Test Partnership

White paper · September 2026

AI-resilience in assessments

A practical framework for minimising the impact of AI cheating in assessment and selection.

Ben Schwencke
Chief Psychologist · MSc Occupational Psychology

Contents

ABOUT THE AUTHOR

Ben Schwencke

Chief Psychologist, Test Partnership

MSc in Occupational Psychology. Ben leads the psychometric research and product science behind Test Partnership's AI-resistant MindmetriQ™ assessment platform.

TEST PARTNERSHIP IN NUMBERS

96%
candidate satisfaction

6,000+
organisations rely on our assessments

3m+
assessments completed

OVERVIEW

Executive summary

Generative AI has changed the conditions under which remote selection takes place. Candidates can now obtain rapid assistance with written applications and many conventional test items.

Research reviews show that current large language models can perform strongly on large sets of multiple-choice examinations (Newton & Xiromeriti, 2024). Field evidence from one online freelance platform indicates that access to an AI writing tool improved and standardised cover letters in ways that reduced their value as signals of applicant ability (Cui et al., 2025). That study concerns a single platform and intervention, but the mechanism it describes applies wherever written applications are used. This does not make structured assessment obsolete. It makes the need for **resilient assessment** greater, particularly because AI lowers the effort required to generate and tailor application materials.

AI resilience is the extent to which a selection process preserves valid information about a candidate by discouraging AI-assisted cheating and limiting the advantage it can provide. It is not a binary property. Methods sit on a continuum from relatively safer to highly vulnerable, and the position of any method depends on how it is designed, administered and verified. Section 1 applies the same idea to an individual assessment; the two levels are related, because a resilient process is built from resilient assessments plus stages that do not share their vulnerabilities.

The central operational risk is shortlist contamination. A small proportion of applicants may cheat, but successful cheaters can become heavily overrepresented among the highest scorers and therefore among those who progress. Research on unproctored employment testing also shows that suspicious score discrepancies require verification and response-pattern evidence rather than simple comparisons of group means (Aguado et al., 2018; Guo & Drasgow, 2010).

THE CENTRAL PRINCIPLE

AI-assisted cheating should be made difficult, risky and unreliable enough that independent completion remains the candidate's most attractive strategy.

5% → 50%

if 5% of applicants cheat successfully and the top 10% are shortlisted, AI-assisted candidates could fill up to half the shortlist (illustrative upper bound)

9 principles

the ACT framework — this paper's proposed organising model: Assessment design, Candidate messaging, Technical solutions

4 bands

selection methods grouped by AI vulnerability, from lower to highest — placement depends on implementation

EXECUTIVE SUMMARY

Key conclusions

1

AI resilience is a system property.

The ACT framework — this paper's proposed organising model — integrates Assessment design, Candidate messaging and Technical solutions.

2

Shortlists are the correct unit of concern.

Low applicant-level cheating can still produce severe contamination among successful candidates.

3

Assessment design should come first.

Surveillance can support resilience but should not compensate for a static, easily outsourced test.

4

Live monitoring is essential.

Upper-tail inflation, verification discrepancies and process data should be reviewed continuously.

5

Familiarity is not evidence of resilience.

Provider size, age and brand establish neither resilience nor vulnerability. Judge providers by what has changed and what evidence shows it works.

6

Assessments are needed more than ever.

The appropriate response to AI is better assessment, not a retreat to highly vulnerable application materials.

SECTION 1

Defining AI resilience

01

CORE DEFINITION

AI resilience is the extent to which an assessment preserves its validity by discouraging the use of AI assistance and limiting the advantage it can provide.

The executive summary applies the same property to the whole selection process. The two are related: a resilient process combines resilient assessments with verification and later stages that do not depend on the same vulnerable evidence, so that assistance at one stage cannot carry through to the hiring decision.

An AI-resilient assessment does not depend on making cheating impossible. Instead, it reduces both the likelihood that candidates will attempt to use AI and the likelihood that doing so will produce a clear, dependable or lasting advantage.

AI resilience therefore concerns the candidate's expected return from cheating. A resilient process increases the effort, time, uncertainty and risk involved, while reducing the probability that AI can solve the task accurately, that the answer can be mapped back to the interface, and that any inflated performance will survive later checks.

What resilience is intended to achieve

- Deter candidates by making rules, monitoring and consequences explicit.
- Reduce the attractiveness of experimentation with AI.
- Limit the opportunity to capture, transfer and reconstruct assessment content.
- Reduce the accuracy, consistency and reproducibility of AI-assisted performance.
- Increase the likelihood that inflated results are exposed through monitoring or verification.
- Ensure that external assistance cannot be relied upon throughout the full selection process.

Resilience, not elimination

Completely eliminating the risk of AI-assisted cheating is neither realistic nor necessarily desirable. Excessively optimising for security could damage validity, reliability, fairness, accessibility and candidate experience to such an extent that the intervention causes more harm than the risk it is intended to address (American Educational Research Association et al., 2014; Coghlan et al., 2021; Hausknecht et al., 2004).

AI resilience should therefore be treated as one design objective among several. Controls should be proportionate to the stakes, expected risk and practical consequences of misuse, and should follow established principles for secure, fair and appropriate test use (International Test Commission, 2016).

> “
>
> The objective is not zero risk. It is a selection process that remains useful, fair and economically valuable in the presence of AI.

Evidence base and limitations

Direct peer-reviewed research on generative AI in employment assessment remains limited. This paper therefore combines evidence from personnel selection, unproctored Internet testing and verification, psychometric test security, and generative AI research in education and labour-market signalling. Proposed measures are identified as design propositions rather than established tests, and the ACT framework and vulnerability bands that follow are the author's proposals informed by that literature.

- Personnel selection & utility research
- Unproctored Internet testing & verification
- Psychometric test security guidance
- Generative AI research in education & labour markets

SECTION 2

Why AI-assisted cheating matters 02

Generative AI is, at its core, a powerful **cognitive offloading tool**: it allows users to transfer part of the information-processing demands of a task to an external system, reducing the cognitive effort required for completion (Gilbert et al., 2023; Risko & Gilbert, 2016). Research also shows that technical tools can improve immediate task performance by externalising cognitive processes, even where this reduces internal processing or subsequent memory (Grinschgl et al., 2021). This creates a fundamental conflict with assessment, which is intended to evaluate the candidate's own cognitive abilities, knowledge or skills. Where AI performs part of the task, the resulting score no longer supports its intended interpretation as evidence of the candidate's own capability. Outcome evidence on the consequences for predictive validity is still emerging, but the threat to what the score means is immediate.

This paper concerns unauthorised assistance on assessments designed to measure unaided capability. Authorised AI use in a task built to assess AI-enabled performance is a different matter: there, AI use is part of what is measured.

Unauthorised use cannot be defended on the basis that candidates will use AI in their future work. Cheating on many assessments requires little or no meaningful AI expertise, often involving no more than submitting a single screenshot and accepting the response produced. The resulting score therefore does not demonstrate effective AI use, judgement or prompt engineering. It simply reflects the ability to transfer assessment content to an external system, making the argument that AI-assisted performance represents a relevant workplace skill difficult to sustain.

Moreover, AI-assisted cheating does not need to be widespread across the applicant pool to cause serious harm. Candidates who gain an artificial score advantage are more likely to reach the upper end of the distribution and may therefore become overrepresented among those who pass, are shortlisted or are hired. This concentration problem is consistent with the wider literature on unproctored employment testing and verification (Aguado et al., 2018; Arthur et al., 2009; Tippins et al., 2006).

The practical question is not only "How many applicants are cheating?" It is "How many shortlisted candidates owe their position to AI-assisted score inflation?"

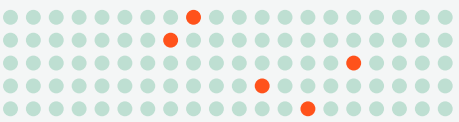
Shortlist amplification

Selection processes concentrate attention on a small proportion of the applicant pool. This means that a low overall cheating rate can translate into substantial shortlist contamination.

ILLUSTRATIVE SCENARIO · UPPER-BOUND, NOT AN ESTIMATE OF PREVALENCE

5%

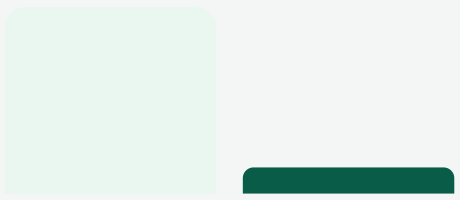
of applicants use AI successfully



The applicant pool — 100 candidates, 5 AI-assisted

10%

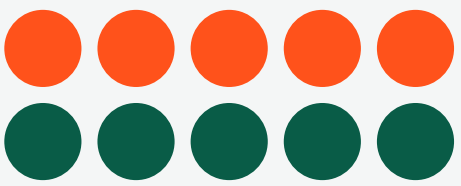
highest scorers are shortlisted — successful cheaters cluster at the top



100 applicants 10 shortlisted

50%

of the shortlist could be AI-assisted candidates



The shortlist — 10 candidates, up to 5 AI-assisted

AI-assisted candidate · Independent candidate · Not shortlisted

The example shows why applicant-level prevalence can be misleading: the effect of cheating is **amplified** when decisions focus on the upper end of the distribution. It also shows why three groups must be kept distinct — five AI-assisted candidates in a shortlist does not mean five places were taken from others.

GROUP 1
Candidates who used unauthorised AI

GROUP 2
Candidates whose scores rose because of it

GROUP 3 · THE MEASURE OF DISPLACEMENT
Candidates who progressed specifically because of that rise

The three drivers of vulnerability

01

Prevalence of cheating

The proportion of applicants who attempt to use AI or another form of unauthorised assistance.

02

Effectiveness of cheating

The size and reliability of the score advantage produced for those candidates.

03

Selectivity of the process

The proportion of applicants who progress, with more selective processes creating greater potential for concentration among successful candidates.

A low cheating rate can still cause substantial damage when AI produces a large score advantage and few applicants progress; conversely, attempted cheating has little effect where the design makes assistance slow, unreliable or unlikely to improve performance.

Shortlist contamination reduces validity

An assessment creates value by ranking candidates according to capabilities that are relevant to future job performance. AI-assisted cheating damages that mechanism because the observed score may no longer reflect the capability the assessment was intended to measure. This threatens the validity of the score interpretation and the utility of the resulting selection decision (American Educational Research Association et al., 2014; Schmidt & Hunter, 1998).

Where an inflated score is unrelated, or only weakly related, to the candidate's true capability, that candidate is placed incorrectly within the ranking and **may displace a stronger independent candidate**.

How much value is lost depends on how far inflated scores diverge from true capability, whether the displaced candidates were marginal or exceptional, and whether later stages correct the error. Displacing one exceptional candidate can cost more than several marginal misplacements, so the loss is not simply proportional to the number of contaminated decisions.

Financial impact

AI-assisted cheating reduces the financial value created by selection because it weakens the relationship between assessment scores and genuine capability. This can lead to stronger candidates being rejected, weaker candidates progressing and poorer hiring decisions. The size of the financial impact depends on:

- the proportion of shortlisted candidates whose scores were artificially inflated;
- the number and expected tenure of hires;
- the costs of poor hiring decisions, including lower productivity, attrition and replacement.
- the extent to which cheating changes progression and hiring decisions;
- the financial value of differences in employee performance; and

Selection-utility models show that assessment value depends partly on validity and on improving the quality of the selected group. As cheating reduces that validity, it reduces expected economic return (Boudreau, 1983; Cronbach & Gleser, 1965; Schmidt et al., 1979).

This paper deliberately does not state a numerical relationship between shortlist contamination and lost value. Cheating candidates may retain substantial genuine capability; some would have progressed regardless; later stages may correct some decisions; and assessment utility and return on investment are different quantities. Organisations that want a figure should build a worked model with explicit assumptions about validity, selection ratio, performance variability and tenure, using established utility formulae. Under most reasonable assumptions, contamination concentrated in the shortlist is **considerably more costly than its applicant-level prevalence would suggest**.

WHAT THE FINANCIAL ARGUMENT CLAIMS

Shortlist contamination reduces the value of selection. How much depends on assumptions that should be stated explicitly — not read from a single ratio.

A selection-system problem

AI cheating should not be treated solely as a test-security issue. Its consequences are shaped by the architecture of the full selection process: where the assessment sits, how heavily its score is weighted, how selective the organisation is, and whether later stages independently reassess the same capabilities.

A vulnerable assessment can distort more than one score. It changes who reaches interview, who consumes recruiter and hiring-manager time, and which legitimate candidates are excluded before stronger evidence can be collected. Where later stages rely on equally AI-vulnerable materials or do not verify the capability assessed, the initial distortion can persist through to the final hiring decision.

Risk is therefore highest when few applicants progress, the assessment carries substantial decision weight, AI provides a large and dependable advantage, and subsequent stages offer little independent verification. In these circumstances, even a small amount of successful cheating can have a disproportionate effect on who advances and ultimately on the quality of the hiring decision.

THE SCALE OF THE SHIFT

Generative AI represents the gravest threat to test integrity since the move from paper-and-pencil to online assessment.

Lessons from the move to online assessment

The transition from paper-and-pencil testing to remote online assessment created the previous major break in test-security assumptions. Tests that had been administered under controlled conditions could now be copied, shared and distributed at scale. This increased the need for large item banks, randomised or adaptive delivery, stronger content controls and verification procedures (International Test Commission, 2006; Tippins et al., 2006; Tippins, 2015).

Generative AI does not merely make assessment content easier to obtain or share; it can interpret that content and produce answers in real time. The threat therefore reaches beyond item exposure and directly challenges whether an observed score reflects the candidate's own capability.

The response must be comparable in ambition to the industry's earlier move towards item-banked online assessment. Just as the move from paper-and-pencil testing to online assessment required a fundamental change in test security, generative AI now requires assessment providers to reconsider how tests are designed and delivered. Item banking remains essential, but it is no longer sufficient on its own. The conditions under which candidates complete assessments have changed so materially that **previous assumptions about test security can no longer be relied upon**.

SECTION 3

The AI-vulnerability continuum

03

No selection tool is completely AI-proof. Even highly controlled methods can be compromised in principle through concealed devices, collusion or other external assistance. AI resilience is therefore a continuum rather than a binary state. The bands below are an **illustrative judgement** of vulnerability to unauthorised AI assistance — not a measured ranking, and not an assessment of predictive validity. They deliberately mix assessment methods, administration conditions and verification procedures, because employers combine all three; where any method sits depends on how it is implemented.

LOWER VULNERABILITY — HIGHEST VULNERABILITY

Lower Live, controlled conditions	**In-person assessment centre** Multiple live exercises and observers make sustained AI assistance difficult, though not impossible with concealed devices.	**In-person structured interview** Live probing limits the usefulness of prepared or AI-generated responses.
	Proctored online assessment Identity, screen and environment checks reduce opportunities for external assistance.	**Supervised verification test** A controlled follow-up that checks whether earlier performance can be reproduced; a procedure applied to other methods rather than a method in itself.
Moderate Remote, but live or design-constrained	**Live video interview** Real-time questioning offers protection, although live AI interview support remains possible.	**Unproctored assessment with resilient design** Dynamic formats, tight timings and fragmented information can slow and degrade AI use. Placement depends on evidence for the specific design (page 25), not the label.
	Personality questionnaire AI cannot simply solve it, but may help candidates imitate a desirable profile.	
Higher Remote, static, unsupervised	**Traditional unproctored assessment** Static questions and generous timings make content easy to capture and submit to AI.	**Asynchronous video interview** Scripts, teleprompters and generated answers can be used without live probing.
	Verified work history or authenticated portfolio Embellishment is possible, but claims can be checked against employers, referees or independently verified work.	
Highest Unsupervised written prose	**CV as written** AI can rewrite and tailor the document, making candidate contribution hard to verify without checking the underlying history.	**Application-form questions** AI can produce polished, role-specific answers with minimal effort.
	Cover letter or personal statement AI can generate the entire submission almost instantly, leaving little reliable candidate signal.	

Placement is conditional on implementation. A loosely proctored test may be no safer than an unproctored one, and a live interview without probing may be little safer than an asynchronous one.

The weakest signals may be the most familiar

Although assessments are often the main focus of discussions about cheating, CVs, cover letters and application forms are generally more vulnerable because large language models can generate, rewrite and tailor them with little effort. Field evidence from one online freelance platform indicates that AI-assisted cover-letter writing reduced the relationship between the written signal and applicant ability (Cui et al., 2025); the setting is specific, but the mechanism generalises. CVs and portfolios should be distinguished: a verified work history or an independently authenticated portfolio remains a stronger source of evidence than polished application prose.

Employers should therefore avoid knee-jerk reactions that abandon assessment and return to CV-led screening. A measured response strengthens AI resilience while preserving the wider quality of the selection process. Structured selection methods remain valuable when they preserve valid information about job-relevant capability (Schmidt & Hunter, 1998).

The knee-jerk reaction

Abandon assessment. Return to CV sifting, cover letters and written application questions — the three methods in the highest-vulnerability band.

The measured response

Strengthen AI resilience while preserving the wider quality of the selection process — structured methods that keep valid information about job-relevant capability.

3 The **highest-vulnerability band** — CV prose, application-form questions and cover letters — consists entirely of traditional application materials.

REMEMBER

The bands concern AI vulnerability — not the predictive validity of each method — and placement depends on implementation.

SECTION 4

Why parts of the market may downplay the problem

04

Parts of the assessment market may understate the threat posed by AI-assisted cheating. Public responses may emphasise **reassurance** rather than the implications for validity, shortlist quality and return on investment.

This does not mean every provider is deliberately concealing evidence. However, narrow analyses, sunk costs and commercial incentives can all encourage conclusions that make the threat appear smaller than it is. What follows is organised around claims buyers should scrutinise, not around any provider's motives.

Using the wrong level of analysis

A common response is to compare average scores before and after the widespread adoption of generative AI. If the mean has changed only slightly, the provider may conclude that AI has had little effect.

This is a weak test of the problem. AI-assisted cheating is likely to be concentrated among a smaller group of motivated or technically capable candidates. Their score gains can be diluted within the full sample while materially changing the upper tail and shortlist. Person-fit, response-time and verification research has long treated aberrant performance as an individual-pattern problem rather than one that can be inferred from a stable group mean (Aguado et al., 2018; Meijer & Sijtsma, 2001; van der Linden & Guo, 2008).

Providers also tend to focus on the proportion of the full applicant pool that may be cheating. This is **the wrong denominator for selection harm**. The relevant question is how many successful candidates achieved progression through inflated performance, which is likely to be more substantial.

Claiming that cheating is nothing new

Candidates have always been able to seek outside help, but generative AI has changed the speed, accessibility, affordability, scale and quality of that assistance. Support that once required planning, specialist knowledge or collaboration can now be attempted opportunistically with a second device. Studies of student use and assessment performance support the conclusion that generative AI represents a material change in the assessment environment (Gruenhagen et al., 2024; Newton & Xiromeriti, 2024).

The existence of earlier forms of misconduct is not evidence that the current threat is unchanged. Treating AI as merely an old problem in a new form avoids the possibility that established assessment formats are no longer fit for purpose.

WHAT HAS ACTUALLY CHANGED

- Speed
- Accessibility
- Affordability
- Scale
- Quality

THE WRONG DENOMINATOR

The relevant question is not how many applicants cheat. It is how many successful candidates progressed through inflated performance.

Protecting legacy products

Much of the assessment market is built around static formats developed before generative AI became widely available. These often include long passages, fully visible information, generous timings and predictable item structures. Such tests may remain valid under honest completion, but their operational security assumptions require renewed scrutiny in light of the demonstrated performance of large language models and the established risks of unproctored Internet testing (Newton & Xiromeriti, 2024; Tippins et al., 2006).

Redesign may require new interfaces, new item banks, platform development, trial data, calibration and renewed validation. Providers with large legacy investments may therefore face practical pressure to defend existing formats. Buyers should **ask what has changed**, rather than infer either resilience or vulnerability from a provider's size or history.

Capability and agility constraints

Meaningful AI resilience requires psychometric expertise, software development, item-bank creation, behavioural insight, security thinking, data analysis and ongoing research. Not every provider has these capabilities internally, and established guidance makes clear that secure online testing is a system-level responsibility rather than a single platform feature (International Test Commission, 2006, 2016; Tippins, 2015).

The industry should be judged by whether it changes the measurement process itself, not merely whether it adds features around the edges. Provider size, age and familiarity establish neither resilience nor vulnerability; the relevant evidence is what the provider has changed and what it can demonstrate.

Defaulting to surveillance

Where providers acknowledge the issue, the first response is often webcam proctoring, browser lockdown, screen recording or automated flagging. These controls can deter misuse and improve assurance, but research also documents privacy, fairness and implementation concerns, and shows that proctoring changes behaviour without eliminating all forms of cheating (Coghlan et al., 2021; Dendir & Maxwell, 2020; Hylton et al., 2016).

More importantly, surveillance does not fix weak assessment design. A static, text-heavy and generously timed assessment remains vulnerable even when browser activity is recorded. Technical controls should support resilient design, not substitute for it.

Claims buyers should scrutinise

Recognising the full problem may require redesign of flagship products, new item banks, research and candour with existing clients; reassurance is cheaper. Three claims therefore deserve particular scrutiny: that stable averages show AI has had little effect; that cheating remains uncommon in the applicant pool; and that monitoring added around an unchanged test is sufficient.

None of this establishes intent. The explanation offered here concerns plausible incentives and analytical choices, not evidence about any named provider. The appropriate response is to ask for evidence, not to assume the worst.

Challenge reassuring claims that rely on stable averages, blunt evidence or unchanged designs.

SECTION 5

The ACT framework for AI resilience 05

AI resilience should not depend on a single intervention. No warning message, assessment format or technical control is sufficient on its own. ACT is this paper's proposed organising framework, informed by existing research on test security and deterrence. It combines three complementary layers: **Assessment design**, **Candidate messaging** and **Technical solutions**. Each layer contains three principles, creating a nine-point approach in line with the broader principle that test security depends on coordinated controls across the assessment system (International Test Commission, 2016; Tippins, 2015).

Assessment design is the primary intervention because the task itself should limit the usefulness of AI. Candidate messaging adds deterrence at low cost. Technical solutions are applied in proportion to risk rather than as a fixed final step: session telemetry, a brief supervised verification test and continuous webcam surveillance impose very different burdens, and a light verification step may sometimes preserve assessment quality better than making the original test ever faster or more fragmented.

A Assessment design	C Candidate messaging	T Technical solutions
1 · Dynamic format Use formats that are harder to capture, transfer and outsource to AI.	**4 · Clear boundaries** State exactly what forms of assistance are prohibited.	**7 · Session monitoring** Use browser, timing and response data to identify patterns requiring review.
2 · Time restriction Make external assistance carry a meaningful time cost.	**5 · Credible monitoring** Create genuine uncertainty about whether misuse will be identified.	**8 · Webcam proctoring** Increase identity and environment assurance where the stakes justify it.
3 · Information fragmentation Prevent one capture from containing everything needed to answer.	**6 · Meaningful consequences** Make the downside of misconduct explicit and credible.	**9 · Verification testing** Check whether high performance can be reproduced under controlled conditions.

HOW ACT WORKS

A reduces the probability that AI assistance will work.

C reduces willingness to attempt cheating.

T increases the probability that misuse will be identified or exposed.

A

ACT · FIRST LAYER

Assessment design

Assessment design should reduce the opportunity, practicality and expected benefit of using AI. It should make external assistance slower, more difficult and less reliable before surveillance or enforcement becomes necessary. This reflects established security thinking that assessment design itself should limit opportunities for misconduct rather than relying solely on detection after the event (Dawson, 2020; Newton & Xiromeriti, 2024).

1 · Dynamic format

Visual, interactive, changing, moving or gamified content, unfamiliar response structures and tasks in which interaction forms part of the measured performance. The objective is not simply to replace text with images, which modern multimodal systems can also process, but to increase the number and complexity of actions required to outsource the task.

2 · Time restriction

Appropriate time limits reduce the opportunity to capture content, consult an AI system, interpret its response and enter the answer. Limits should be tight enough to create a meaningful cost for external assistance, but realistic for legitimate candidates, supported by trial data and justified for the intended construct and population (AERA et al., 2014).

3 · Information fragmentation

Reveal information progressively rather than in full: one question or option at a time, separate stages, no return to previous screens, unique response options. Fragmentation reduces the value of a single screenshot and increases the actions needed to reconstruct the task (Dawson, 2020; ITC, 2016).

DESIGN CONSIDERATION	VULNERABLE DESIGN	RESILIENT DESIGN
Dynamic format	• Static, text-heavy questions • Complete item captured in one screenshot • Minimal interaction required	• Visual, interactive or changing content • Active engagement is required • Format is harder to capture or describe
Time restriction	• Generous or unlimited time • Enough time to experiment with prompts • Limits ignore task complexity	• Tight but realistic limits • External help carries a meaningful time cost • Timings reflect reading and reasoning demands
Information fragmentation	• All information visible simultaneously • Question and options copied together • AI answer maps easily to the interface	• Information revealed progressively • Multiple captures needed to reconstruct the task • Answers are harder to transfer back accurately

C

ACT · SECOND LAYER

Candidate messaging

Candidate messaging acts primarily as a deterrent. It should remove ambiguity, raise the perceived risk of misconduct and make the consequences clear before the assessment begins. Evidence from academic-integrity and deterrence research suggests that clear norms, perceived detection risk and credible consequences can influence dishonest behaviour (McCabe et al., 2002; Nagin & Pogarsky, 2003; Ortiz-Bonnin & Blahopoulou, 2025).

4 · Clear boundaries

Tell candidates directly what is prohibited, including generative AI, online searches, assistance from another person, screenshots, second devices and the copying or sharing of content. Clear, explicit rules are preferable to vague appeals to honesty, particularly where norms around acceptable AI use are still developing.

5 · Credible monitoring

Candidates should understand that activity and results may be reviewed using session data, unusual response patterns, score anomalies, proctoring, later-stage comparisons or supervised verification. The aim is not to claim perfect detection, but to create credible uncertainty about whether misconduct will succeed undetected.

6 · Meaningful consequences

Tell candidates what may happen if misconduct is identified: invalidating the result, supervised reassessment, removal from the process, withdrawal of an offer or further action after appointment. The perceived certainty and meaningfulness of a sanction can matter more than vague or purely symbolic warnings.

MESSAGING CONSIDERATION	VULNERABLE APPROACH	RESILIENT APPROACH
Clear boundaries	• Refers vaguely to honest completion • Does not explicitly prohibit AI use • Rules are buried or incomplete	• Explicitly prohibits AI and outside help • Covers screenshots, searches and second devices • Uses direct language and acknowledgement
Credible monitoring	• Provides no reason to expect review • Uses generic statements with no follow-through • Makes experimentation appear low risk	• States that activity and results may be reviewed • Links suspicious results to further checks • Reflects monitoring actually in use
Meaningful consequences	• Uses vague phrases such as "action may be taken" • Provides little downside to attempting cheating • Does not connect misconduct to progression	• States specific consequences in advance • Includes invalidation or supervised reassessment • Makes clear that serious misconduct can end the process

ACT · THIRD LAYER

Technical solutions

Technical controls provide an additional layer of deterrence, evidence and verification. They should be proportionate and should support resilient design rather than compensate for a fundamentally vulnerable assessment. This layered approach is consistent with international guidance for computer-based testing and assessment security (International Test Commission, 2006, 2016).

7 · Session monitoring

Record browser focus loss, tab switching, exits from full-screen mode, copy-and-paste activity, unusual pauses, irregular response times and abrupt changes in performance. These indicators should not be treated as proof in isolation; combine multiple indicators to identify aberrant patterns for further review.

8 · Webcam proctoring

Identity and environment assurance through live or recorded footage, screen recording and review of flagged events. It can deter obvious misuse but cannot eliminate second devices or concealed assistance — and raises privacy, fairness and candidate-experience concerns that require proportionate implementation.

9 · Verification testing

Examine whether a candidate can reproduce earlier performance under more controlled conditions. It should measure the same capability using different items, be shorter where possible and apply clear rules to substantial discrepancies (Aguado et al., 2018; Guo & Drasgow, 2010; Makransky & Glas, 2011).

TECHNICAL CONSIDERATION	WEAK IMPLEMENTATION	STRONG IMPLEMENTATION
Session monitoring	• Records events without interpretation • Relies on a single flag • Treats flags as proof	• Combines focus, timing and response indicators • Uses patterns to trigger review • Communicates monitoring clearly
Webcam proctoring	• No identity or environment assurance • Data collected but rarely reviewed • Used inconsistently	• Confirms identity where appropriate • Monitors candidate, screen or environment • Uses clear criteria for flagged sessions
Verification testing	• No verification testing for shortlisted candidates • Repeats the same uncontrolled conditions • Large discrepancies have no consequence	• Uses different items under tighter controls • Targets shortlisted, high-scoring or flagged candidates • Meaningful discrepancies trigger review

Proportionate implementation

Not every assessment requires all nine ACT measures at maximum strength. The appropriate combination depends on the stakes, vulnerability of the format, candidate volume, selectivity, expected prevalence and effectiveness of cheating, privacy and accessibility, and the cost of poor decisions. Proportionate use is central to responsible assessment practice (American Educational Research Association et al., 2014; International Test Commission, 2016).

For lower-risk uses, resilient Assessment design and clear Candidate messaging may be sufficient. For high-volume or high-stakes selection, employers may also require stronger Technical solutions, including session monitoring, proctoring and verification. ACT should be applied proportionately rather than as a rigid checklist.

Controls that are too restrictive or intrusive can undermine validity, reliability, fairness, accessibility, completion rates and candidate experience. Applicant reactions can influence organisational attractiveness and the practical success of selection systems, while surveillance can create distinct ethical and privacy burdens (Coghlan et al., 2021; Hausknecht et al., 2004). The strongest strategy preserves assessment quality while materially reducing the expected value of cheating.

COMBINE THE LAYERS PROPORTIONATELY

A **Assessment design — the primary intervention**
The task itself limits the usefulness of AI.

C **Candidate messaging — the second layer**
Clear rules and credible consequences deter experimentation.

T **Technical solutions — scaled to the risk**
Controls differ greatly in burden; choose the lightest that gives the assurance the stakes require.

Lower-risk uses: **A + C** may be sufficient. High-volume or high-stakes selection: add **T** — session monitoring, proctoring and verification.

PROPORTIONATE BY DESIGN

Not every assessment needs all nine measures at maximum strength. The strongest strategy preserves assessment quality while materially reducing the expected value of cheating.

SECTION 6

Cheating personas and targeted interventions

06

Candidates who cheat should not be treated as a homogeneous group. One of the earliest experimental classifications of examination misconduct distinguished **individualistic-opportunistic** cheating from **individualistic-planned** cheating, alongside social forms of cheating. Individualistic-opportunistic cheating was described as unplanned and impulsive, whereas individualistic-planned cheating involved foresight and activity before the test situation (Hetherington & Feldman, 1964).

Although this distinction was developed before the advent of AI-assisted cheating, it is nevertheless a useful operational lens because the two groups are likely to respond differently to the ACT interventions. Wider research also supports separating deliberate intention from situational opportunity: intentions, attitudes, perceived control and social norms can predict dishonest behaviour, while experimental evidence indicates that increasing the perceived certainty of detection can reduce cheating (Beck & Ajzen, 1991; Nagin & Pogarsky, 2003).

Individualistic-planned cheaters

Individualistic-planned cheaters decide before the assessment that they intend to obtain unauthorised assistance. They may practise prompts, compare AI systems, seek leaked questions, prepare multiple devices, recruit collaborators and rehearse how content and answers will be transferred. This population is likely to be **smaller but more insidious** because its members are prepared to dedicate significant time, energy and resources to defeating the selection process. Evidence from contract cheating provides an imperfect but relevant parallel: only a small minority in a combined student sample reported using paid ghostwriting, but most of those who had done so reported doing it repeatedly (Curtis & Clare, 2017).

Candidate messaging alone is unlikely to deter this group. Clear rules and consequences remain necessary for fairness and enforcement, and a specific, credible statement that high scores will be verified may still change some decisions — but a generic integrity warning is unlikely to persuade someone who has already decided to cheat.

Assessment design can still reduce the efficacy of cheating by increasing the number of actions required, limiting the time available and making AI answers less reliable. However, determined candidates may bypass even strong design through extensive preparation, prompting, collaboration or technical workarounds. Design raises the cost of cheating, but it does not necessarily remove the intention to cheat.

Technical solutions are therefore proposed as the priority ACT layer for individualistic-planned cheaters. Session monitoring and proportionate proctoring can identify suspicious activity during the assessment, while identity checks and verification testing can expose performance that cannot be reproduced under controlled conditions. These controls are not infallible, but they create the strongest prospect of catching or exposing a candidate who is committed to proceeding despite the design and messaging.

Proposed priority: T. Monitoring, proctoring and verification create the strongest prospect of exposing a candidate committed to cheating despite design and messaging.

Individualistic-opportunistic cheaters

Individualistic-opportunistic cheaters do not necessarily begin the assessment with a fixed plan to cheat. They experiment with AI or other assistance when the task becomes difficult, when failure appears likely or when an easy opportunity presents itself. Generative AI has made this form of misconduct easier because a candidate can move from temptation to attempted assistance with little preparation.

Candidate messaging is likely to be effective for this group. Explicitly prohibiting AI use, explaining that activity and results may be reviewed, and stating meaningful consequences changes the immediate calculation at the point of temptation. The message should remove ambiguity and make experimentation feel risky rather than harmless. This is consistent with evidence that clear norms and a greater perceived certainty of detection can influence dishonest behaviour (McCabe et al., 2002; Nagin & Pogarsky, 2003).

Assessment design is also likely to be highly effective because it removes the flexibility needed to experiment. Tight but fair timings, dynamic formats and fragmented information reduce the time available to capture content, construct prompts, interpret AI output and transfer an answer back to the assessment. Where attempting to use AI creates an immediate time penalty and an uncertain result, independent completion remains the more attractive strategy.

Technical solutions are comparatively less important for this group. Where assessment design and candidate messaging are strong, they may already prevent the attempted behaviour, making intrusive monitoring redundant for many lower-risk uses. Technical controls can still operate as a backstop in high-stakes settings, but they should not be the default response when the first two ACT layers are sufficient.

PLANNED CHEATERS · PROPOSED PRIORITIES

- **T** Monitoring, proctoring, identity checks and verification
- **A** Design that raises the cost and unreliability of cheating

OPPORTUNISTIC CHEATERS · PROPOSED PRIORITIES

- **A** Tight timings and fragmented information that remove room to experiment
- **C** Clear boundaries and consequences at the point of temptation

Mapping ACT interventions to cheating personas

The priorities below are **proposed**, derived from the reasoning in this section. They are not effectiveness ratings: direct comparative evidence on these interventions in AI-assisted recruitment does not yet exist, and they should be revised as it accumulates.

CHEATING PERSONA	ACT INTERVENTION	PROPOSED PRIORITY	RATIONALE
Individualistic-planned	A Assessment design	SECONDARY	Reduces the efficiency and reliability of cheating, but may be bypassed through extensive preparation and resources.
	C Candidate messaging	SUPPORTING	Limited deterrent once the decision to cheat is made, although credible verification and consequences may still alter it; supports fair enforcement.
	T Technical solutions	PRIMARY	Monitoring and proctoring can identify misuse, while verification can expose performance that cannot be reproduced.
Individualistic-opportunistic	A Assessment design	PRIMARY	Tight timings, dynamic formats and fragmented information remove the room needed to experiment.
	C Candidate messaging	SECONDARY	Clear boundaries, credible monitoring and consequences can shift the risk calculation at the point of temptation.

Some candidates will cheat regardless

The ACT framework does not assume that every candidate can be deterred. Some people will attempt to cheat regardless of the rules, format or stated consequences. Given sufficient motivation, preparation and resources, a determined candidate could compromise even a robust remote selection tool in principle. Assessment providers and employers should therefore avoid claims that cheating can be reduced to zero (Dawson, 2020; International Test Commission, 2016).

That limitation does not make intervention futile. A substantial reduction can still be achieved by dissuading situational cheaters, removing the opportunity to experiment, reducing the advantage available to planned cheaters and increasing the probability that remaining misconduct is identified or exposed. Even if a small determined group remains, reducing both the number of attempts and the success rate of those attempts can materially protect shortlist quality and selection validity.

INTERVENTION OBJECTIVE

The goal is not to eliminate cheating. It is to deter candidates who can be dissuaded, constrain those who cannot, and detect or verify the remaining risk if needed.

SECTION 7

Monitoring live assessment data 07

AI resilience should not be judged solely through laboratory trials, supplier claims or theoretical design features. These can show that an assessment is likely to make cheating more difficult, but they cannot establish how candidates behave once it is deployed at scale. Operational verification studies illustrate the importance of examining real score discrepancies and response behaviour in live selection contexts (Aguado et al., 2018; Sanz et al., 2020).

Providers should examine live operational data to determine whether score distributions, response patterns and selection outcomes are changing in ways that may indicate successful AI use. Monitoring should be **continuous** because technology and candidate strategies will evolve, and security evidence can become obsolete as delivery conditions change (Tippins, 2015).

Why group averages are not enough. Comparing average scores before and after the adoption of generative AI is unlikely to reveal the full problem. If cheating is concentrated among a small proportion of candidates, large individual gains can be diluted within the overall mean. The relevant psychometric problem is aberrant person-level performance, not only movement in a group average (Meijer & Sijtsma, 2001; van der Linden & Guo, 2008).

A group mean is therefore a blunt indicator, poorly suited to detecting concentrated, high-impact misuse.

Analyse the upper tail

Monitoring should focus on the upper end of the distribution, where successful AI-assisted candidates are most likely to appear. The central question is whether unusually high scores are occurring more frequently than expected. Useful indicators include:

- The proportion scoring above the 90th, 95th and 99th percentiles
- The ratio of observed high scorers to the number expected from historical data
- Changes in the size and shape of the upper tail
- Unusual clusters of very high or near-perfect scores
- Inflation concentrated within particular roles, locations, devices or campaigns
- Gaps between assessment performance and later controlled stages

WHY AVERAGES MISLEAD

A stable average does not demonstrate that an assessment remains secure. The mean may barely move while the shortlist changes materially.

ILLUSTRATIVE METRIC · UPPER-TAIL INFLATION RATIO

10%
observed proportion of high scorers

÷

5%
expected proportion of high scorers

=

2.0
inflation ratio

The ratio should be calculated across several thresholds. A genuine improvement in applicant quality may shift much of the distribution, whereas cheating may produce a narrower increase among unusually high scorers.

Shortlist distortion matters more than average inflation

The most important operational risk is the proportion of shortlisted candidates whose scores were materially inflated. A relatively small cheating rate can create a serious selection problem when those candidates become concentrated among the highest scorers and displace independent candidates. Track:

- Estimated proportion of anomalous high scorers
- Shortlisted candidates drawn from an inflated upper tail
- Changes in pass rates and shortlist composition
- Score discrepancies vs later selection stages
- High scorers failing supervised verification

Combine score and process evidence. Upper-tail inflation is a warning signal, not proof that an individual has cheated. Combine it with irregular response times, repeated browser exits, component inconsistencies and failure to reproduce results under supervision (Aguado et al., 2018; Kantrowitz & Dainis, 2014; Wright et al., 2014). No single indicator should normally determine an individual outcome (American Educational Research Association et al., 2014).

Establish a credible baseline with fixed thresholds. Compare like with like: similar roles, populations, languages, channels, assessment versions and selection thresholds. Define "high scorer" against a fixed historical reference; recalculating percentiles within each new cohort can conceal inflation.

Interpret verification discrepancies carefully. Judge a gap between unproctored and supervised scores against expected measurement error, regression to the mean, practice effects and changed testing conditions before treating it as evidence of misuse.

Include a random sample in verification. Checking only flagged candidates cannot show how much misuse the flagging system misses. Where feasible, verify a random sample as well, to estimate the base rate.

A clean upper tail is not proof of security. Assistance may move candidates just across a pass threshold without producing exceptional scores. Monitor the pass-mark region and threshold crossings as well as the extreme tail.

Demonstrating that resilient design works

The ACT framework explains what providers should change. Buyers also need a standard for how providers show that the changes succeed. Claims of AI resilience should rest on evidence gathered under realistic delivery conditions, not on design descriptions alone (Aguado et al., 2018; Kantrowitz & Dainis, 2014; Lievens & Burke, 2011; Sanz et al., 2020). An adequate evidence base would meet five conditions.

1. **Compare unaided and deliberately AI-assisted performance under realistic conditions**
   Use the same device, timing and interface that candidates actually use, with testers instructed to obtain assistance as a motivated candidate would.

2. **Report score gains and changes in progression rates, not just AI answer accuracy**
   A model answering items correctly in isolation is not the same as a candidate gaining an advantage within the live assessment. Report the effect on scores and on who would have been shortlisted.

3. **State which AI systems, assistance methods and testing dates were used**
   Results are specific to the tools and strategies tested, and go stale as those change.

4. **Check that resilience has not been bought at the cost of measurement quality**
   Tighter timing and fragmented presentation must be shown to preserve reliability, the intended construct, accessibility and candidate experience.

5. **Repeat the evaluation as tools and candidate strategies change**
   AI-resilience evidence has a shelf life. State when it was last refreshed.

DESIGNED TO IMPROVE RESILIENCE

Features with a clear rationale — dynamic formats, time limits, fragmentation, monitoring — but without comparative evidence yet. These are **design propositions**, and should be described as such.

DEMONSTRATED REDUCTION IN AI-ASSISTED ADVANTAGE

Features whose effect on score gains and progression has been measured against the five conditions above, with tools, methods and dates stated.

Providers — Test Partnership included — should be explicit about which column each of their claims sits in. Where comparative evidence is not yet available, a feature should be described as designed to improve resilience, not as proven protection. Buyers should treat the absence of that distinction as a warning sign in itself.

A HIGHER STANDARD OF EVIDENCE

Claims of AI resilience should rest on evidence from live operating conditions — the upper tail, score and process indicators together, reproducibility under supervision, and effects on real selection outcomes.

SECTION 8

Recommendations for employing organisations

08

Employers should reassess whether their current assessment providers remain a safe option. Many organisations pay substantial sums for long-established tests because recognised providers historically appeared dependable and low risk. That choice was understandable, but **familiarity is not evidence of resilience**.

The assessment environment has changed. A provider may now represent greater risk if its products were designed for a pre-generative-AI world and the organisation lacks the mobility or willingness to adapt. Established research on unproctored testing and newer evidence on generative AI both support renewed scrutiny of legacy delivery assumptions (Newton & Xiromeriti, 2024; Tippins et al., 2006).

Reconsider what "safe" means. A provider should not be considered safe merely because it is large, established or expensive. The relevant question is whether its assessments remain appropriate for the conditions in which candidates now complete them. A familiar product can create false reassurance when its security assumptions are outdated.

WARNING SIGNS OF OUTDATED SECURITY ASSUMPTIONS

- ⚠ Static, text-heavy content
- ⚠ Predictable response formats
- ⚠ Generous time limits
- ⚠ Designs that have changed little in years
- ⚠ Fully visible information
- ⚠ Technical controls added around an otherwise unchanged test

Size and familiarity are not evidence either way

Large providers often have substantial resources, and may also carry large existing item banks, established platforms and long governance processes. Neither fact settles the question: resources can fund redesign, and legacy can slow it.

Employers often pay a premium for established brands on the assumption of lower risk. That premium is justified only if the provider can show what it has changed and what evidence supports the change. Where it cannot, the premium is **buying reputation rather than resilience**.

The issue is not size itself. It is whether the provider is willing and able to dedicate research, product and development resources to the new threat, and whether it can demonstrate the result. A smaller provider with strong technical expertise may now be the safer choice; so may a large one that has genuinely redesigned. Buyers should test the evidence, not the brand.

RECONSIDER WHAT SAFE MEANS

A provider is not safe merely because it is large, established or expensive.

Broaden the vendor search

Employers should consider providers that are agile enough to redesign quickly, willing to challenge established formats, investing in psychometric and product research, transparent about limitations, and grounded in how candidates actually use technology.

Innovation should not be judged by branding or feature lists. Employers should ask **what has materially changed in the assessment itself** and what evidence shows that the change matters. Credible innovation requires substantive design changes and empirical support, not reassuring platitudes or unrealistic claims of complete protection.

Buyers have a legitimate concern about a serious issue that can materially affect the validity, fairness and return on investment of assessment-based selection. These concerns should be treated with appropriate attention and respect, rather than dismissed, minimised or answered with vague reassurance.

The ground has shifted

Generative AI has disrupted assumptions that once underpinned recruitment. Polished written applications may not reflect the candidate's communication skills, tailored cover letters may not demonstrate motivation, and conventional remote test performance may not always reflect unaided capability. Emerging field evidence on AI-assisted cover letters and broader assessment research support this shift in interpretation (Cui et al., 2025; Newton & Xiromeriti, 2024).

This is a disruptive period, and agility will be essential because the threat is not static. Employers should expect assessment strategy to evolve more quickly and should work with providers that monitor, test, learn and adapt continuously. Moreover, as AI systems become more capable, faster, cheaper and more widely used, the pressure on assessment security will continue to grow. **Delay is likely to increase exposure** as AI capabilities and candidate familiarity with them grow, although the pace of that change is uncertain.

Assessments are needed more than ever

Importantly, the rise of generative AI does not reduce the need for assessment; **it increases it**. AI lowers the cost of producing and tailoring applications, weakening the differentiation provided by CVs, cover letters and written application questions at the earliest stages of selection, as the freelance-platform evidence illustrates (Cui et al., 2025). Employers may therefore face larger applicant pools while having less reliable written evidence on which to base shortlisting decisions.

Abandoning assessment and returning to CV-led screening would move the process towards methods that are even more vulnerable to AI assistance. The appropriate response is not to use fewer assessments, but to use better assessments that preserve meaningful evidence of candidate capability and retain established criterion-related validity and utility (Schmidt & Hunter, 1998).

Addressing the issue early also makes assessment easier to defend internally. If senior leaders first encounter the threat through a visible failure, one possible reaction is to remove assessments altogether — which would leave the organisation relying on the methods AI has made least dependable. The strategic objective should therefore be to strengthen assessment, not abandon it, preserving robust evidence of candidate capability while reducing reliance on the weakest signals.

The strategic objective is to strengthen assessment, not abandon it — preserving robust evidence of capability while reducing reliance on the methods AI has made least dependable.

Assess whether current providers remain suitable

Employers should not renew solely because switching feels risky. The risk of change must be weighed against the risk of remaining with a product that has not adapted.

AREA TO REVIEW	QUESTION FOR THE PROVIDER
Assessment design	What has changed in the item format, timing and information delivery in response to AI?
Evidence	Are live distributions, upper-tail inflation and shortlist outcomes monitored?
Verification	Can unusually high performance be checked under controlled conditions?
Transparency	Does the provider acknowledge limitations and explain where risk remains?
Technical controls	Are monitoring and proctoring proportionate, interpretable and accessible?
Research & development	Is the provider investing in redesign or relying mainly on marketing and surveillance?
Agility	How quickly can the provider respond when new vulnerabilities emerge?

A practical direction

1. **Audit the full selection process.** Include CVs, cover letters, application forms, interviews and assessments.
2. **Challenge existing providers.** Ask for evidence of design changes, live monitoring, upper-tail analysis and verification outcomes.
3. **Compare a broader range of vendors.** Include agile providers with demonstrable assessment-design and research capability.
4. **Prioritise resilient design.** Look beyond proctoring and browser tracking to the structure of the task itself.
5. **Use proportionate technical controls.** Apply stronger assurance where the stakes and risks justify it.
6. **Monitor shortlist outcomes.** Focus on whether anomalous candidates are concentrated among those progressing.
7. **Review regularly.** AI resilience is an ongoing governance issue, not a one-off procurement question.

CONCLUSION

Neither denial nor panic

Generative AI has altered the meaning of security in recruitment. Selection methods that once appeared safe may now be highly vulnerable, while familiar application materials can be generated or improved in ways that weaken their evidential value (Cui et al., 2025; Newton & Xiromeriti, 2024).

The correct response is neither denial nor panic. Employers should not abandon assessment, and providers should not rely on surveillance alone. A balanced ACT strategy begins with resilient **Assessment design**, reinforces it through clear **Candidate messaging**, and applies proportionate **Technical solutions** alongside verification and continuous evidence gathering (International Test Commission, 2016; Tippins, 2015).

The organisations that navigate this period best will be those willing to question inherited assumptions, broaden their vendor options and prioritise evidence, innovation and adaptability over reputation alone. Those that cannot respond convincingly are likely to face harder questions from senior leaders about whether assessment still earns its place — questions that are far easier to answer with evidence gathered in advance.

> “
>
> The old rules no longer apply. The need for valid assessment has not diminished, but the standard for delivering it has changed.

References

Aguado, D., Vidal, A., Olea, J., Ponsoda, V., Barrada, J. R., & Abad, F. J. (2018). Cheating on unproctored Internet test applications: An analysis of a verification test in a real personnel selection context. *The Spanish Journal of Psychology, 21*, e62. https://doi.org/10.1017/sjp.2018.50

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

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored Internet-based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology, 2*(1), 39–45. https://doi.org/10.1111/j.1754-9434.2008.01105.x

Beck, L., & Ajzen, I. (1991). Predicting dishonest actions using the theory of planned behavior. *Journal of Research in Personality, 25*(3), 285–301. https://doi.org/10.1016/0092-6566(91)90021-H

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Boudreau, J. W. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology, 36*(3), 551–576. https://doi.org/10.1111/j.1744-6570.1983.tb02236.x

Coghlan, S., Miller, T., & Paterson, J. (2021). Good proctor or "Big Brother"? Ethics of online exam supervision technologies. *Philosophy & Technology, 34*(4), 1581–1606. https://doi.org/10.1007/s13347-021-00476-1

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). University of Illinois Press.

Cui, J., Dias, G., & Ye, J. (2025). Signaling in the age of AI: Evidence from cover letters on an online freelance platform [Preprint]. arXiv.

Curtis, G. J., & Clare, J. (2017). How prevalent is contract cheating and to what extent are students repeat offenders? *Journal of Academic Ethics, 15*(2), 115–124. https://doi.org/10.1007/s10805-017-9278-x

Dawson, P. (2020). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education.* Routledge. https://doi.org/10.4324/9780429324178

Dendir, S., & Maxwell, R. S. (2020). Cheating in online courses: Evidence from online proctoring. *Computers in Human Behavior Reports, 2*, 100033. https://doi.org/10.1016/j.chbr.2020.100033

Gilbert, S. J., Boldt, A., Sachdeva, C., Scarampi, C., & Tsai, P.-C. (2023). Outsourcing memory to external tools: A review of "intention offloading." *Psychonomic Bulletin & Review, 30*(1), 60–76. https://doi.org/10.3758/s13423-022-02139-4

Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology, 74*(9), 1477–1496. https://doi.org/10.1177/1747021821100060

Gruenhagen, J. H., Sinclair, P. M., Carroll, J.-A., Baker, P. R. A., Wilson, A., & Demant, D. (2024). The rapid rise of generative AI and its implications for academic integrity. *Computers and Education: Artificial Intelligence, 7*, 100273. https://doi.org/10.1016/j.caeai.2024.100273

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored Internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment, 18*(4), 351–364. https://doi.org/10.1111/j.1468-2389.2010.00518.x

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Hetherington, E. M., & Feldman, S. E. (1964). College cheating as a function of subject and situational variables. *Journal of Educational Psychology, 55*(4), 212–218. https://doi.org/10.1037/h0045337

Hylton, K., Levy, Y., & Dringus, L. P. (2016). Utilizing webcam-based proctoring to deter misconduct in online exams. *Computers & Education, 92–93*, 53–63. https://doi.org/10.1016/j.compedu.2015.10.002

International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing, 6*(2), 143–171. https://doi.org/10.1207/s15327574ijt0602_4

References

International Test Commission. (2016). International guidelines on the security of tests, examinations, and other assessments. *International Journal of Testing, 16*(3), 181–204. https://doi.org/10.1080/15305058.2015.1111221

Kantrowitz, T. M., & Dainis, A. M. (2014). How secure are unproctored pre-employment tests? Analysis of inconsistent test scores. *Journal of Business and Psychology, 29*(4), 605–616. https://doi.org/10.1007/s10869-014-9345-x

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67*(1), 241–293. https://doi.org/10.1111/peps.12052

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability. *Journal of Occupational and Organizational Psychology, 84*(4), 817–824. https://doi.org/10.1348/096317910X522672

Makransky, G., & Glas, C. A. W. (2011). Unproctored Internet test verification using adaptive confirmation testing. *Organizational Research Methods, 14*(4), 608–630. https://doi.org/10.1177/1094428110370715

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2002). Honor codes and other contextual influences on academic integrity. *Research in Higher Education, 43*(3), 357–378. https://doi.org/10.1023/A:1014893102151

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Nagin, D. S., & Pogarsky, G. (2003). An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology, 41*(1), 167–194. https://doi.org/10.1111/j.1745-9125.2003.tb00985.x

Newton, P. M., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review. *Assessment & Evaluation in Higher Education, 49*(6), 781–798. https://doi.org/10.1080/02602938.2023.2299059

Nye, C. D., Do, B. R., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment, 16*(2), 112–120. https://doi.org/10.1111/j.1468-2389.2008.00416.x

Ortiz-Bonnin, S., & Blahopoulou, J. (2025). Chat or cheat? Academic dishonesty, risk perceptions, and ChatGPT usage in higher education students. *Social Psychology of Education, 28*, 113.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Sanz, S., Luzardo, M., García, C., & Abad, F. J. (2020). Detecting cheating methods on unproctored Internet tests. *Psicothema, 32*(4), 549–558.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*(6), 609–626. https://doi.org/10.1037/0021-9010.64.6.609

Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology, 59*(1), 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

Wright, N. A., Meade, A. W., & Gutierrez, S. L. (2014). Using invariance to examine cheating in unproctored ability tests. *International Journal of Selection and Assessment, 22*(1), 12–22. https://doi.org/10.1111/ijsa.12053

ABOUT TEST PARTNERSHIP

Faster, more accurate assessments that candidates actually enjoy.

Test Partnership delivers cognitive and behavioural assessments that balance scientific rigour with genuine candidate experience. Our MindmetriQ™ platform is designed around the principles in this paper — dynamic formats, tight timings, fragmented information and live monitoring — and we hold it to the evidence standard the paper sets out.

96%	6,000+	3m+
candidate satisfaction	organisations	assessments completed

Talk to our team about AI-resilient assessment ›

testpartnership.com