

Aincrad

Playable work samples for hiring, training, and qualification

Working paper, September 2026

Aincrad is all about turning real work into a video game. Instead of asking a candidate to tell a camera what they would do, we put them inside a version of the job, let the system change around them, and record what they actually do.

An investment-banking candidate should be able to repair a model, review a deal room, respond to a changed assumption, and explain a recommendation. A software candidate should be able to inspect a failing service, form a hypothesis, ship a fix, and respond when the first fix breaks something else. A data-center technician should be able to read the work order, choose the right tools, isolate a fault, stop when the condition is unsafe, and leave behind a clean record. These are all games in the useful sense of the word: systems with state, rules, tools, constraints, decisions, and consequences.

The goal is to replace indirect screening with observable work wherever a simulation can reproduce the important part of the job. The game is the interface. The product underneath it is a versioned evidence system that connects the job to the scenario, the scenario to behavior, behavior to a score, and the score to a decision that a person can inspect.

Proving Ground is the first product under Aincrad. It applies this system to data-center hiring and qualification. The same platform can later support software companies, game studios, investment banks, industrial services, and other jobs where the order and quality of decisions matter.

The problem

Most hiring asks people to describe work instead of doing it.

A resume is a compressed story about prior opportunity. A one-way video interview is a rehearsed answer recorded under an artificial set of constraints. A generic reasoning test samples broad ability. A personality questionnaire asks the candidate to report on the candidate. A structured interview can be useful, but it still observes an account of behavior rather than the behavior itself.

Each method can contribute useful information. None should be dismissed casually. The problem is the distance between the evidence collected and the work the employer needs done.

That distance becomes obvious in jobs with visible operating state. A data-center technician does not succeed by saying the words “safety first.” The technician succeeds by noticing that the supposedly dead circuit is still energized, refusing to continue, documenting the condition, and escalating through the right path. An analyst does not succeed by saying “I pay close attention to detail.” The analyst succeeds by finding that the debt schedule and cash-flow statement no longer reconcile after a late change to the financing case. A software engineer does not succeed by describing debugging. The engineer succeeds by choosing a high-information test, reading the result, and changing course.

The U.S. Office of Personnel Management defines work samples as tasks that are identical or highly similar to the work itself. OPM notes their strong job relevance and warns that they can be expensive to develop and maintain.[1] A classic meta-analysis found meaningful criterion-related validity for work-sample tests, while a later re-analysis of selection research cautioned against treating any old headline validity coefficient as universal.[2, 3] Applicant-reaction research has also found that interviews and work samples tend to be perceived more favorably than more abstract selection tools.[5]

So the opening is not that hiring science has ignored work samples. The opening is that rich, repeatable work samples have historically been too expensive to build, operate, vary, score, and maintain for most jobs.

What HireVue solved, and what remains

HireVue solved enterprise distribution. It put video interviewing, structured content, assessments, scheduling, integrations, governance, and fraud controls into one procurement surface. Its current products already include game-based psychometric assessments, Virtual Job Tryouts, and an AI-assisted Assessment Builder. HireVue says it offers more than 40 job-specific Virtual Job Tryouts across roles and industries, and that its game-based products measure broad competency domains.[6, 7, 8] HireVue also publishes a substantial science and governance position, including validation, calibration, fairness review, ongoing monitoring, and human decision authority.[9]

HireVue’s current surface also includes an AI Interviewer that creates job-description-derived questions and rubrics, conducts probing conversations, evaluates transcript content against role criteria, and sends results to the applicant-tracking system. HireVue states that it does not score facial expressions, body language, appearance, background, or tone, and advertises opt-out, accommodation, and human-review paths.[10, 9] These are vendor-reported capabilities, not independent evidence that the product predicts job performance in every use.

That is the standard Aincrad has to beat. It is not enough to put an avatar in front of a quiz. It is not enough to generate questions from a job description. It is not enough to collect more clicks.

Aincrad is built around a narrower bet. When the job can be represented as a stateful episode, the employer should be able to watch the candidate perform the work, inspect the path, and connect that path to later outcomes. The replacement target is the part of the funnel where a generic screen or recorded answer stands in for a job-relevant demonstration.

Method	What it observes well	What Aincrad tries to add
Resume	Prior roles, credentials, written career history	Direct evidence from a standardized job episode
One-way video	Communication and prepared responses	Decisions inside changing state, not only explanations
Generic psychometric game	Broad constructs through short abstract tasks	Concrete, employer-specific work and artifacts
Virtual job tryout	Realistic role scenarios, job preview, judgment, and a multi-measure fit profile	Proposed differentiation through deterministic replay, semantic trace, and version-specific inspectable evidence
Coding test or take-home	Code production and technical artifacts	Stateful incidents, system effects, prioritization, and recovery

The aim is not to remove every interview. It is to make interviews downstream of evidence. A manager

should be able to ask, “Why did you choose this test after the alert changed?” instead of “Tell me about a time you solved a hard problem.”

Why video games

The word “game” creates the wrong picture for some people. Aincrad does not require points, mascots, fantasy worlds, twitch controls, or photorealistic 3D. A game is useful here because it is a controlled system.

Every Aincrad episode has:

- a starting state;
- a goal or work order;
- information the candidate may inspect;
- tools and actions the candidate may use;
- constraints on time, money, safety, quality, or capacity;
- events that change the state;
- consequences for each action;
- an end state and a record of how it was reached.

That structure is the backbone of the product. It gives Aincrad four things a static screen cannot give as easily.

Process, not only outcome

Two candidates can reach the same answer in different ways. One forms a plausible hypothesis, runs the cheapest discriminating test, updates after new evidence, and documents the result. The other clicks through every option until something works. A final score can hide that difference. A semantically defined event trace can make the paths inspectable. Telemetry does not validate the interpretation by itself.

Evidence-centered design gives this idea a formal shape: define the claim about a person, define the evidence that would support it, then design the task that can elicit that evidence.[11] Educational research on stealth assessment applies the same logic to game environments, where actions can provide evidence without interrupting the experience with a separate test question. It is a measurement framework, not employment-validity evidence.[12]

Changing state

Real work changes while it is being done. The server that looked healthy produces a second alert. The client cuts the purchase price. The supposedly isolated bus still shows voltage. A game can introduce the new fact after the candidate has committed to a plan. That lets the employer observe behavior after a programmed change. Calling that behavior adaptability requires separate construct and criterion evidence.

Equivalent variation

The same underlying mechanic can generate several forms. The fault can move. The budget can change. The evidence can arrive in a different order. The candidate should face the same construct without receiving the same route. McKinsey reported that its Imbellus-built Solve assessment generated thousands of permutations, and also said that proving comparability across those forms required extensive testing.[13]

The second half matters. Randomness is not equivalence. Aincrad has to know which parameters changed, why the forms should be comparable, and whether real completion data supports that claim.

A reusable mechanic layer

Different jobs share operating patterns. A technician isolates a failing component. An engineer isolates a failing service. An analyst isolates the assumption that broke the model. The skin is different. The mechanic is diagnosis.

The first reusable mechanic library is deliberately small:

- diagnose and test hypotheses;
- allocate scarce resources;
- sequence dependent work;
- prioritize under interruption;
- inspect, verify, and document;
- stop, escalate, or ask for missing authority;
- understand second-order effects in a system;
- collaborate when information is split across people.

Aincrad shares mechanics, evidence, runtime, and governance. It does not pretend that a data center, an integrated development environment, and a financial model need the same renderer.

What the research actually supports

Research supports the direction, but it does not support the claim that games are automatically better or impossible to game.

Finding	What it supports	What it does not prove
Work samples mirror job tasks and can show strong job relevance.[1, 2]	Build from critical work, not generic game mechanics.	Any attractive simulation predicts performance.
Updated meta-analytic estimates place structured interviews near .42 mean operational validity and work samples near .33, with substantial variation.[4]	Use simulation and structured follow-up as complementary evidence, then test local incremental value.	A simulation is inherently more predictive than a structured interview.
Applicants tend to react well to interviews and work samples.[5]	Job relevance can improve trust and face validity.	Every candidate will enjoy a hiring game.
Serious-game meta-analyses report positive average learning effects in studied education and training settings, with important design moderators.[14, 15]	Games can be useful for the separate training lane.	Learning effects automatically transfer to workplace performance or establish hiring validity.
Across 52 samples in 44 papers, game-related indicators correlated with traditional cognitive tests at observed ($r=.30$) and corrected ($r=.45$), with substantial heterogeneity.[16]	Game behavior can contain cognitive signal.	A game is interchangeable with a validated cognitive battery.
The personnel-selection literature is still heterogeneous and construct evidence is inconsistent.[17]	Every pack needs its own evidence and limits.	The category has a universal validity advantage.
In one cognitive game-based assessment, prior video-game usage was positively associated with performance.[18]	Practice, device checks, and nuisance-variable studies matter.	A familiar interface removes all measurement error.
In one honesty-humility experiment, the game-based version showed less intentional distortion, but participants could distort both formats.[19]	Integrity should be measured and managed.	A game is cheat-proof or fake-proof.

The strongest evidence for Aincrad is therefore a combination of two established ideas: work samples can be job-relevant, and digital environments can capture a detailed process trace. The hard work is showing that the trace means what Aincrad says it means for a specific job, version, population, and decision.

The company should never sell “scientifically validated games” as a blanket property. Validity belongs to a score interpretation and use. Change the job, instructions, time limit, language, interface, scoring weights, threshold, population, or customer process, and the evidence may no longer transfer.

The precedent: Imbellus, McKinsey, and Roblox

Imbellus is the clearest historical precedent. It combined learning science, psychometrics, data science, engineering, and game design in assessments designed to infer problem-solving constructs from simulation behavior. Public research on its work with McKinsey describes cognitive task analysis, expert models, telemetry, and feature selection rather than a simple win-loss score. The preliminary study involved 527 optional McKinsey candidates and compared scores with the existing Problem Solving Test and hiring progression, not later job performance.[20]

McKinsey began working with Imbellus in 2016. By 2020 it reported that more than 15,000 people in 30 countries had played Solve. McKinsey later described a game-based innovation lab of about 25 people and said Solve was reaching roughly 300,000 early-career candidates per year.[13, 21] These company-reported figures establish deployment scale, not predictive validity or quality of hire.

Roblox acquired substantially all of Imbellus’s assets in 2020. Its SEC filing valued the acquired technology at \$11.7 million. Roblox then used the assessment technology in its own hiring.[22, 23] Roblox’s current early-career process combines problem-solving games built on Roblox, workplace scenarios, and coding or algorithmic assessments. It also provides untimed practice games so candidates can learn the interface before the scored experience.[24, 25] Public sources document the process and company claims, but do not disclose independent evidence linking scores to later performance, retention, or productivity.

Roblox matters for another reason. Roblox Studio demonstrates that a playable system can be built, tested, published, updated, emulated across devices, and distributed through one platform.[27] Roblox reported an average 127 million daily active users across 2025, a company metric that should be read as platform-distribution evidence rather than hiring evidence.[26]

The lesson is not that Aincrad should become Roblox. The lesson is that the game layer and the assessment layer can be treated as infrastructure. Taken together, the multi-year development history, extensive testing, and dedicated multidisciplinary teams suggest that bespoke assessment development remained specialized.

Aincrad’s product hypothesis is that AI can lower parts of the authoring cost. It can draft scenes, rules, variants, dialogue, assets, tests, and scoring proposals. It cannot manufacture the missing job analysis, outcome data, legal basis, or human accountability. Aincrad should use AI to author, then compile the result into a deterministic, reviewable runtime.

What Aincrad builds

Aincrad has four products inside one system.

Aincrad Studio

The Studio turns a role, source documents, and desired evidence into a playable episode. It asks for the job description, standard procedures, tools, common errors, safety boundaries, artifacts, and examples of excellent and unacceptable work. It then proposes observable claims, mechanics, variants, and a scoring plan.

The first Studio is constrained. It does not accept arbitrary customer code. It composes approved mechanics, state objects, actions, faults, documents, and evidence rules. A customer can change context and parameters without silently inheriting the validity of a different assessment.

Aincrad Runtime

The Runtime delivers the experience, maintains authoritative state, records semantic events, applies the versioned rules, and can replay the session from the same seed and input log. It manages identity, time, disconnects, accommodations, data retention, and integrations.

The candidate's computer can render a rich environment. The decision evidence should not depend on an opaque client-side score. Hidden rules stay on the server. Every result is tied to the exact scenario, engine, rubric, threshold, seed, and accommodation state that produced it.

Aincrad Evidence

The Evidence layer maps events to claims. It can say that the candidate opened the one-line diagram before isolation, tested for absence of voltage, encountered an unexpected reading, stopped, and escalated with the correct equipment identifier. It can also show the missing evidence.

This layer produces a replay, a timeline, evidence items, dimension scores, safety gates, uncertainty, reviewer notes, and a signed decision record. An LLM may summarize the trace or help a reviewer navigate it. It should not be the final, unreproducible pass-fail authority.

Aincrad Registry

The Registry states what each pack and version is allowed to claim. It records the intended role and use, job analysis, content owner, scoring specification, reliability and validity evidence, population, accessibility review, subgroup monitoring, approvers, known limits, and expiry.

This is the trust boundary. A beautiful prototype can publish immediately as a prototype. A training experience can publish when it is safe and instructionally reviewed. A consequential hiring score needs a governed path through shadow use, evidence collection, approval, and monitoring.

Job and source material → observable claims → game mechanics → versioned episode → event trace → evidence packet → human decision → later work outcome

The end-to-end workflow

The customer should be able to begin with a sentence close to this:

Build a 25-minute browser simulation for a junior data-center technician. Test safe isolation, fault diagnosis, information search, documentation, and escalation. The real job is open-book. Do not require prior gaming experience.

For an entry-level role, the episode may test only proficiency genuinely required on day one. If the employer expects a skill to be developed after hire, Aincrad should use a standardized learn-then-perform design or place the episode in Ramp or qualification instead of screening novices on prior opportunity.[4]

Aincrad then follows a controlled workflow.

1. **Collect the work.** Import the job description, work orders, SOPs, diagrams, equipment list, real artifacts, critical incidents, and examples of good and bad execution.
2. **Write observable claims.** Replace “critical thinking” with statements such as “chooses the highest-information safe test before replacing equipment” or “stops when the isolation boundary is contradicted.”
3. **Choose mechanics.** Match each claim to diagnosis, sequencing, allocation, prioritization, verification, documentation, or escalation.
4. **Generate the episode.** Draft state, objects, actions, constraints, events, documents, assets, variants, and evidence rules.
5. **Compile and test.** Run schema checks, deterministic replays, scripted agents, reachability tests, exploit probes, performance checks, and candidate-safe bundle checks.
6. **Review the work.** A domain expert confirms that the environment behaves like the job. An assessment reviewer confirms that the evidence supports the intended claims. An accessibility reviewer checks irrelevant barriers.
7. **Freeze a version.** The approved scenario, engine, rubric, thresholds, forms, and allowed uses become immutable.
8. **Run in the right mode.** Prototype, training, shadow assessment, or decision use each has a separate data and guidance policy.
9. **Inspect the evidence.** The employer sees the result, replay, critical moments, score derivation, uncertainty, and review history.
10. **Connect later outcomes.** The customer imports training, supervisor, quality, safety, or productivity outcomes so Aincrad can test whether the score adds useful information.

The customer does not get a universal magic score. The customer gets a documented instrument that can improve as evidence accumulates.

Publication levels

One creator can support several trust levels without pretending they are equivalent.

Level	What the customer can do	What Aincrad can say	Decision use
Prototype	Generate, play, share privately, edit	Playable scenario	None
Training-ready	Teach, hint, retry, explain, report completion	Learning activity for a defined objective	Learning only
Shadow assessment	Run without affecting the decision	Evidence is being collected and calibrated	No pass-fail effect
Assessment-grade	Use an approved version inside its documented scope	Version-specific score with stated evidence and limits	Bounded hiring or qualification use

This solves a central product conflict. A customer should not need an industrial-organizational psychologist to make a prototype. A customer should not be able to make a consequential employment decision from an unreviewed prototype.

The candidate experience

The candidate should know what is happening.

Before the episode, Aincrad explains what type of work will appear, what data will be captured, how long the experience should take, what device is needed, how the result will be used, and how to request an accommodation. The candidate gets an untimed interface practice that uses different content from the assessment. Roblox’s current practice-game approach is an implementation precedent intended to build interface familiarity through separate content.[24] Roblox research also found that public rewards attracted non-candidate and off-task play, a warning that incentives can contaminate telemetry and participant populations.[28]

During the episode, the interface should reproduce only the fidelity that matters. If spatial behavior is not scored, the candidate should not lose points for navigating a 3D room. A fixed workbench, spreadsheet, terminal, diagram, inbox, or conversation may be more faithful to the job than a free-roaming world.

If the real work is open-book, the episode should be open-book. A technician should be able to read the MOP. An analyst should be able to inspect the source file. An engineer should be able to search the repository and use ordinary documentation. The evidence is often whether the candidate knows what to inspect, how to interpret it, and when the written procedure no longer fits the state.

After the episode, the candidate receives a clear completion record and the process for technical review, accommodation, correction, or appeal. Feedback must be designed carefully. Training feedback can be detailed. Hiring feedback cannot reveal protected routes or scoring keys, but it can still state the assessed areas and whether a technical issue affected the run.

Aincrad should not infer emotion, honesty, disability, personality, anxiety, or “culture fit” from face, voice, posture, or mouse movement. The Department of Justice specifically warns that online interactive games used in hiring must measure relevant job skills rather than unrelated sensory, manual, or speaking limitations.[29] The interface must support keyboard access, remapping, captions, non-color cues, adjustable time where appropriate, assistive technology, and an equivalent alternative when the

interface itself is not essential to the job. WCAG 2.2 is the baseline, not the complete employment-accommodation policy.[30]

How scoring works

Aincrad separates observation from inference.

An event is an observation: “candidate opened E-401 at 03:14.” An evidence item is a rule-backed statement: “candidate inspected the source document before selecting the isolation point.” A dimension score combines evidence under a published version. A reviewer decision adds accountable judgment. A hiring decision still belongs to the employer.

The score should prefer a small number of legible claims over hundreds of latent features. Rich telemetry is useful when it preserves the path. It becomes dangerous when Aincrad turns every hesitation into an unsupported psychological inference.

The scoring stack has five layers:

1. **Outcome.** Was the work product correct, complete, safe, and on time?
2. **Critical gates.** Did the candidate cross a boundary that should override the total score, such as energizing unsafe work or fabricating a test record?
3. **Process evidence.** Which information, tests, actions, revisions, and escalations appeared in the trace?
4. **Efficiency and quality.** Did the candidate use proportionate steps, avoid rework, preserve service, and leave a useful record?
5. **Human review.** Does the replay support the automated derivation, and is there context the rules missed?

Speed is never a default virtue. It counts only where pace is part of the job and the interface does not make speed a proxy for gaming history, reading rate, disability, network quality, or device performance.

Reviewers need calibration cases, double-scored samples, disagreement tracking, and periodic rechecks. A signed review is useful only when the reviewer role, rubric, evidence, and confidence are visible.

Proving Ground: the first Aincrad product

Proving Ground applies Aincrad to data-center hiring, qualification, and ramp. It is the first pack, not a separate company and not a dead-end vertical.

The product thesis is simple: candidates complete realistic data-center work orders in a browser-based simulation. The system records the work, re-simulates the session, derives inspectable evidence, and lets an expert sign the result. Employers decide from the replay and evidence packet instead of a resume claim or a generic screen.

The current prototype already contains important platform primitives: a deterministic simulation core, versioned scenario definitions, semantic telemetry, evidence rules, replay, a reviewer workbench, an employer console, separate practice and assessment modes, and a public verification surface. The implemented scenarios cover orientation, rack execution, structured cabling, lockout and tagout judgment, and parts of UPS fault isolation. This is product implementation evidence. It is not yet evidence of

predictive validity, a live credentialed reviewer network, or comprehensive data-center readiness.

That distinction stays inside the white paper because it is part of the product truth.

The external task base is concrete. Microsoft’s current data-center-technician description includes procedure-based diagnosis, component replacement, grounding and cabling quality checks, incident reporting, and escalation.[33] Proving Ground does not copy one employer’s procedure. It turns a customer-approved task analysis and source set into a versioned episode.

A representative hiring episode

The candidate receives a work order to replace a branch breaker on a power distribution unit. The documents include the method of procedure, a one-line diagram, the equipment label, the permit state, and test records. The environment includes a gear rack, lockout devices, meter, panel, radio, and several plausible wrong paths.

The candidate has to:

1. inspect the work order and identify the correct equipment;
2. select the required protective equipment;
3. identify the upstream source and establish the isolation boundary;
4. apply lockout and tagout in the correct ownership and sequence;
5. perform a live-dead-live meter check and verify absence of voltage;
6. respond when a supposedly dead point still shows voltage;
7. stop the work, preserve the safe state, and escalate with useful detail;
8. complete the test record and explain the decision.

The strongest candidate may never replace the breaker. If the state is unsafe, stopping is the correct work. A conventional task that scores only completion can reward the most dangerous person in the room. A stateful simulation can make refusal and escalation part of the success condition.

Claim	Observable evidence	Evidence packet
Reads before acting	Opens work order, one-line, and equipment data before isolation	Document timeline and identifiers
Establishes safe boundary	Selects correct source, PPE, and lock sequence	State replay and gate result
Tests rather than assumes	Performs live-dead-live and places leads on defined points	Instrument trace and readings
Adapts to contradiction	Stops after unexpected voltage instead of forcing completion	Critical-moment replay
Escalates well	Names asset, condition, test, and current safe state	Timestamped radio or written record
Documents the job	Leaves an accurate test record with unresolved condition	Completed artifact and audit trail

NVIDIA’s current GB200 and GB300 documentation shows why the domain is becoming more stateful, not less. A rack-scale system combines compute trays, NVLink switch trays, power shelves, management networks, high-speed fabric, and liquid-cooling manifolds. NVIDIA’s validation checklists require

power, cooling, networking, hardware inventory, and system state to be verified together.[31, 32] These systems create good future episodes because the technician has to reason across several coupled sub-systems.

Those future episodes must be labeled honestly. The present Proving Ground build does not yet demonstrate open-ended rack optimization, high-speed throughput diagnosis, complete liquid-cooling commissioning, generators, or full OEM qualification. They are roadmap directions that require new content, domain review, and evidence.

Verify and Ramp

Proving Ground has two lanes with a hard boundary.

Verify is the pre-hire or qualification lane. It has no procedural hints, no answer-revealing prompts, no visible rubric, protected variants, deterministic scoring, replay, and accountable review. Documents remain available when the real job is open-book.

Ramp is the learning lane. It can use hints, checklists, worked examples, explanations, retries, and customer procedures. The learner sees gaps and practices them. A later qualification uses a fresh locked form.

Practice attempts never affect hiring readiness, enter the selection review queue, or reveal the protected scoring content. The shared environment creates continuity from hiring to onboarding. The data policy keeps the two uses separate.

Training is a real product, but the research claims are different. Meta-analyses of serious and digital games report positive average learning effects in studied educational and training contexts, often moderated by instructional support, repeated sessions, and design quality.[14, 15] Aincrad still has to measure whether a particular Ramp pack improves learning, retention, transfer, and later workplace performance.

The data-center economic case

The first commercial question is not whether the 3D scene looks good. It is whether the evidence changes an expensive decision.

The cleanest pilot begins in shadow mode. A design partner supplies known-skill technicians or candidates, domain experts, and an independent later outcome. Aincrad measures reviewer time, candidate completion, technical failures, agreement with experts, false-pass and false-fail patterns, and whether the simulation adds information beyond the existing process.

A current planning scenario uses 450 annual hires and four \$2,000 travel-week equivalents per hire. That is $450 \times 4 \times \$2,000 = \3.6 million in annual travel spend. If a later validated training and qualification program truly removes one trip for every hire, the gross addressable amount is \$900,000. If it removes one trip for 25 percent of the cohort, the amount is \$225,000. These are modeled scenarios, not realized savings, and the hiring-validation pilot does not itself prove a travel reduction.

The same discipline applies to time. Fewer field-engineer screening hours create capacity unless the company actually avoids cost or converts the time to billable work. Earlier billability creates revenue only when authorized work and demand exist. Travel, recruiting, field-engineer time, payroll, and earlier billability cannot be added together without checking overlap.

The current commercial discussion anchor is a bounded paid shadow pilot followed by a conditional production program. Setup, content, expert calibration, validation, integration, support, and risk drive

the price. Cloud rendering and replay are a small part of the cost because the simulation runs primarily on the candidate’s device.

The same backbone in other industries

The data-center example is deep because it is the first product. The other examples below are proposed Aincrad packs, not validated products.

Software companies

O*NET describes software work in terms of analyzing requirements, developing tests, modifying systems to correct errors, and evaluating changes against constraints.[34] A useful Aincrad episode would begin with a repository, an incident, logs, a service map, a ticket queue, a test environment, and a time limit.

The candidate might receive an intermittent checkout failure. The first obvious error is not the root cause. A deploy fixes one metric and worsens another. The episode records which logs were inspected, which hypothesis was tested, which change was made, how the candidate scoped risk, whether tests were added, whether the rollout was guarded, and how the incident was documented.

The correct environment may be a real code editor and container rather than a 3D world. The game is the changing system state. Aincrad adds value only if the episode measures more than an ordinary coding test or paid work sample at a reasonable candidate burden.

Game companies

O*NET’s current video-game-designer profile includes balancing gameplay, designing mechanics and challenges, integrating feedback, maintaining design documentation, creating test plans, and working with artists and technical staff.[35]

A game-studio pack could present a small live build with retention, economy, bug, and player-feedback signals. The candidate has a limited sprint budget. A change that improves short-term engagement can damage the economy or frustrate a different player segment. A designer must identify the design problem, propose a change, tune parameters, plan the test, and explain the trade-off. A producer must replan dependencies after an asset slips. A QA candidate must reproduce a fault, reduce the case, prioritize severity, and write a useful report.

This is a natural vertical because the work product is already a game. Aincrad still has to avoid confusing familiarity with the specific genre or engine for general design ability.

Investment banks

O*NET describes financial and investment analysis as valuation, financial modeling, analysis of company and market information, preparation of transaction materials, and communication of recommendations.[36] Current analyst job descriptions also emphasize modeling, valuation, due diligence, and transaction execution.[37] FINRA’s current Series 79 outline organizes entry-level investment-banking knowledge around collecting and evaluating data, underwriting and financing transactions, and mergers, acquisitions, tender offers, and restructuring. It is a licensing outline, not a work-sample design, but it supplies a narrower task taxonomy than a generic finance test.[38]

Recorded video is still part of some major-bank early-career processes. Goldman Sachs describes a first video interview and explicitly refers to its HireVue interview. Bank of America tells student applicants to become familiar with the HireVue platform.[39, 40] That supports a narrow claim: the screen exists. It does not prove that every bank uses it or that video causes worse hiring outcomes.

An investment-banking pack can replace part of the one-way screen with a controlled deal room. The candidate receives historical statements, management projections, debt terms, comparable companies, a draft model, and a short client request. The state changes: management revises revenue, a debt covenant becomes binding, or the purchase-price assumption moves.

The candidate has to find the model break, repair the schedules, choose and defend valuation methods, identify missing diligence, update the output, and write the senior-review note. Aincrad records the workbook states, cell lineage, source use, reconciliation checks, time allocation, revisions, and explanation.

The episode should use a spreadsheet because the spreadsheet is the job. Turning it into a cartoon trading floor would reduce fidelity.

Pack	Primary surface	Shared mechanics	Example evidence
Proving Ground	3D workbench, diagrams, tools, records	Sequence, diagnose, verify, stop, escalate	Safe state, tests, documentation, replay
Software incident	Repository, terminal, service map, tickets	Diagnose, test, prioritize, recover	Hypothesis path, code, tests, rollout choice
Game studio	Playable build, economy sheet, issue board	Balance, allocate, experiment, communicate	Change set, test plan, trade-off record
Investment banking	Spreadsheet, deal room, messages	Reconcile, model, prioritize, revise	Formula lineage, source use, review note

The common product is the event and evidence system. The surface changes with the work.

Platform architecture

The technical plan follows the evidence chain.

Control plane

The control plane manages organizations, users, candidates, roles, campaigns, consent, accommodations, packs, versions, rubrics, thresholds, reviewers, appeals, retention, and audit exports. It connects to applicant-tracking and learning systems. Hiring and training records are separate by default.

Runtime contract

Every runtime exposes stable object and action identifiers, authoritative state transitions, seeded variation, ordered events, state hashes, disconnect markers, and deterministic replay fixtures. The browser receives only the candidate-safe bundle. It never receives hidden scoring keys or protected answer routes.

The shared contract lets Aincrad support different surfaces. SpatialOps can run a physical workbench. CodeOps can run repositories and containers. DealOps can run spreadsheets and documents. A later DialogueOps can run multi-party conversations. They share identity, versioning, telemetry, evidence, review, and outcome linkage.

Scenario definition

A scenario defines objects, state, actions, constraints, documents, event injections, goals, variants, evidence-bearing transitions, allowed use, accessibility mappings, content ownership, and regression tests. It separates coaching content from protected assessment content.

Material edits create a new version. If an edit changes the work, difficulty, evidence, or interface, the new version loses inherited assessment status until reviewed.

Automated quality assurance

Agents can test reachability, broken states, trivial routes, impossible goals, variant extremes, scoring inconsistencies, unsafe content, and performance. They can generate candidate personas with different strategies and compare evidence profiles. These checks reduce production cost. They do not replace real participants, domain experts, accessibility testing, or outcome validation.

Security and integrity

Integrity is a risk system, not a promise that cheating disappears.

The platform uses equivalent forms, content exposure tracking, signed event logs, server-side rules, session hashes, identity controls, anomaly detection, protected practice content, and red-team testing. It records the confidence of the run and routes suspicious cases to review. It does not claim to infer cheating from one unusual mouse path.

AI assistants create a second integrity problem. Some real jobs permit AI, search, templates, and collaboration. The assessment should reproduce the actual policy. If the job allows an assistant, evaluate the candidate's use of it. If the job requires unaided execution, enforce and disclose that boundary, then treat evasion as an integrity signal rather than a secret psychological score.

Competitive landscape

The market is not empty. Aincrad sits between several categories.

- **Enterprise hiring platforms** such as HireVue combine interviews, assessments, job tryouts, science, integrations, and governance.[6, 7, 8]
- **Rich proprietary assessments** include McKinsey Solve, which McKinsey says reaches roughly 300,000 candidates a year, and Roblox's sustained internal game-based process, for which Roblox does not publicly report annual assessment volume.[21, 24]

- **Game-based assessment vendors** such as UnCircuit, The Talent Games, Harver, Arctic Shores, Criteria, and Equalture sell games or task-based measures.[41, 42, 43, 44]
- **Work-sample platforms** such as CodeSignal, Vervoe, Canditech, and Hatchways let employers test role-specific artifacts and skills.[45, 46, 47]
- **Simulation-authoring platforms** such as FacilityVR, Attensi, Warp VR, and Uptale help companies create and deploy interactive training.[48, 49, 50, 51]
- **Data-center and trades training** such as BILT with Salute, Interplay Learning, and Pixaera already covers guided procedures, simulations, assessment modes, and readiness reporting.[52, 53, 54]

No individual feature is the moat. “Game-based” exists. “AI-generated assessment” exists. “No-code 3D” exists. “Assessment mode” exists. “Every click becomes data” is not a defensible claim.

The possible moat is the maintained evidence system: the mapping from work to mechanics, the semantic trace, the version registry, comparable variants, calibrated reviewers, accessibility and governance controls, and the longitudinal link between simulation behavior and later field performance.

A direct HireVue comparison

Dimension	HireVue today	Aincrad thesis
Core enterprise value	End-to-end screening, interviewing, assessment, automation, and integration	Playable work episodes plus evidence and replay
Assessment surface	Video and AI interviews, psychometric games, virtual job tryouts, role assessments	Stateful job environments using the native work surface
Primary evidence	Structured responses, assessment scores, job-tryout measures	Semantic action trace, artifacts, state changes, and critical moments
Authoring	Guided AI Assessment Builder and science framework	Mechanic-based simulation compiler with trust levels
Validation	HireVue-published I-O validation and governance program	Pack and version registry that Aincrad still has to prove
Current advantage	Distribution, integrations, data, science, procurement trust	Potentially deeper job specificity and replayable process evidence
Failure mode	Generic content may remain distant from employer-specific work	Bespoke production and weak validity can destroy the platform model

The goal to “kill HireVue” becomes concrete here. Aincrad wins a use case when a buyer replaces the recorded screen or generic assessment with a playable job episode because it supplies incremental decision evidence that predicts independent later outcomes at acceptable cost, candidate burden, fairness, and operational complexity. Anything less is a design preference, not displacement.

Business model

The customer may create once, but the assessment runs every cohort. The recurring product is the evidence lifecycle.

1. A customer defines a role and creates or commissions a pack.
2. Every candidate or learner run uses delivery, identity, telemetry, scoring, review, and retention.

3. Exposed content and changed work create new variants and versions.
4. Outcome data updates the evidence and monitoring record.
5. Hiring can lead to separate ramp, qualification, promotion, and recertification use.
6. A successful role expands to another site, equipment family, language, or team.

The early model is managed, not self-serve fantasy economics:

- a paid job-analysis, content, implementation, and validation engagement;
- an annual organization, runtime, evidence, security, and integration fee;
- per candidate, learner, or expert-verified result;
- an annual maintained-pack fee for assessment-grade versions;
- optional expert review, calibration, localization, outcome studies, and premium integrations;
- later, revenue share with approved domain creators and verified pack owners.

A free or low-cost prototype tier can make creation legible. Consequential use should be paid because it requires accountable work.

The unit economics must include content labor, domain experts, validation, integration, reviewer time, candidate support, security, insurance, and maintenance. Rendering and storage are unlikely to be the main cost. If every customer requires a new engine, new art pipeline, new scoring system, and months of expert labor, Aincrad is a bespoke studio. The platform story becomes true only when mechanics, runtime, evidence, tests, and integrations are reused enough to reduce marginal delivery.

Go-to-market

Start with Proving Ground.

Data centers offer concrete work, expensive mistakes, visible safety gates, repeatable procedures, domain experts, and a hiring and ramp problem that managers can describe. The initial customer is a commissioning firm, critical-infrastructure service provider, operator, training pipeline, or staffing partner with enough volume and enough outcome access to run a serious shadow study.

The first sale is a bounded decision problem, not a platform license. One role. Two or three episodes. Named outcomes. Known participants. Expert access. A fixed pilot fee. Explicit exclusions. No unsupported promise that the pilot will save money on day one.

The expansion sequence is:

**one job works → one customer pays → outcomes support the score → variants reuse the engine
→ a second pack works → customers author safely → a governed ecosystem**

The second pack should remain adjacent, such as fiber, high-density rack commissioning, critical power, or incident response. The first new vertical should reuse a visible mechanic, perhaps diagnosis in software operations, before Aincrad tries to serve every profession.

Distribution can come through employers, training providers, staffing firms, equipment partners, and domain consultancies. A partner can teach the work. Aincrad can verify independent execution. The

clean data-center positioning is: guided platforms standardize how work is taught; Proving Ground standardizes the evidence that someone can perform it independently.

Validation plan

Every assessment-grade pack earns its claim in stages.

Job analysis and content evidence

Identify critical tasks, conditions, tools, errors, and competencies with incumbents and supervisors. Collect real artifacts. Write the claim and evidence model before finalizing the scene. Confirm that important content is present and irrelevant difficulty is absent.

Usability and nuisance variables

Test whether participants understand the interface. Measure prior gaming, device, language, accessibility needs, and technical failures. Provide practice. Remove navigation, reaction time, graphics, or reading demands that are not part of the job.

Known-groups study

Compare experienced and novice participants under a preregistered plan. Confirm expert review agreement and equivalent-form behavior. Known-groups separation is useful construct evidence. It is not proof of future job performance.

Shadow hiring

Candidates complete the episode, but the result does not affect the decision. Compare Aincrad with the existing process and independent later outcomes. Estimate score consistency or generalizability, criterion validity with confidence intervals, incremental validity over the existing process, predictive and measurement bias, subgroup selection rates with sample-size warnings, technical reliability, candidate reactions, completion, accommodation outcomes, and reviewer burden.[56, 57]

Bounded decision use

Only an approved version inside a documented role, population, and use can affect a decision. Keep human review, notice, accommodation, appeal, monitoring, and a kill switch. Continue field confirmation for safety-critical authorization.

Continuous monitoring

Track form exposure, score drift, outcome relationships, subgroup results, reviewer disagreement, complaints, incidents, support load, and changed work. Expire evidence when the job or instrument changes materially.

The Uniform Guidelines remain codified at 29 CFR Part 1607, and the EEOC’s 2007 technical-assistance document describes validation, adverse-impact review, and less-discriminatory alternatives. A June 9, 2026 Department of Justice Office of Legal Counsel opinion rejects major parts of that interpretation and calls the existing validation regulations unlawful. The opinion is not a court judgment and does not repeal Title VII or state and local requirements. Aincrad should preserve job analysis, validation, subgroup monitoring, and documentation as scientific and risk controls, with deployment-specific legal review.[55, 58, 59]

Rules we cannot ignore

This section is product-risk guidance, not legal advice.

- **Employment discrimination.** Intentional discrimination remains prohibited. Neutral selection procedures can also face disparate-impact challenges under Title VII’s statutory framework, but the current federal executive-branch interpretation is narrower than the still-published 2007 EEOC guidance. State and local law may impose independent requirements.[55, 59]
- **Disability and accommodation.** The interface must measure the job, not an irrelevant sensory, manual, communication, or cognitive barrier. Employers must provide reasonable accommodations where required.[29]
- **Automated-employment rules.** New York City’s AEDT law can require an annual bias audit, public results, and advance notice for covered tools. California regulations effective October 1, 2025 apply existing employment-discrimination law to automated decision systems and require records. Illinois AI employment provisions took effect January 1, 2026. Colorado requirements for covered employment decision technology begin January 1, 2027. Coverage depends on each law’s definitions.[60, 61, 62, 63]
- **High-risk AI.** The EU AI Act lists many AI systems used for recruitment, selection, and candidate evaluation as Annex III high risk, subject to scope exceptions. As of September 2026, the AI Omnibus delays those obligations until December 2, 2027. Aincrad should still design for risk management, data governance, documentation, logging, human oversight, accuracy, robustness, and cybersecurity.[64, 65]
- **Privacy.** Collect only data needed for the stated purpose. Set retention and deletion. Restrict reuse. Do not create a universal labor score or sell applicant telemetry.
- **Accessibility.** Follow WCAG 2.2, test with assistive technology and real users, and provide equivalent paths where the visual or motor interaction is not essential.[30]
- **Intellectual property.** Customer SOPs, equipment models, source code, financial data, and test content need clear rights, isolation, and export controls.

Governance belongs in the product. Aincrad needs candidate notices, accommodation workflows, human review, appeals, immutable versions, access logs, retention controls, audit exports, model cards, subgroup monitoring with sample-size warnings, complaint logs, and a kill switch.

What could go wrong

Risk	How Aincrad fails	Earliest honest test
Wrong construct	Score reflects gaming, device, navigation, reading, or language instead of work	Collect nuisance variables and compare alternative interfaces
No incremental validity	Replay looks persuasive but adds nothing beyond the current process	Shadow study with preregistered independent outcomes
False anti-cheat promise	Routes leak, proxies play, or AI agents perform the task	Red-team bundles, forms, identity, collusion, and exposure
Bespoke-services trap	Every customer needs a new engine and months of manual content	Measure reusable rules, assets, tests, and hours across deployments
UGC validity explosion	Unreviewed creators publish consequential tests under the brand	Gate hiring publication and keep unreviewed content non-consequential
Training contamination	Practice reveals scored routes or influences hiring readiness	Separate content, forms, data, permissions, and downstream calculations
Candidate resentment	The episode feels childish, opaque, invasive, long, or unpaid	Candidate research, completion, withdrawal, trust, and appeal data
Opaque scoring	A model issues confident results that cannot be reproduced	Re-run score derivation across versions and require evidence citations
False realism	A detailed 3D environment reproduces appearance but not important work	Domain task analysis and criterion-linked evidence review
Weak economics	Buyers want one export, not a maintained evidence service	Contract for annual minimum, usage, outcomes, and version maintenance

Hard no-go conditions are useful. Aincrad should stop or change the thesis if serious pilots show that the score adds no information, gaming or device variables explain the result, candidates reject the burden, a simpler work sample performs as well for far less cost, employers will not share outcomes, or delivery remains permanently bespoke.

MVP scope

The first Aincrad version does not need a universal game engine or a public marketplace.

It needs:

- input for role, claims, context, constraints, time, and source documents;
- five to seven observable mechanic primitives;
- diagnosis, resource-allocation, and sequencing shells;
- AI-assisted scenario configuration, narrative, assets, and parameter sets;
- browser delivery by link;
- semantic telemetry and deterministic replay;
- a reviewer timeline and evidence packet;
- prototype, training, shadow, and assessment publication levels;
- a pack and version registry;
- outcome import and basic monitoring;

- notice, accommodation, appeal, audit, and retention controls.

Do not build arbitrary scripting, a public Roblox-style marketplace, universal personality inference, VR as a requirement, an opaque LLM scorer, or a portable global talent score. Do not build the software, game-studio, and banking packs at the same time. Use them to pressure-test the architecture, then earn the second vertical.

Technical plan

This is the plan for the next 6 to 12 months.

Product and architecture

- Move the current Proving Ground runtime, evidence, review, and registry concepts under Aincrad.
- Define the runtime adapter contract and stable event vocabulary.
- Lock the separation between practice, shadow, assessment, qualification, and field confirmation.
- Remove protected scoring content from candidate bundles and close practice-to-readiness leakage paths.
- Add immutable engine, scenario, rubric, threshold, form, and result versions.
- Build the first constrained Studio on top of approved mechanics and parameters.
- Add outcome import, monitoring, accommodations, appeals, and audit packets.

Content

- Select one data-center role and two or three independent episodes.
- Complete external domain review for safety, terminology, equipment behavior, and scoring.
- Build equivalent variants without changing the intended claim.
- Add one adjacent pack only after the first study works.
- Write one design-only specification each for software, game development, and investment banking to test whether the mechanic layer travels.

Evidence

- Preregister the first known-groups and shadow studies.
- Calibrate reviewers and measure agreement.
- Collect prior-gaming, device, language, accessibility, and technical-failure variables.
- Define independent later outcomes before viewing the simulation results.
- Publish a version-specific assessment card with evidence, limitations, and allowed use.

Commercial

- Recruit one paid shadow design partner with an operational champion and outcome access.
- Price the pilot around bounded deliverables, not cloud usage.
- Track content labor, reviewer time, support, reuse, integration, and maintenance.
- Test annual runtime plus per-session economics before building a marketplace.
- Do not claim realized savings until the customer measures them.

What done looks like in 6 to 12 months

- Proving Ground operates as an Aincrad pack on the shared runtime and evidence model.
- One data-center role has a documented job analysis and at least two versioned episodes.
- A domain expert can correct and approve a scenario without editing engine code.
- The same mechanic can be reskinned across three role contexts without changing deterministic behavior.
- A paid partner completes a shadow study with independent later outcomes.
- Reviewers can explain every consequential result from the trace and rubric.
- Equivalent forms show acceptable behavior and exposure can be monitored.
- Candidate support, accommodations, appeals, retention, and audit exports work end to end.
- One customer agrees to recurring platform and usage economics.
- A second pack reuses most of the runtime, evidence, and authoring primitives.

Questions and answers

Is Aincrad just a game-based assessment company?

It is a work-simulation and evidence company. Games are the operating medium because they preserve state, accept actions, enforce rules, and record consequences. Some packs will look like 3D games. Others will look like terminals, spreadsheets, inboxes, diagrams, or live service consoles.

Why is this better than a structured interview?

Structured interviews remain one of the strongest established predictors of overall performance on average. Aincrad is better only when the episode reproduces important work that the interview cannot observe efficiently and local evidence shows useful incremental value. The simulation gives the interviewer something concrete to discuss.^[4]

Why not use a paid take-home project?

For some jobs, a paid take-home is the right answer. Aincrad matters when state changes, standardization, equivalent forms, time-bounded decisions, safety, replay, or high-volume delivery are important. If a simpler work sample produces equal or better evidence at lower burden, use it.

Can candidates game the game?

They can study, rehearse, collude, use proxies, inspect bundles, stream, automate, or ask AI for help. Aincrad reduces and measures those risks with variants, practice separation, server rules, logs, exposure monitoring, and review. It never promises an ungameable test.

Does more telemetry make the score more accurate?

No. Telemetry becomes useful only after a claim and evidence model explain why the event matters. Collecting everything can create privacy risk and spurious correlations. Aincrad keeps raw events for replay and promotes only defined evidence into the score.

Will AI build every game from one prompt?

AI can generate a strong first draft. High-stakes content still needs source material, domain correction, deterministic compilation, accessibility review, participant testing, and version-specific validation. The product should make that work faster and visible.

Why begin with data centers?

The work is procedural but not mindless. Safety, sequence, diagnosis, documentation, and escalation can be observed in a bounded environment. Errors are expensive. The current Proving Ground prototype already supplies the first runtime and content base.

What keeps this from becoming a consulting firm?

The mechanic library, scenario schema, runtime, evidence rules, QA agents, integrations, registry, and outcome system must be reused. Aincrad should track reuse and hours from the first contract. If marginal delivery does not fall, the honest business is a studio and should be priced that way.

What is the moat?

Not 3D and not the first generated game. The moat can become the accumulated map from work to mechanics, versioned evidence, comparable forms, reviewer calibration, outcome data, governance, integrations, and customer trust.

Does a simulation authorize someone to work on live equipment?

No. A simulation can support hiring, training, and qualification decisions inside a defined scope. Safety-critical authorization may still require supervised field confirmation, site-specific approval, and current credentials.

What we are asking for

- A data-center design partner willing to fund a bounded shadow pilot.
- Domain experts who can supply real tasks, artifacts, failure modes, and review time.
- Access to known-skill participants and an independent later work outcome.
- Industrial-organizational psychology, accessibility, employment-law, privacy, and security review.

- Software, game-studio, and investment-banking experts willing to pressure-test the reusable mechanic layer without pulling the team away from the first pack.

The bet is simple. A company should be able to describe the work and the evidence it needs, turn that into a playable experience, and test whether the resulting evidence improves the decision. Aincrad earns the right to replace a HireVue screen one job, one version, and one validated outcome at a time.

Sources

- [1] U.S. Office of Personnel Management. [Work Samples and Simulations](#).
- [2] Roth, P. L., Bobko, P., and McFarland, L. A. [A Meta-Analysis of Work Sample Test Validity: Updating and Integrating Some Classic Literature](#). *Personnel Psychology*, 2005.
- [3] Sackett, P. R., Zhang, C., Berry, C. M., and Lievens, F. [Revisiting Meta-Analytic Estimates of Validity in Personnel Selection: Addressing Systematic Overcorrection for Restriction of Range](#). *Journal of Applied Psychology*, 2022.
- [4] Sackett, P. R. et al. [Revisiting the Design of Selection Systems in Light of New Findings Regarding the Validity of Widely Used Predictors](#). *Industrial and Organizational Psychology*, 2023.
- [5] Hausknecht, J. P., Day, D. V., and Thomas, S. C. [Applicant Reactions to Selection Procedures: An Updated Model and Meta-Analysis](#). *Personnel Psychology*, 2004.
- [6] HireVue. [Game-Based Assessments](#). Current product page, accessed September 2026. Product and outcome statements are vendor-reported.
- [7] HireVue. [Virtual Job Tryout](#). Current product page, accessed September 2026. Product and outcome statements are vendor-reported.
- [8] HireVue. [HireVue Launches Assessment Builder to Scale Validated Skills Hiring Across Enterprises](#). February 19, 2026.
- [9] HireVue. [Our Science](#). Current product and governance page, accessed September 2026. Counts and outcome statements are vendor-reported.
- [10] HireVue. [AI Interviewer: Built for Trust at Enterprise Scale](#). Current product page, accessed September 2026. Capabilities and controls are vendor-reported.
- [11] Mislevy, R. J., Almond, R. G., and Lukas, J. F. [A Brief Introduction to Evidence-Centered Design](#). ETS Research Report RR-03-16, 2003.
- [12] Shute, V. J., and Ventura, M. [Stealth Assessment: Measuring and Supporting Learning in Video Games](#). MIT Press, 2013.
- [13] McKinsey and Company. [Innovating Recruiting Through Online Gaming](#). May 1, 2020.
- [14] Wouters, P., van Nimwegen, C., van Oostendorp, H., and van der Spek, E. D. [A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games](#). *Journal of Educational Psychology*, 2013.
- [15] Clark, D. B., Tanner-Smith, E. E., and Killingsworth, S. S. [Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis](#). *Review of Educational Research*, 2016.
- [16] Bipp, T., Wee, S., Walczok, M., and Hansal, L. [The Relationship Between Game-Related Assessment and Traditional Measures of Cognitive Ability: A Meta-Analysis](#). *Journal of Intelligence*, 2024.
- [17] Ramos-Villagrasa, P. J. et al. [Game-Related Assessments for Personnel Selection: A Systematic Review](#). 2022.
- [18] Ohlms, M. L., Hohner, A., and Melchers, K. G. [Is Test Takers’ Video Game Usage a Game Changer? Criterion-Related Validity and Potential Bias of a Cognitive Game-Based Assessment](#). *Applied Psychology*, 2025.

- [19] Ohlms, M. L., Melchers, K. G., Kanning, U. P., and Barends, A. J. [Game on, Faking off? Are Game-Based Assessments Less Susceptible to Faking Than Traditional Assessments?](#) *Journal of Business and Psychology*, 2025.
- [20] Kantar, R., McNulty, K., Snow, E., Emery, M., Wainess, R., and Doshi, V. [Constructing Cognitive Profiles for Simulation-Based Hiring Assessments](#). Proceedings of Educational Data Mining, 2018.
- [21] McKinsey and Company. [From Apprenticeship in Space to Selecting Microbes: Meet McKinsey’s Game-Based Innovation Lab](#). Current company profile, accessed September 2026.
- [22] Roblox Corporation. [2022 Annual Report, Imbellus acquisition note](#). SEC filing.
- [23] Roblox. [Accelerating Innovation: How Eight Startup Founders Are Continuing to Innovate at Roblox](#). October 2023.
- [24] Roblox Careers. [Early Careers: The Roblox Assessment](#). Current careers page, accessed September 2026.
- [25] Roblox. [Fair Play: Roblox’s Game-Based Talent Assessment](#). July 29, 2025.
- [26] Roblox Corporation. [2025 Annual Report on Form 10-K](#). Filed February 11, 2026.
- [27] Roblox Creator Hub. [Roblox Studio](#). Current documentation, accessed September 2026.
- [28] Emery, M. et al. [Identifying Off-Task Users in a Large-Scale, Game-Based Practice Assessment](#). Proceedings of Educational Data Mining, 2024.
- [29] U.S. Department of Justice, Civil Rights Division. [Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring](#). May 12, 2022.
- [30] World Wide Web Consortium. [Web Content Accessibility Guidelines 2.2](#). W3C Recommendation.
- [31] NVIDIA. [DGX GB Rack Scale Systems User Guide](#). Current documentation, accessed September 2026.
- [32] NVIDIA. [GB200 and GB300 Deployment Summary Validation Checklist](#). Current documentation, accessed September 2026.
- [33] Microsoft. [Data Center Technicians](#). Current careers description, accessed September 2026.
- [34] O*NET OnLine. [Software Developers, 15-1252.00](#). Updated occupational profile, accessed September 2026.
- [35] O*NET OnLine. [Video Game Designers, 15-1255.01](#). Updated occupational profile, accessed September 2026.
- [36] O*NET OnLine. [Financial and Investment Analysts, 13-2051.00](#). Updated occupational profile, accessed September 2026.
- [37] CFA Institute job board document. [Analyst, Investment Banking Group: responsibilities and requirements](#). Accessed September 2026.
- [38] Financial Industry Regulatory Authority. [Investment Banking Representative Qualification Examination, Series 79, Content Outline](#). Current outline, accessed September 2026.
- [39] Goldman Sachs. [Prepare for Your Interview](#). Current student careers page, accessed September 2026.
- [40] Bank of America. [Student Application Process](#). Current student careers page, accessed September 2026.
- [41] UnCircuit. [Game-Based Hiring Assessments](#). Current vendor page, accessed September 2026.
- [42] Harver. [Gamified Assessments](#). Current vendor page, accessed September 2026.
- [43] Arctic Shores. [What Is a Task-Based Assessment?](#) Current help documentation, accessed September 2026.
- [44] Criteria. [Game-Based Assessments](#). Current vendor page, accessed September 2026.
- [45] CodeSignal. [Assessment Generation and Agentic Capabilities](#). Current vendor announcement, accessed September 2026.
- [46] Vervoe. [Assessment Builder](#). Current vendor page, accessed September 2026.
- [47] Canditech. [AI Features in Canditech](#). Current vendor documentation, accessed September 2026.

- [48] FacilityVR. [AI-Assisted Industrial Simulation Authoring and Delivery](#). Current vendor page, accessed September 2026.
- [49] Attensi. [Attensi Creator](#). Current vendor page, accessed September 2026.
- [50] Warp VR. [Scenario Authoring and Immersive Learning Platform](#). Current vendor page, accessed September 2026.
- [51] Uptale. [Immersive Learning Platform Features](#). Current vendor page, accessed September 2026.
- [52] BILT. [BILT and Salute Partner to Scale Data Center Training and Operations](#). August 27, 2026.
- [53] Interplay Learning. [Navigating Assessment Mode Simulations](#). Current help documentation, accessed September 2026.
- [54] Pixaera. [Immersive Training and Readiness Platform](#). Current vendor page, accessed September 2026.
- [55] U.S. Equal Employment Opportunity Commission. [Employment Tests and Selection Procedures](#). Technical assistance document.
- [56] Society for Industrial and Organizational Psychology. [Principles for the Validation and Use of Personnel Selection Procedures](#). 2018.
- [57] Society for Industrial and Organizational Psychology. [Considerations and Recommendations for the Validation and Use of AI-Based Assessments for Employee Selection](#). January 2023.
- [58] Electronic Code of Federal Regulations. [29 CFR Part 1607, Uniform Guidelines on Employee Selection Procedures](#).
- [59] U.S. Department of Justice, Office of Legal Counsel. [Constitutionality of Disparate-Impact Liability Under Title VII](#). June 9, 2026.
- [60] New York City Department of Consumer and Worker Protection. [Automated Employment Decision Tools](#).
- [61] California Civil Rights Department. [Civil Rights Council Secures Approval for Regulations to Protect Against Employment Discrimination Related to Artificial Intelligence](#). June 30, 2025.
- [62] Illinois General Assembly. [Public Act 103-0804](#). Effective January 1, 2026.
- [63] Colorado General Assembly. [SB26-189, Automated Decision-Making Technology](#). Signed May 14, 2026; employment provisions begin January 1, 2027.
- [64] European Union. [Consolidated Regulation \(EU\) 2024/1689, Annex III](#). Consolidated July 27, 2026.
- [65] European Commission. [AI Omnibus Enters into Force](#). July 27, 2026.