

podium

Interpreting GCAT Confidence Scores in the Age of AI

Whitepaper
2025

Citation ISO Certification
ISO 27001: 2022
Information security management
REGISTERED

INTRODUCTION: CONFIDENCE SCORES

The GCAT (General Cognitive Ability Test) is a robust psychometric assessment, registered with the British Psychological Society (BPS) and rated 4 out of 4 stars for technical quality. It is designed to measure General Mental Ability (GMA), including Verbal, Numerical, and Abstract reasoning skills.

With the evolution of digital testing and the increasing presence and access to AI tools, the landscape of online assessments has changed.

> *Can I trust this result?*

> *Did the candidate complete the test under fair conditions?*

> *Should I follow up with the candidate or consider a supervised retest?*

These are natural and necessary reflections in a high-stakes testing environment. The GCAT Confidence Score is designed to support these questions with evidence, helping administrators reach informed and fair conclusions.

The Rise of Unsupervised Assessment and Emerging Challenges

In recent years, there has been a substantial shift towards unsupervised online assessments. This shift, driven by the need for scale, convenience, and accessibility, has brought with it a new layer of complexity in maintaining test security and result validity. While unsupervised testing has clear advantages, it also raises questions about the authenticity of candidate performance.

It is not uncommon for organisations to request a follow-up, face-to-face assessment when stakes are high, particularly if results appear inconsistent or unexpected. This is considered best practice for validating online test results.

However, it is not always feasible due to constraints such as time, budget, and logistical complexity. The GCAT Confidence Score helps to bridge the gap between efficiency and rigour, allowing test administrators to better assess whether an online result is likely to reflect a candidate's true ability, even when a follow-up session is not possible.

© Podium Systems Limited

AI and High-Stakes Testing: A Growing Concern

With the rapid advancement of generative AI technologies, there is an increasing risk of candidates using AI-based tools (e.g., ChatGPT, Google Gemini) during assessments to gain an unfair advantage. This concern is particularly acute in high-stakes recruitment scenarios, where a strong score can secure a position, and a weak one can end a candidate's prospects.

Candidates are under more pressure than ever to perform well, and this pressure can lead to behaviour that compromises the integrity of assessments - whether through deliberate misuse of technology or unintended breaches such as opening another tab to search for an answer. Understanding and identifying these behaviours is essential to protect both the organisation and the candidate from inaccurate outcomes.

However, it is not always feasible due to constraints such as time, budget, and logistical complexity. The GCAT Confidence Score helps to bridge the gap between efficiency and rigour, allowing test administrators to better assess whether an online result is likely to reflect a candidate's true ability, even when a follow-up session is not possible.

This document outlines how to understand and use the GCAT Confidence Score, providing practical guidance for identifying potential anomalies without jumping to premature conclusions.

> ***Using AI can be a helpful way to help people prepare for interviews and build their confidence. However, using a real-time AI assistant to respond in a live job interview simply isn't ethical. It's like cheating during an exam.***[1]
>
> — Hayfa Mohdzaini, Technology Research Adviser at CIPD (The Times)

[1] The Times (2024). "Cheating AI tool prompts candidates during job interviews." https://www.thetimes.co.uk/article/cheating-ai-tool-prompts-candidates-during-job-interviews-m0n76zwvl

© Podium Systems Limited

1. WHAT IS A CONFIDENCE SCORE?

The Confidence Score is a metric ranging from 0 to 100 that reflects the system's level of confidence that the test was completed under standard, fair conditions.

This score is generated using:

- **Candidate behaviour**: when a candidate clicks away from the test window.
- **Timing data**: how much time a candidate spends on each question compared to the normed average.
- **Clipboard monitoring**: attempts to copy/paste during the test are flagged.
- **Test recovery**: the number of times a candidate re-starts their test before completing it.

Additional metrics may also be considered, such as **Podium's Facial Validation** technology, which can help verify candidate identity and detect environmental anomalies during the assessment process.

These technologies, along with the behavioural indicators above, contribute to the overall confidence score. Together, they provide a comprehensive view of the test-taking experience, offering greater assurance - or greater caution - around the validity of the candidate's result.

These inputs are synthesised by an algorithm that weighs patterns of behaviour.

But Remember:
A low score **DOES NOT** *confirm* misconduct, but it does raise questions about the reliability of the result.

2. SCORE INTERPRETATION FRAMEWORK

Confidence Scores are grouped into five qualitative bands:

- **Very High (90-100%)**: Candidate demonstrated typical desired behaviour. No anomalies detected.
- **High (70-89%)**: Some deviation, but within expected bounds. Likely a reliable result.
- **Medium (55-69%)**: Moderate anomalies present. Result may be impacted.
- **Low (40-54%)**: Significant irregularities. Confidence in the result is reduced.
- **Very Low (0-39%)**: Multiple concerning behaviours. Caution advised.

Alongside the band and percentage score, administrators will be shown key flags such as:

- Number of Click Away events.
- Whether questions showed significant time deviations when the candidate clicked away.

© Podium Systems Limited

3. INTERPRETING THE DATA

When reviewing a Confidence Score, administrators should:

- **Look at the context**: Evaluate the candidate's role, background, and any other relevant assessment data. For example, a software developer applying for an analytical role might reasonably spend more or less time on numerical questions based on their experience.
- **Consider the frequency of Click Away events**: A small number of events may be benign - for example, triggered by innocent pop-up notifications or accidental clicks. However, a higher number of Click Away events could be indicative of distractions or attempts to access external content.
- **Assess timing deviations**: The system will indicate whether there were significant deviations from the norm on questions where the candidate clicked away. While individual deviations can occur naturally, repeated or extreme differences may suggest irregular behaviour.
- **Use supporting data**: Consider whether the performance aligns with other recruitment indicators such as interviews, CV claims, or references. For example, if a candidate's test score appears unusually high compared to their academic or professional background, a closer review may be warranted.
- **Combine indicators**: While individual Click Away events or time deviations may not be significant alone, their combination with a low confidence score and other factors (such as facial validation alerts) should encourage a more cautious interpretation.

Example scenarios:

a) A candidate clicks away three times, and these questions also show significant time deviations. This combination may suggest external assistance.

b) A candidate with a low confidence score reports experiencing Wi-Fi dropouts. Follow-up discussion or supervised retesting may clarify the issue.

c) A candidate has a high score but flags for both Click Away events and timing deviation. If the score seems inconsistent with their application profile, a retest or further discussion may be advisable. Caution should be taken.

© Podium Systems Limited

4. HOW TO USE THIS INFORMATION

This Confidence Score should not be used in isolation to make pass/fail decisions. Instead, it is designed to:

- **Flag results for review**: Invite further scrutiny or discussion.
- **Inform re-testing decisions**: Where a candidate reports technical issues or suspicion is high, a supervised re-test might be appropriate.
- **Guide post-assessment discussions**: If there are concerns, administrators may choose to follow up with the candidate.

Example follow up questions:

> *Did you encounter any technical issues during the test? Could you describe them and how you were affected by them?*

> *Can you explain why you left the assessment (x) times during the assessmnet?*

> *Are you aware that switching away from the assessment window is tracked as part of our test security measures? Was there a particular reason you might have needed to leave the test screen during the assessment?*

> *Can you confirm that you completed the assessment independently, without assistance from other individuals or tools such as AI platforms during the test?*

© Podium Systems Limited

5. IMPORTANT CONSIDERATIONS

- **This is not a cheating detection tool**: While suggestive of cheating, a low score is not proof of misconduct, but rather an alert to potential irregularities.

- **Technical issues can impact scores**: A poor internet connection, power cuts, or accidental window switches are all plausible explanations.

- **Performance may still be valid**: Some candidates may perform well despite anomalies/technical issues. Others may underperform due to legitimate disruptions that are not possible to detect.

- **Understand timing variance**: Being significantly slower or faster than average on a question is not, in itself, an indicator of dishonest behaviour.

  A slow response may indicate the candidate was thoughtfully considering their answer or struggling with the content.

  Conversely, a fast response may result from guessing, a high level of confidence, or even accidentally skipping over a question.

  These behaviours are not inherently suspicious but are taken into account when combined with other indicators.

6. ETHICAL USE AND COMMUNICATION

Administrators should handle confidence scores with care, fairness, and transparency. While browser technology can confirm that a candidate has left the test screen, it cannot determine where they went or what they did. A Click Away event could therefore be entirely innocent - such as responding to a notification or accidental clicking - though it remains unadvisable during a timed test. Candidates should not be accused of dishonesty without clear evidence, and any concerns should be addressed respectfully and sensitively.

If a candidate disputes a result, administrators are encouraged to:

- Offer an explanation of the confidence score system.
- Consider allowing a retest under monitored conditions.
- Document the process and rationale for any follow-up actions.

Full details of any Click Away events and time deviations that contributed to the confidence calculation are securely stored and can be made available to administrators upon request.

This allows for deeper insight into the specific behavioural indicators behind a confidence score, supporting informed follow-up where necessary.

© Podium Systems Limited

CONCLUSION

In an increasingly digital and unsupervised testing environment, the Confidence Score provides essential support for test administrators in interpreting candidate performance.

Built on robust data collection and analysis - including blur events, timing deviations, clipboard monitoring, and optional facial validation technologies - the Confidence Score offers a nuanced view of test integrity.

While the score does not serve as a cheating detection mechanism, it is a valuable tool for identifying potential anomalies and guiding thoughtful, evidence-based decisions. It helps flag results that may warrant additional review or candidate follow-up, particularly in high-stakes scenarios where fairness and rigour are paramount.

By incorporating this score into the broader assessment process, administrators can:

- Make informed decisions even when follow-up testing is not feasible.
- Address integrity concerns while maintaining fairness and objectivity.
- Support candidates by acknowledging the impact of possible disruptions.

Ultimately, the GCAT Confidence Score empowers organisations to uphold the quality and validity of their assessments - balancing convenience with rigour, and technology with sound professional judgement.

For any further guidance or support, please contact our Psychometrics and Assessment Integrity team.

© Podium Systems Limited

podium

AUSTRALIA | BAHRAIN | CHINA | CROATIA | CZECH REPUBLIC | EGYPT
FINLAND | GERMANY | INDIA | ITALY | NEW ZEALAND | OMAN
REPUBLIC OF IRELAND | SINGAPORE | SOUTH AFRICA | SWEDEN
SWITZERLAND | THE NETHERLANDS | TURKEY | UNITED ARAB EMIRATES
UNITED KINGDOM

www.podium365.com

© Podium Systems Limited