognition of Assumptions, Evaluation of Arguments, Inference, Deduction, and Interpretation. Pearson conducted factor analyses of the original Watson-Glaser assessment and found that the Inference, Deduction, and Interpretation scales factored together. Since the release of the Watson-Glaser II Critical Thinking Appraisal in 2010, the results of these three measures have been combined under the heading Draw Conclusions.

The RED model informs the structure of the assessment and its scoring. Each test-taker is presented with 40 items: 12 testing their ability to recognize assumptions, 12 testing their ability to evaluate arguments, and 16 testing their ability to draw conclusions. As well as an overall score, administrators can review a development report that breaks down test-takers’ scores in each of the three aptitudes separately. They can also choose to share this development report with test-takers.

When the assessment is used for selection purposes, administrators might only compare candidates’ overall scores. But when it is used for development purposes, the development report can reveal areas where the test-taker should concentrate, or might need support, in order to improve their critical thinking skills.

The Watson-Glaser test subscales



Recognition of Assumptions



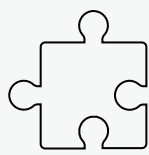
Deduction



Evaluation of Arguments



Interpretation



Inference

Each test-taker is presented with 40 items

12

testing their ability to recognize assumptions

16

testing their ability to draw conclusions

12

testing their ability to evaluate arguments



Over-exposure

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 **Over-exposure**
- 4 Biases
- 5 Norms
- 6 History and reach of Watson-Glaser

When an assessment is identical for all test-takers, items in the assessment can become over-exposed as time goes by. For example, someone who applies to work at two different companies that both use Watson-Glaser could encounter the same test items both times, giving them an unfair advantage in their second test.

This is a particular issue in assessments used for high-stakes purposes like recruitment. High stakes increase the incentive to cheat, and ways of cheating include smuggling item information out of the test. If an item that always or regularly appears in the test becomes publicly available, the test is no longer fair.

Watson-Glaser avoids over-exposure, and so aims to improve fairness, by drawing assessment items from a large item bank. This means each item is presented to test-takers less often overall, and the chances of two test-takers being presented with the same set of items is small.



Biases

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 **Biases**
- 5 Norms
- 6 History and reach of Watson-Glaser

Strong attitudes, opinions, and biases affect some people’s ability to think critically (Klaczynski, Gordon, & Fauth, 1997; Nickerson, 1998; Sa, West, & Stanovich, 1999; Stanovich & West, 1997, 2008; West, Tolplak, & Stanovich, 2008).

Watson-Glaser takes this into account by presenting test-takers with two types of scenario: neutral and controversial. Neutral scenarios deal with subject matter that development testers rate as less controversial than other issues, such as the weather, scientific facts, or common business situations. Controversial scenarios refer to political, economic, and social issues that frequently provoke emotional responses.

Because the assessment includes both neutral and controversial material, the results indicate how well a test-taker can think critically whether or not they have strong feelings about the subject matter.

Watson-Glaser tests users with two different types of scenario



Neutral scenarios

- Weather
- Scientific facts
- Common business situations



Controversial scenarios

- Political
- Economic
- Social issues



Norms & scoring

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 **Norms**
- 6 History and reach of Watson-Glaser

Norms are sets of scores derived from a specific group of test-takers – managers, for example. Comparing an individual test-taker’s score to a relevant norm provides more meaningful information than their raw score alone.

Watson-Glaser presents each test-taker with 40 scenarios. Test-takers’ responses to each item can be either correct or incorrect. Adding up the number of correct responses gives a raw score between 0 and 40.

Raw scores can be used to rank test-takers, but little else can be inferred from them. So Watson-Glaser also allows administrators to evaluate an individual test-taker’s raw score relative to a large sample of others who took the same test. The sample is known as a normative group, and the score derived from their performance is known as a norm.

For example, imagine a test-taker achieves a raw score of 35 out of 40. Without applying a norm, we cannot know whether this is a “good” score or not. But say the assessment is being administered to assess candidates for a managerial position. Comparing this test-taker’s score to Watson-Glaser’s Manager norm, we would see that they have achieved a percentile rank of 86. This means that the test-taker scored equal to or higher than about 86% of managers. This is a more meaningful result, which is more useful for the administrator’s decision making process.

Without applying a norm, we cannot know whether a score is “good” or not.



Norms & scoring

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 **Norms**
- 6 History and reach of Watson-Glaser

The ideal normative group for a given context is one that is representative of those taking the test in that context. The right norm also depends on the reason for administering the assessment. An administrator interested in intelligence testing might want to compare test-takers’ scores to the general population, for example, while one selecting candidates for potential employment might want to compare them to their peers in the profession.

Administrators can use their own Watson-Glaser results to construct their own norms. A sample size of at least 200 test-takers is ideal for constructing a norm. For administrators who do not yet have a big enough sample, or who do not have the resources to construct their own norms, Watson-Glaser comes with 18 different norms constructed by Pearson, including for occupations (accountant, engineer and human resource professional, for example) and positions or organizational levels (from entry-level to executive). These are updated and added to frequently.



History and reach of Watson-Glaser

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 Norms
- 6 **History and reach of Watson-Glaser**

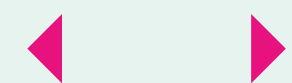
The Watson-Glaser Critical Thinking Appraisal was first published in 1964 by Goodwin Watson and Edward M. Glaser, after a development period dating back to 1926. Since then, it has become the leading method of assessing critical thinking globally. It is available in French, English, Dutch and Spanish, and used in many countries including Australia, Canada, India, France, Japan, The Netherlands, Mexico, Singapore, the USA and the UK.

A revised version of the assessment, the Watson-Glaser II Critical Thinking Appraisal, was released in 2010. This revision incorporated enhancements requested by customers, and introduced the RED model.

Both the original Watson-Glaser and Watson-Glaser II used fixed forms. That is, all people taking the tests responded to the same set of items. The current version of the assessment, the Watson-Glaser III Critical Thinking Appraisal, instead presents each test-taker with a selection of items drawn from a large item bank.

Watson-Glaser III also introduces a more sophisticated scoring system, more contemporary and business-relevant scenarios, enhanced reporting, and other improvements requested by customers.

Today, Watson-Glaser is part of Pearson's TalentLens portfolio of products. Thousands of organizations and schools use Watson-Glaser alongside other TalentLens tests to hire great managers, develop high-potential employees, and admit students into challenging programs.



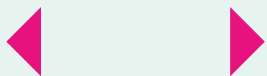
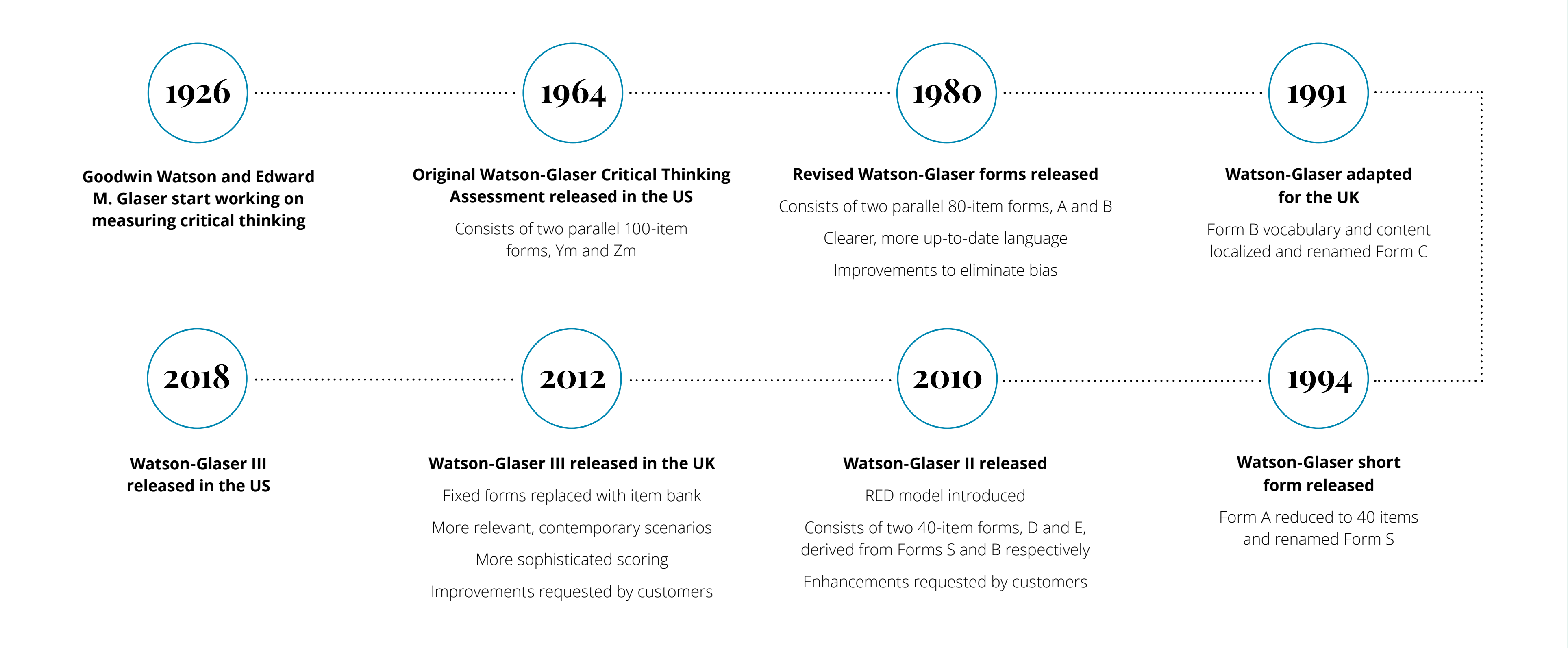
History and reach of Watson-Glaser

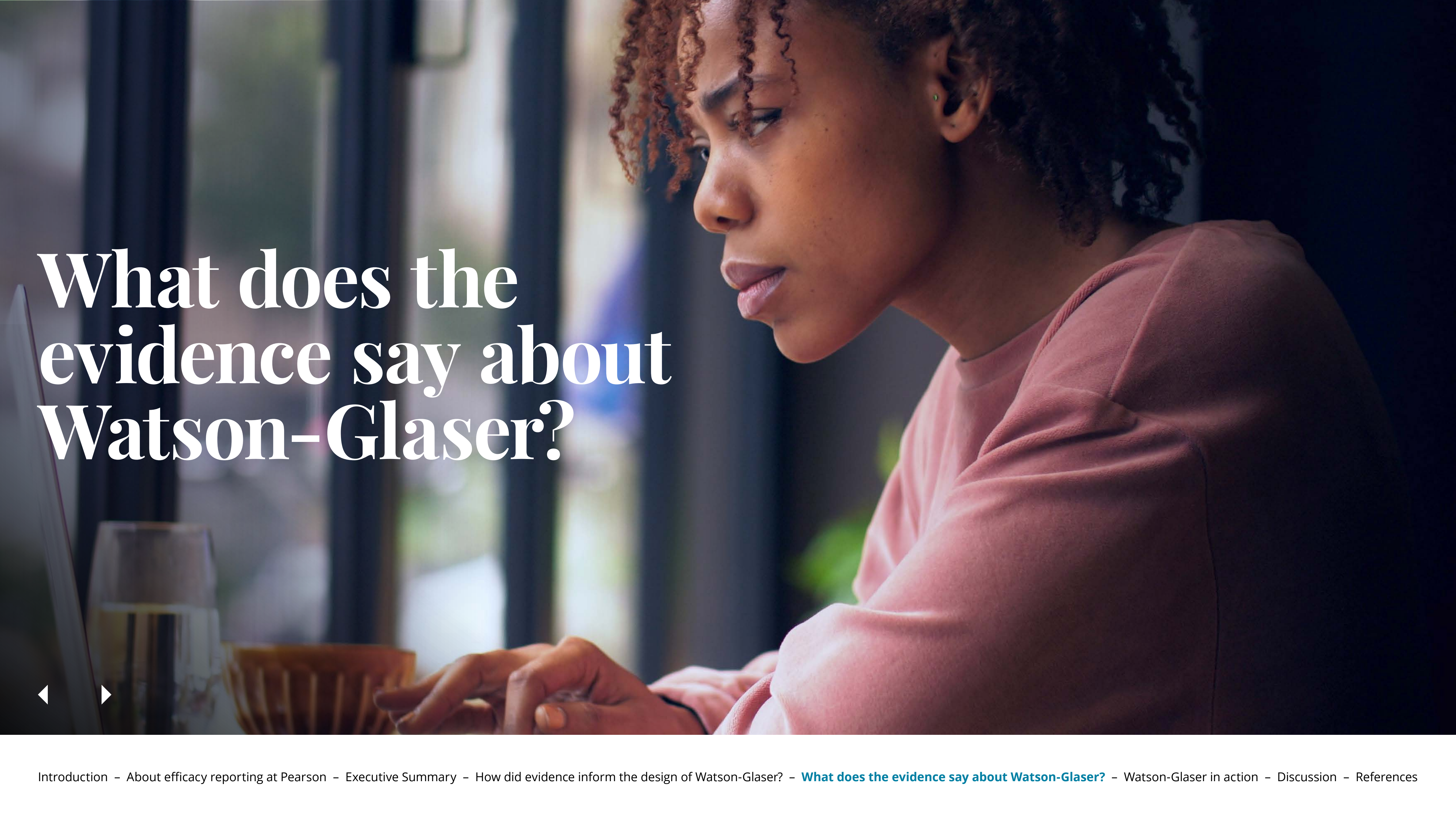
For more details about the development of Watson-Glaser and the ways the assessment can be used, see the technical manual and [frequently asked questions](#) for the product.

The Technical Manual is available for current users or prospective users. Please get in touch with your TalentLens representative to request it.

In this section

- 1 The importance of critical thinking
- 2 The RED model
- 3 Over-exposure
- 4 Biases
- 5 Norms
- 6 History and reach of Watson-Glaser





What does the evidence say about Watson-Glaser?



The efficacy of an assessment considers its capability to measure a trait or ability in a valid, reliable, and fair manner. The objective of Watson-Glaser is to measure an individual's critical thinking ability: their ability to look at a situation and clearly understand it from multiple perspectives while separating facts from opinions and assumptions. To determine the efficacy of Watson-Glaser, we need to review evidence related to its capability to measure a trait or ability in a valid, reliable, and fair manner.

In this section

- 1** Validity
- 2** Reliability
- 3** Fairness

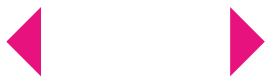


We judge the efficacy of assessments like Watson-Glaser against three Assessment Quality Indicators (AQIs): validity, reliability, and fairness.

In this section 1 Validity 2 Reliability 3 Fairness	Validity	Can the assessment be used for its intended purpose, and can we interpret the results as intended?	Watson-Glaser’s intended purpose is to help organizations select candidates for employment and development, and to help academic instructors select students for particular programs. The results are intended to indicate how well test-takers can recognize assumptions, evaluate arguments, and draw conclusions, in order to measure their overall critical thinking ability. We look at how the evidence supports this indicator.
	Reliability	Are the results consistent over time, over different forms of the assessment, and/or over different scorers?	To measure this, researchers evaluate if all the items in the test measure the same thing in the same way. Researchers also measure reliability using the test-retest method, where the same participants take the same test on two separate occasions to see if they get a consistent score.
	Fairness	Can the results be used the same way for all test-takers?	For all test-takers, the results of Watson-Glaser should indicate critical thinking ability relative to an appropriate norm. Included are multiple types of evidence and methods to demonstrate this.

Different versions of the assessment

This report collects evidence spanning a number of years and all three versions of the Watson-Glaser assessment. Except where specifically stated otherwise, the findings presented here relate to elements of the assessment that have not changed between versions. Therefore we have not noted which version of Watson-Glaser specific findings relate to. If you are interested in finding out which studies and findings relate to which version of Watson-Glaser, please request the technical manual for the assessment.



Validity

In this section

1 Validity

2 Reliability

3 Fairness

Can the assessment be used for its intended purpose, and can we interpret the results as intended?

There is evidence that the Watson-Glaser Critical Thinking Appraisal measures the cognitive abilities that underlie critical thinking skills, and that scores on the assessment can predict attainment in education and in the workplace, as summarized in the studies below.

Can Watson-Glaser scores be interpreted as a measure of critical thinking ability?

To start investigating whether Watson-Glaser can be used for its intended purpose – to measure critical thinking ability – we need to assess whether it is structured in a way that supports this purpose. Two separate **confirmatory factor analyses** provide evidence that Watson-Glaser’s RED model of critical thinking (recognize assumptions, evaluate arguments, draw conclusions) supports the assessment’s intended purpose (Watson & Glaser, 2019).

Confirmatory factor analysis is used to test how well a theoretical model explains the relationships between variables. Pearson carried out two of these analyses during the development of Watson-Glaser II, one at the tryout stage and one during standardization. Compared to two other possible models – one with critical thinking as the only factor, and one using the five subscales from the original Watson-Glaser ([see The RED model](#)) – these analyses confirmed that the RED model is the most valid way to interpret results from Watson-Glaser.



Validity

In this section

1 Validity

2 Reliability

3 Fairness

To continue investigating whether Watson-Glaser can be used to measure critical thinking ability, we can compare results from Watson-Glaser to the results of other tests intended to measure the same or similar things. Positive correlations between the results provide evidence in support of **convergent validity**.

Confirmatory factor analysis confirmed that the RED model is the most valid way to interpret results from Watson-Glaser.

Over the years, studies have revealed positive correlations between results from Watson-Glaser II and a number of tests measuring cognitive ability, including achievement tests and reasoning tests. Scores range from 0.39 to 0.70, on a scale of 0.0 to 1.0, where 1.0 would mean the two scores were measuring exactly the same thing (Watson & Glaser, 2019). For example:

- Watson-Glaser has a 0.53 correlation with scores on Raven's Advanced Progressive Matrices (Watson & Glaser, 2006).
- A study involving 91 working adults in the USA compared Watson-Glaser III with the Numerical Data Interpretation Test (NDIT), a measure of numerical reasoning. The study produced a correlation of 0.47. Because both tests measure reasoning ability, we might expect them to be related; but NDIT uses mainly numerical content while Watson-Glaser is mainly a verbal reasoning test, so we would only expect the relationship to produce a moderate correlation (Pearson, 2017).

Additional correlations are included in the technical manual.

Studies comparing Watson-Glaser scores with other measures of cognitive ability have produced moderate to high correlations: 0.39-0.70.



In this section

1 Validity

2 Reliability

3 Fairness



Can organizations use Watson-Glaser scores as intended?

Watson-Glaser is intended to help organizations select and develop employees. To investigate whether Watson-Glaser is suitable for this purpose, we can compare the results of the assessment with indicators of on-the-job performance – such as job performance ratings, supervisor ratings, and training course grades. Positive correlations between the results provide evidence in support of ***criterion-related validity***.





Table 1: *Correlations between Watson-Glaser scores and on-the-job performance indicators*

Comparison indicator	Correlation
Leadership and middle management assessment center ratings of judgment, analysis, openness to experience and similar	0.16-0.58
Supervisory ratings of judgement, decision-making, problem-solving and similar	0.23-0.44
Level in organization attained	0.33

Studies have compared Watson-Glaser scores with a number of on-the-job performance indicators, including performance in assessment center exercises, supervisory ratings of critical thinking behavior, and organization level attained. The correlations found in these studies range from 0.16 (depends on the circumstances) to 0.58 (very beneficial) (Watson & Glaser, 2019). See table 1.

We use the US Department of Labor’s guidelines to judge how strongly a correlation supports the criterion-related validity of an assessment.

Table 2: *US Department of Labor criterion-related validity guidelines (US Department of Labor, 1999)*

Correlation coefficient score	Usefulness interpretation
 > 0.35	Very beneficial
 0.21-0.35	Likely to be useful
 0.11-0.20	Depends on the circumstances
 < 0.11	Unlikely to be useful

We would expect Watson-Glaser scores to correlate more strongly with measures of cognitive ability, because the things they are measuring should be very similar. The correlations with on-the-job performance measures still suggest that test-takers who perform well on the Watson-Glaser are also likely to perform well at work.

In this section

1 Validity

2 Reliability

3 Fairness

Can academic institutions use Watson-Glaser scores as intended?

Watson-Glaser is intended to help academic institutions select students who are likely to do well on coursework. To investigate whether Watson-Glaser is suitable for this purpose, we can compare the results of the assessment with other indicators of course success. Positive correlations between the results provide evidence in support of **crit**erion-related validity.

Studies have compared Watson-Glaser scores with final course grades in both a business degree and the Bar Professional Training Course, a post-graduate vocational training course for aspiring barristers in England and Wales. The correlations found in these studies range from 0.38 to 0.62, suggesting that Watson-Glaser is a very beneficial predictor of likely course success (Watson & Glaser, 2019). Table 3 shows these correlations with final course grades, along with correlations found with other indicators of academic success.

Table 3: Summary of correlations between Watson-Glaser scores and indicators of academic success (Watson & Glaser, 2019)

Correlation	Comparison indicator	Study sample
0.28	Grade point average	139 educational psychology students
0.30	Grade point average	147-194 education students
0.38	Final course grades on a business degree	Business school students
0.41	Grade point average	114 education students
0.42	Exam 1 score	158-164 educational psychology students
0.51	Semester 1 grade point average	37 first year students on a Pennsylvania, USA nursing program
0.53	Semester 1 grade point average	31 first year students on a Pennsylvania, USA nursing program
0.57	Exam 2 score	158-164 educational psychology students
0.59	Semester 1 grade point average	41 first year students on a Pennsylvania, USA nursing program
0.62	Final course grades	123 legal training course students

Correlation coefficient score	Usefulness interpretation
> 0.35	Very beneficial
0.21-0.35	Likely to be useful
0.11-0.20	Depends on the circumstances
< 0.11	Unlikely to be useful

The correlations range from 0.38 to 0.62, suggesting that Watson-Glaser is a very beneficial predictor of likely course success.



Reliability

Are the results consistent over time, over different forms of the assessment, and/or over different scorers?

There is evidence that Watson–Glaser scores are internally consistent, and consistent both over time and across different forms of the assessment.

We use the US Department of Labor’s guidelines to judge how strongly a correlation supports the reliability of an assessment.

Table 4: Summary of reliability scores for Watson-Glaser

Limited applicability	Adequate	Good	Excellent
		Correlations between testing occasions (0.73-0.89)	
		Consistency of test items (0.83)	
		Correlations between different forms of the test (0.82-0.88)	

Table 5: US Department of Labor criterion-related reliability guidelines (US Department of Labor, 1999)

Correlation coefficient	Interpretation
< 0.70	May have limited applicability
0.70-0.79	Adequate
0.80-0.89	Good
> 0.89	Excellent

In this section

1 Validity

2 Reliability

3 Fairness



Reliability

In this section

1 Validity

2 Reliability

3 Fairness

Are Watson-Glaser’s results internally consistent?

An assessment is *internally consistent* if all the items in the test measure the same thing in the same way. We measure internal consistency using Cronbach’s alpha, an index that measures how closely related a set of items are as a group.

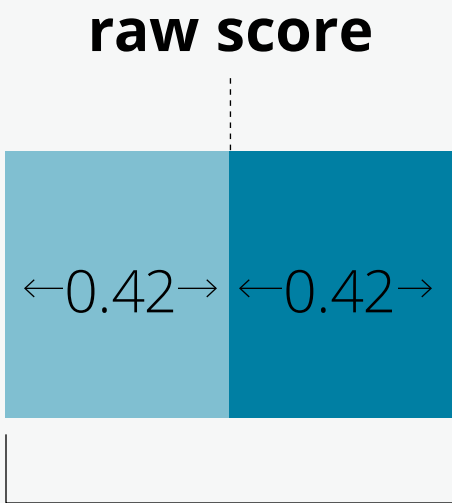
Cronbach’s alpha was computed based on a sample of 147 working adults in the US. This analysis put Watson-Glaser III’s internal consistency at 0.83 – a good level of reliability (Watson & Glaser, 2019).

No test is 100% accurate, so all test scores are estimates of the test-taker’s “true” score. The standard error of measurement (SEM) is a measure of the accuracy of this estimate. About 68% of the time, the raw score is within ± 1.0 SEM of the ‘true’ score, and about 96% of the time, the raw score is within ± 1.96 SEM of the ‘true’ score.

The SEM for Watson-Glaser III was estimated to be 0.42. This means that if a test-taker achieves a raw score of 25, for example, we can be:

- 68% confident that their ‘true’ score is 25 ± 1.0 SEM (0.42) – so between 24.58 and 25.42
- 96% confident that their ‘true’ score is ± 1.96 SEM ($1.96 \times 0.42 = 0.82$) – so between 24.18 and 25.82

Tests on previous versions of Watson-Glaser have placed its internal consistency between 0.75 and 0.86 (adequate to good) and the SEM between 0.32 and 3.6 (Watson & Glaser, 2019).



68%

confidence that the 'true' score lies within this band



96%

confidence that the 'true' score lies within this band



Reliability

In this section

1 Validity

2 **Reliability**

3 Fairness

Are Watson-Glaser results consistent over time?

We can investigate this question using the test-retest method, where the same participants take the same test on two separate occasions (assuming the participants' critical thinking skills do not improve between occasions). If there is a positive correlation between the scores from the two occasions, we can say that Watson-Glaser has ***test-retest reliability***.

A number of studies using various versions of Watson-Glaser produced correlations between testing occasions ranging from 0.73 to 0.89 – adequate to excellent test-retest reliability (Watson & Glaser, 2019).

Are Watson-Glaser results consistent over different forms of the test?

Watson-Glaser III presents test-takers with 40 items randomly selected from a large bank. It is important to test whether different selections of items lead to similar scores. To investigate this, we can build two different tests, each using a different selection of items from the bank, ask a group of study participants to complete both tests, and compare their results. If there is a positive correlation between the scores on the two tests, we can say that Watson-Glaser has ***alternate form reliability***.

Two such studies in the UK, one with 355 participants and one with 318, produced correlations of 0.82 and 0.88 respectively – good alternate form reliability (Watson & Glaser, 2019).



Fairness

In this section

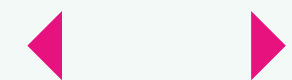
- 1 Validity
- 2 Reliability
- 3 **Fairness**

There is evidence that Watson-Glaser does not favor or disadvantage any particular group of test-takers in any way that could influence the real-world decisions the test is intended to support about staffing and development. The evidence also shows that the assessment can be provided in different modes, to suit different administrators' and test-takers' needs, without influencing the results. We summarize this evidence and mitigation strategies in this section.

What steps does Pearson take to mitigate against test bias when writing the test items?

The items in the test are developed and reviewed by experts who have decades of test development experience and are trained to create content that is not biased against protected groups. The development team includes an industry expert on equal opportunities and fairness in testing, who helps to ensure the content is developed with these crucial issues in mind.

After the item development and review phase, each item is trialled on large groups of participants. We then perform a statistical analysis to review the quality of the items. Those found to potentially be biased against a protected group are excluded from the final test.



Fairness

Do groups perform differently on the Watson-Glaser test?

To investigate whether or not particular groups of people perform differently on the Watson-Glaser test, we can break down the groups of test takers by different demographic factors and compare their Watson-Glaser results.

The difference between the groups can be shown as a Cohen’s *d* statistic. A Cohen’s *d*:

- above 0.8 indicates a large difference between the groups
- above 0.5 is a moderate difference
- above 0.2 is a small difference
- below 0.2 is a negligible difference

These are not just theoretical or academic measurements; when the test is used to make hiring decisions, as Watson-Glaser is intended to be, a large difference can have material effects on the lives of people from the affected groups.



In this section

- 1 Validity
- 2 Reliability
- 3 **Fairness**



Fairness

In this section

- 1 Validity
- 2 Reliability
- 3 **Fairness**

All study samples are different, so finding differences in how particular groups of test-takers perform does not necessarily mean that difference will always appear. If studies consistently find differences between groups, this does not necessarily mean that the test favors or disadvantages certain groups; it could indicate that some other related factor is having an effect. That means we cannot conclude that the test is biased unless the source of the performance differences can be traced to the design of the test.

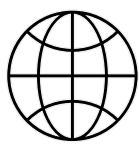
As such, several different samples of UK test-takers were combined and analyzed to find evidence about Watson-Glaser's fairness. The differences between those with and without disabilities, between men and women, and between different age groups were small to negligible, and there is evidence that differences between age groups were related to sampling factors, not to the design of the assessment (Watson & Glaser, 2019).



Fairness

Students whose primary language is English performed better than those for whom English is a second language. Taking the Watson-Glaser test involves a good deal of reading, so some difference is to be expected; in fact, the difference found was small to moderate. Watson-Glaser is available in a number of languages, and is written for readers at or below the 9th grade reading level. The Watson-Glaser technical manual recommends administering the assessment in the test-taker’s first language or, if this is not possible, to take this into account when interpreting the results.

When the sample was divided into White, Black, and Asian students, Black students tended to have the lowest scores. These differences are in line with differences found on other cognitive ability tests (Hough, Oswald, & Ployhart, 2001). Neisser et al. (1996) cite factors such as socio-economic and cultural differences between groups as explanations for findings such as this. Other research has identified test taker concern with conforming to negative racial stereotypes on test performance (Steel and Aronson, 1995), whilst more recent research has uncovered differences in familiarity with ability tests between groups as a possible explanation (Hinton, 2015). Nonetheless, we wanted to explore this finding further to ensure no bias exists on any ethnic group.



The Watson-Glaser test and profile reports are available in following languages:

- US English
- UK English
- Australian English
- Indian English
- French
- French Canadian
- Castilian Spanish
- US Spanish
- Dutch

Table 6: Summary of fairness evidence for Watson-Glaser

Negligible difference	Small difference	Moderate difference	Large difference
With and without disabilities (-0.18-0.09)			
Men and women (-0.06-0.37)			
16-24 year olds and 45+ year olds (-0.37-0.32)			
	Primary language (0.28-0.75)		
			Ethnic groups (0.75-1.31)

In this section

1 Validity

2 Reliability

3 Fairness



Fairness

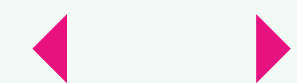
In this section

- 1 Validity
- 2 Reliability
- 3 **Fairness**

We therefore carried out a hierarchical regression on students taking a vocational law course. This analysis involved examining ***differential validity***, which is the extent to which the relationship between test scores and course performance is consistent across different groups. A test is biased if it is not similarly predictive of performance across groups -- that is, if those from different groups who have the same score on the test are not equally likely to succeed. After completing this analysis, we found that the Watson-Glaser test scores were consistent with course performance, regardless of ethnicity, indicating that there is no evidence Watson-Glaser is biased against a particular group.

The analysis also confirms that there is no evidence of bias against men or women, against people whose primary language is not English, or against people with disabilities. The analysis indicates that the test marginally favored younger candidates, but the effect size was so small that it is unlikely to have a meaningful real-world effect (Watson & Glaser, 2019).

After completing differential validity analysis, we found that the Watson-Glaser test scores were consistent with course performance, regardless of ethnicity, indicating that there is no evidence Watson-Glaser is biased against a particular group.



Fairness

In this section

- 1 Validity
- 2 Reliability
- 3 **Fairness**

While group differences in average scores can be justified if analyses like this find no evidence of statistical bias in a test, differences in actual pass rates between groups can be mitigated by careful selection of pass marks. ***Customers are advised to review pass rates for different groups and select a pass mark with minimum pass rate differences across groups.***

Table 6 summarizes these studies and results for Watson-Glaser.

Looking at all the research together, there is evidence that Watson-Glaser does not favor or disadvantage any particular group of test-takers in any way that could influence the real-world decisions, although it does show group differences, so users should consider this during decision making for staffing and development.

In addition, it is important for users to take each test-taker's individual attributes, and how those attributes may interact with the testing context, into consideration when interpreting test scores. And these scores should form only one part of the judgment and decision-making around selection for employment or educational opportunities.



Fairness

In this section

- 1 Validity
- 2 Reliability
- 3 **Fairness**

Does the way Watson-Glaser is administered hinder any test-takers in demonstrating their skills?

The form of the test – for example whether it is administered on paper or online – should not affect a test-taker’s results. If we have the same participants sit the two different forms of the test, and the correlation between the two sets of results is similar to the correlation found when participants sit exactly the same form of the test twice, it suggests that the different forms of the test are having little effect on the results.

Watson-Glaser can be administered with or without supervision, and with or without a time limit. For example, in a hiring context, Watson-Glaser is primarily administered without supervision, to screen out individuals who do not demonstrate the required level of critical thinking ability. While Watson-Glaser III is an exclusively online assessment, Watson-Glaser II is still available as a supervised online or pencil and paper test.



Fairness

There is evidence that the different ways of administering Watson-Glaser do not help or hinder any test-takers in demonstrating their skills. This means test-takers who need to use a particular mode or version of the test, who only have access to it in one mode, or who require accommodations, are not hindered compared to other test-takers. Table 7 summarizes the correlations.

Table 7: *The summary of correlations of test forms for Watson-Glaser.*

Correlation	Test form	Findings
0.73-0.89	Test-retest reliability (benchmark correlation)	This shows how similar the results are when the same participants take the same form of the test twice, on two different occasions.
0.87	On paper vs online	This shows how similar the results are when the same participants take two different forms of the same test: once on paper, and once online.
0.73	Timed vs untimed	This shows how similar the results are when the same participants take two different forms of the same test: once with a time limit, and once without.

In this section

1 Validity

2 Reliability

3 Fairness



Fairness

In this section

- 1 Validity
- 2 Reliability
- 3 **Fairness**

A 2005 study investigated whether administering Watson-Glaser on paper or online has any effect on the results. The study involved 226 adult participants from a variety of occupations, divided into two groups. One group took the Watson-Glaser assessment on paper and then online, and the other group took it online and then on paper. The study found a strong correlation (0.87) between the results from the two different ways of taking the test (Watson & Glaser, 2006).

The time limit for completing Watson-Glaser is designed to be generous, because the test is intended to measure ability, not speed. There is evidence that the time limit has no significant effect on the results for test-takers, suggesting that Watson-Glaser is indeed measuring ability, and not inadvertently measuring speed as well.

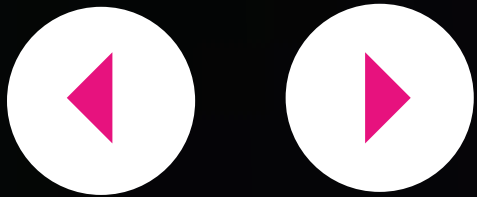
The study involved 137 participants, who each took the test twice online, once with a 30-minute time limit and once with no time limit. To counterbalance the study, half took the timed version first, and half took the untimed version first. There was no significant difference between average scores for the timed and untimed assessments. All except one participant completed the assessment fully both times. The single exception completed 39 of the 40 items when timed, and all 40 when not timed (Watson & Glaser, 2019). This suggests that test-takers who read more slowly are not hindered, for example, and that if a test-taker requires an extended time limit as a disability-related accommodation, this would not be expected to affect their results.

The correlations in both these studies were very close to the correlation found in a simple test-retest study of Watson-Glaser, where participants took the same form of the test both times (Watson & Glaser, 2019).





Watson-Glaser in action



For higher education

Institution	Length of time using Watson-Glaser	About
University of South Florida (USF)	4 years	Jenny Post is the administrator for the university's Incredible Critical Thinking Program and Michael Gillespie is the Program's Director and an Associate Professor of Psychology. Together, they talk about how asking their students to complete Watson-Glaser assessments has helped embed critical thinking skills as a core component of the university's offering.

In this section

1 For higher education

— University of South Florida (USF)

— Ritsumeikan Asia Pacific University

2 For professional development

3 For recruitment

Critical thinking at the heart of USF

Mike and Jenny are in agreement that critical thinking is an absolutely key skill required in the modern world. It is also the most important skill employers look for and lack of it amounts to a critical skills gap for students moving into the world of work.

USF uses Watson-Glaser to test students at the beginning and end of their programs, allowing them to monitor student improvement and evaluate how their courses have improved students' critical thinking skills.

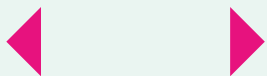
Why Watson-Glaser

Jenny explains that in practical terms, the test's automatic scoring is essential given that USF requires all students to complete the test. She also highlights the development report as a real strength because it uses straightforward, everyday language that is easy for students to understand.

“It gives them a map for improvement.”

Mike feels that the pedigree of the Watson-Glaser test also inspires confidence because it is *“based on a history of research that is supportive of the test.”*

“It is widely used in industry and just has the best reputation for critical thinking.”



For higher education

Aiding student development

Mike and Jenny agree that in terms of feedback to students, *“it’s unhelpful just to say yes you are good at critical thinking, or no you are not.”* The test’s detail on individual strengths and weaknesses contained within the development report is key, because this is the data students need to improve their skills; it gives them actionable content that allows them to practice and develop critical thinking throughout their education. Mike comments that many students find the test a *“reality check”* as it challenges their beliefs about their critical thinking skills.

Mike is clear that the USF policy of tracking critical thinking development through Watson-Glaser testing helps students gain the confidence, motivation, and proficiency they need for employment. As he says, the questions asked of them when they leave university will not be answered in a textbook: *“Critical thinking gives them the tools and structured thought processes that they need to solve novel problems in the workplace.”* And it’s exactly these skills that USF is nurturing in students and that will give them a competitive advantage in the work environment. *“We want to be known as the university that puts out strong critical thinking students,”* he says.

“We have made it part of the campus culture... this is what we do here.”

Getting the best from Watson-Glaser

- Mike and Jenny gave us their top tips for implementing the test, so that they and their students get the most out of it.
- They deliberately use it in a non-threatening manner: it’s easy to frame Watson-Glaser in a way that makes people open, interested, and curious.
 - They have made it compulsory for all students, but also take the time to explain to them why it is important and why they need to do it.
 - They capitalize on the development report and have resources in place to support student critical thinking development.

In this section

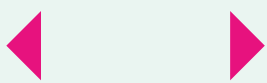
1 For higher education

— University of South Florida (USF)

— Ritsumeikan Asia Pacific University

2 For professional development

3 For recruitment



For higher education

Organization	Length of time using Watson-Glaser	About
Ritsumeikan Asia Pacific University, Japan	1 year	Mr Hirokazu Taoguchi is Assistant Manager/Admissions Office (International), in charge of the screening process for the university; Miss Amelie Chenet Smith is Admissions Counselor and Samuel Beddow is International Admissions Counselor. Here they discuss how adopting Watson-Glaser testing as part of the application process will enable the university to identify the best potential students.

In this section

1 For higher education

- University of South Florida (USF)
- **Ritsumeikan Asia Pacific University**

2 For professional development

3 For recruitment

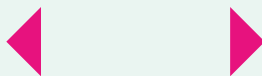
Transforming the application process

Ritsumeikan Asia Pacific University in Japan has a specific problem in that their wide and varied intake of students from around the world makes it hard to compare and evaluate individual applications from very different educational and examination backgrounds. In 2019, the university started to require Watson-Glaser testing as part of the application process for an initial cohort of 500 international students, hoping that the test used alongside more traditional markers of academic performance – such as a GPA score – would help them fairly and accurately identify those students who will excel at the institution.

In 2019, Watson Glaser was used on an initial cohort of

500

International students



For higher education

In this section

1 For higher education

— University of South Florida (USF)

— **Ritsumeikan Asia Pacific University**

2 For professional development

3 For recruitment

Setting the bar high

Samuel comments that in the last year, the amount of interest and speculative enquiry from students for the graduate school has drastically risen. An unanticipated benefit of having the Watson-Glaser test as part of the application procedure is that it automatically selects those candidates with the tenacity and persistence to pursue this more rigorous application process. In this way, the Watson-Glaser test operates not just as a screening tool, but also gives the university insight into how motivated the applicant is.

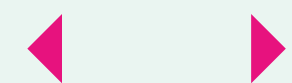
What's next for the Admissions Team and Watson-Glaser?

The university wants students who are multi-disciplinary – who are able to pull back to gain a new perspective and take a critical and holistic view on global issues.

“Critical thinking is critical”

Jokes Samuel, before rephrasing his thoughts. There is, he says, an expectation that in today's workplace candidates will have the intellectual flexibility to shift between roles. People are trained as generalists and need to be able to extrapolate and adapt.

The university's ambition is to gather more data from successive intake to map how Watson-Glaser performance predicts student success on their courses and beyond. The results from the pilot have been encouraging and the university is considering how to expand further its use of the appraisal.



For professional development

Organization	Length of time using Watson-Glaser	About
US Air Force	Over 10 years	BobbieAnn Meyer-Piper is Curriculum Developer for the Chief Leadership course for the US Air Force. Her job involves running training courses for Senior Enlisted Leaders and she explains how Watson-Glaser has helped inform and develop the courses she leads.

In this section

1 For higher education

2 For professional development

— US Air Force

— Steelecase

3 For recruitment

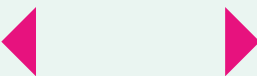
Developing leadership skills

Critical thinking is an extremely important skill for the US Air Force. Bobbie jokes that the people on her courses have been successful all their working lives and can be *“pretty full of themselves.”*

Watson-Glaser becomes an important eye-opener that makes them realize *“they may not be as hot as they think they are!”*

Taking the test at the start of their course helps open participants up to the idea that they need to learn to *“stop and think”* and increases their willingness to be receptive. As Bobbie comments, there’s a perception that military personnel need to be good at following orders, but this is not the case today; the Air Force needs people with mental agility who can think through situations and adapt to rapidly changing circumstances. Bobbie needs the people on her courses to *“think about their thinking”* and Watson-Glaser helps them do just that.

The development report in the Watson-Glaser appraisal allows Bobbie to build lessons that are firmly anchored to the test results and feedback. She explains how she always emphasizes in her courses that emotions and biases are the biggest barriers to critical thinking, and the test results open her students’ eyes to this. It helps them appreciate that *“taking time to think is an effortful thing.”*



For professional development

In this section

1 For higher education

2 For professional development

— **US Air Force**

— Steelecase

3 For recruitment

Improving critical thinking ability

Bobbie says that all their instructors are asked to take the test and, once they have, they want to take it again. These second tests after training show significant improvements in their scores (from around 36 to 59%). Research has found scores to be consistent across different testing occasions, but improvements in results like this may be possible with specific critical thinking training.

“These people are used to making decisions, but they’re also used to following a rule book. They don’t tend to dwell so much on the thinking through but go straight to solutions. Watson-Glaser proves to them that they can become better thinkers.”

Bobbie’s advice to get the best from Watson-Glaser

Bobbie feels that the following tips help her students get the fairest possible results from the testing process.

- Really hammer home the fact that they need to take time and care over the test – really try to hit hard on taking their time with it.
- If you can give them some form of practice questions this can really help. The set-up test can be a really good lesson as it helps them take it seriously.
- Emphasize they need to carefully read the directions: there are different approaches for each section and they need to pay close attention to the instructions.



For professional development

Organization	Length of time using Watson-Glaser	About
Steelecase	6 years; more intensively for the last 2	Oana Stefanescu is a Consultant on HR Operational Excellence & Assessment Research. Steelecase is a global company that has recently decided to roll out Watson-Glaser testing for all senior employees. Oana talks about how the company’s desire for an audited, evidence-based selection procedure has resulted in a “journey of adaptation” across the corporation and how Watson-Glaser data is underpinning this transformation.

In this section

1 For higher education

2 For professional development

— US Air Force

— Steelecase

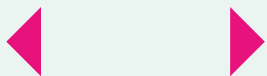
3 For recruitment

Watson-Glaser as part of Steelecase’s recruitment journey

Part of Oana’s job remit is to look at HR innovation and operational excellence within Steelecase. She explains that while the company has been using Watson-Glaser for about six years, the test was used inconsistently across the company within the hiring process. From her point of view, a key part of moving to recruitment excellence was applying the same hiring standards throughout the company – and Watson-Glaser has enabled Steelecase to do this.

Two years ago, the company decided to scale up its use of Watson-Glaser to provide a company-wide metric, increasing the amount of reliable data they were gathering about the quality and success of their new hires. Steelecase chose Watson-Glaser for reasons of trust and validity, in particular its test/re-test reliability. For Oana, the fact that Watson-Glaser is backed by scientific research within the social science domain means it has “unarguable” usefulness as a recruitment tool. Steelecase uses the assessment as part of the selection procedure; the new hire is then evaluated by management throughout their early days on the job.

While data collection is still a work in progress, initial results suggest a positive correlation between performance on the Watson-Glaser assessment and performance on the job – which is really “good news” for Oana, as it is exactly the kind of evidence-based and convincing argument for use of the assessment that will sway initial skeptics.



For professional development

How Watson-Glaser has helped refine and improve processes at Steelecase

Oana explains how the Watson-Glaser assessment introduces fairness into the recruitment procedure by mitigating for any hiring bias. As she says, countless studies show that this is prevalent across managers “who are always biased toward selecting personnel akin to them – though they don’t like to be told that!”

The assessment helps draw attention to areas the interviewer or recruiter may otherwise have overlooked – if a candidate scores in the lower ranges, it “raises a little red flag or at least a question mark” over what this could mean. This then leads on to a more intensive interview process with the aim of surfacing more information about the candidate’s strengths and weaknesses; the Watson-Glaser scores confirm to the company that this is worth spending time on.

Watson-Glaser testing also saves the company time and money; not just at the recruitment stage, but also over the longer term. For Oana, this has possibly been the key argument in persuading Steelecase to roll out Watson-Glaser universally:

“If your hire is successful then you don’t need to spend so much time and resource developing them.”

Oana’s top tip

Oana explains that some candidates can initially be resistant to the idea of the assessment. As she says: “for them it’s just a test – if there are questions to answer: it’s a test!” She has had success with reframing the description of the assessment as a “simulation” and concentrating on explaining that Watson-Glaser looks at broader, real-world scenarios and skills rather than testing technical or job-specific knowledge.

In this section

1 For higher education

2 For professional development

— US Air Force

— **Steelecase**

3 For recruitment



For recruitment

Organization	Length of time using Watson-Glaser	About
Kearney	3 years	Kearney believes strongly in the value of critical thinking as a key skill in today’s competitive marketplace. Recruitment specialists Nabilah Tan and Teresa Yap explain how using Watson-Glaser enables them to identify candidates that will excel in their job roles.

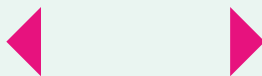
In this section

- 1 For higher education
- 2 For professional development
- 3 **For recruitment**

Critical Thinking in the Modern World

As a key component of problem-solving, critical thinking is a skill that has always been crucial for management consultants, so it is no surprise that Kearney regards it as a foundational competency for its employees. However, in today’s data-rich world the nature of problem-solving is evolving as the variety and depth of information available increases.

Recognition of this change is driving company culture in Kearney, as their employees are also increasingly expected to display a range of multi-disciplinary skills that span both the competencies required in innovative technologies and the people skills required on the job – making selection of the “right” hires both more important and more challenging.



For recruitment

Watson-Glaser as part of the Kearney recruitment journey

With years of experience in Kearney’s recruitment department, Nabilah and Teresa are clear that you cannot gauge someone’s critical thinking ability from their CV – and this is why they need an independent assessment from a test like Watson-Glaser. From their point of view, Watson-Glaser not only helps identify candidates with a strong suite of critical thinking skills but, crucially, does this in a way that introduces fairness to their selection procedure. Because the score takes into account the level of difficulty and the test questions are randomized from item banks, Teresa adds that there is less bias or subjectivity in the test results.

Using the test as a first gate for candidate selection means that Kearney saves considerable time and money within recruitment as it allows them to focus only on those who have already proved their ability from the Watson-Glaser assessment. Once applicants have cleared this critical bar, Kearney is then free to concentrate on case interviews, which highlight a very different skill set in candidates, and allow the interviewer to focus on other – more diverse – attributes.

Taking the Watson-Glaser test can also benefit the applicants themselves, as Nabilah and Teresa both agree that there is often a pool of candidates that look “borderline” from their CVs:

“Taking the WG assessment is an opportunity for them to ‘wow’ us and progress to interview.”

Teresa’s top tip

Teresa recommends trialling the assessment in house; she says that putting your company’s “high-flyers” through it gives a really good gauge of what the range of scores might mean for individual competencies – and allows you identify an appropriate internal benchmark that can then be used going forward in determining the talent pool to interview. In addition this also allows Kearney to monitor trends from selection to development.

In this section

- 1 For higher education
- 2 For professional development
- 3 **For recruitment**



Discussion



Discussion

To evaluate the efficacy of an assessment like the Watson-Glaser Critical Thinking Appraisal, we must consider its reliability, validity, and fairness.

The purpose of Watson-Glaser is to measure an individual's critical thinking ability: their ability to look at a situation and clearly understand it from multiple perspectives while separating facts from opinions and assumptions.

Critical thinking is an essential skill for higher education, employment, and positive participation in society. An effective measure of critical thinking ability is valuable to:

- Higher education institutions, which need to be able to measure critical thinking ability, as a first step toward supporting its development, for their incoming, and sometimes existing, students' critical thinking skills
- HR professionals and employers, who need to be able to assess candidates' critical thinking ability so they can select suitably skilled individuals for open positions and develop current staff

Since work began on the first version of Watson-Glaser in the 1920s, the assessment has been extensively studied. In this report, we have reviewed the body of evidence to build up a picture of Watson-Glaser's efficacy, with a particular focus on whether the assessment can be said to offer validity, reliability and fairness.

There is good evidence that the RED model, the model of critical thinking ability underlying the design of Watson-Glaser, supports the assessment's intended purpose, and that all the items in the test measure the same thing in the same way. The evidence also shows, to a moderate to high degree of confidence, that Watson-Glaser is measuring what it sets out to measure.

As we would expect given these findings, the evidence shows that the test offers good alternate form reliability, and adequate to excellent test-retest reliability. That is, whichever test items are randomly selected from the item bank, the results are largely consistent; and the same test-taker will receive consistent results from taking the test on different occasions, assuming their critical thinking ability does not improve in the interim.



Discussion

Watson-Glaser can be administered in various different modes, to suit administrators' needs and to accommodate test-takers with particular needs. There is evidence that the choice of mode does not help or hinder any test-takers in demonstrating their critical thinking skills.

In-depth investigations of different groups' Watson-Glaser scores indicate that the test is not biased for or against any particular group. While initial comparative studies seemed to show that Black students tended to get lower scores than other ethnic groups, further investigation showed that Watson-Glaser scores were consistent with academic performance regardless of ethnicity, indicating that differences in scores were not a result of bias in the test. There is evidence that the test marginally favors younger candidates, but the effect size is so small that it is unlikely to skew higher education providers' understanding of their courses' effectiveness, or candidates' employment prospects.

Higher education institutions and workplaces alike are making effective use of Watson-Glaser to discover individuals' skills and qualities in ways a traditional résumé cannot reveal, allowing these organisations to recruit and begin developing their people for success.

Colleges have discovered that Watson-Glaser is most effective when the test is mandatory for students, but instructors take time to explain how it is being used and how it will help them. When it is implemented in this way, people respond to Watson-Glaser with openness and curiosity, and it is possible to make the test an integral part of campus culture.

Workplaces have reported that Watson-Glaser is most effective when implemented early in the recruitment process, before interview, as a way of screening candidates and discovering appropriate areas of focus for interviews.

Studies comparing Watson-Glaser scores with measures of on-the-job performance, such as organizational level attained and supervisory ratings of critical thinking behavior, have produced a range of correlations. The evidence from these studies nevertheless suggests that test-takers who perform well on Watson-Glaser are also likely to perform well at work, validating employers' use of the test in recruitment and development.

Similarly, studies comparing Watson-Glaser scores with academic course performance have indicated that the test is a very beneficial predictor of likely course success.



Discussion

In the future, we will continue to reinforce the body of research evidence supporting the validity of the Watson-Glaser, by working with our customers to gather data on key outcomes. The ongoing development of the Watson-Glaser will continue, with item bank refreshes, expansion of the current norm group offering and new language versions planned.

We are exploring new avenues for both test delivery mechanisms and new uses for the product. There is increasing demand to have assessments optimized for mobile delivery, which would make tests more accessible to those without access to a computer, or simply with a preference for using a mobile over other devices. We are aiming for Watson-Glaser to be mobile optimized in the near future to meet this need and will be conducting the required investigation into the product design and the impact on validity and usability of this delivery method.

Critical thinking is often seen as a crucial skill for success in many roles and as a result, it is no surprise that this skill features in many employability frameworks which guide students and educators in the skills needed for future employment. We are currently working on projects to implement the Watson-Glaser in an educational context, as part of broader Pearson projects to assess and develop the employability skills of students around the world, such as linking our Talent assessments to the Pearson employability framework.





References



ACT. (no date). Collegiate assessment of academic proficiency technical manual. Iowa City, IA: ACT. Retrieved from <https://www.act.org/content/dam/act/unsecured/documents/CAAP-TechnicalHandbook.pdf>

American Management Association staff. (2019). American Management Association critical skills survey: workers need higher level skills to succeed in the 21st century. January 24, 2019. <https://www.amanet.org/articles/ama-critical-skills-survey-workers-need-higher-level-skills-to-succeed-in-the-21st-century/>

Association of American Colleges and Universities. (2011). The LEAP vision for learning: *Outcomes, practices, impact, and employers' view*. Washington, DC: Association of American Colleges and Universities.

Butler, H. A., Dwyer, C. P., Hogan, M. J., Franco, A., Rivas, S. F., Saiz, C., & Almeida, L. S. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity*, 7(2), 112–121.

Davies, W., & Stevens, M. (2019). *The importance of critical thinking and how to measure it*. Pearson TalentLens.

Educational Testing Service. (2013). *Quantitative market research* [PowerPoint slides]. Princeton, NJ: Educational Testing Service.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership*, 43(2), 44–48.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Millbrae, CA: The California Academic Press.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4), 449–455.

Hinton, D. (2015). Uncovering the root cause of ethnic difference in ability testing: differential test functioning, test familiarity and trait optimism as explanations of ethnic group differences (Doctoral dissertation, Aston University).

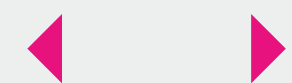
Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: issues, evidence, and lessons learned. *International Journal of Selection and Assessment*, 9(1/2), 152–194.

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.

Korn, Melissa. (2014). Bosses seek 'critical thinking,' but what is that? *The Wall Street Journal*. October 21, 2014. <https://www.wsj.com/articles/bosses-seek-critical-thinking-but-what-is-that-1413923730>

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51(2), 77–101.

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



OECD. (2012). *Education at a Glance 2012: OECD Indicators*. Paris: OECD Publishing.

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

Rayner, G., & Papakonstantinou, T. (2015). Employer perspectives of the current and future value of STEM graduate skills and attributes: An Australian study. *Journal of Teaching and Learning for Graduate Employability*, 6(1), 100-115.

Steele, C.M., & Aronson, J. (1995). Stereotype Threat and the intellectual test-performance of African-Americans. *Journal of personality and Social Psychology*, 69 (5): 797-811.

Rowe, M. P., Gillespie, B. M., Harris, K. R., Koether, S. D., Shannon, L.-J. Y., & Rose, L. A. (2015). Redesigning a general education science course to promote critical thinking. *CBE-Life Sciences Education*, 14(3), 1–12.

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

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

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

Stewart, William. (2014). School leavers lack the critical thinking skills needed for university, exam board warns. TES. January 25, 2014. <https://www.tes.com/news/school-leavers-lack-critical-thinking-skills-needed-university-exam-board-warns>

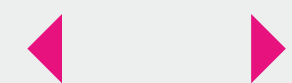
US Department of Labor, (1999). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author.

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

Watson, G., & Glaser, E. M. (2019). *Watson-Glaser III Critical Thinking Appraisal: User's guide and technical manual*.

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.



Pearson
80 Strand
London
WC2R 0RL

pearson.com
@Pearson

Download this report ▶