Structured Work Samples and Interviews as Alternatives to Degree Screens Under AI-Contaminated Assessment

Lauren Mercer
Simon Fraser University
Canada
lauren.mercer@sfu.ca

Maya Kalyani Reed
University of Missouri, Columbia
United States
maya.kalyani.reed@umsystem.edu

Hoang Justin Tran
Université Grenoble Alpes
France
hoang.justin.tran@etu.univ-grenoble-alpes.fr

Alexander J. Garnier
Concordia University
Canada
alexander.j.garnier@concordia.ca

ABSTRACT

Generative AI complicates the inference that assessed academic work provides evidence of individual competence, but removing degree requirements may shift selection toward discretionary interviews, referrals, or network advantage. We develop a stylized simulation of screening for non-licensed occupations in which employers choose among degree-only, interview-only, degree-plus-interview, degree-plus-work-sample, work-sample-only, direct-hybrid, all-signal augmented-hybrid, and network-fallback regimes. Across 600 design scenarios with 5,000 synthetic applicants each, signal validity, AI contamination, degree costs, direct-assessment bias, direct-signal dependence, referral inequality, and hire rates vary over literature-informed ranges. The direct-hybrid regime combining work samples and structured interviews maximized a performance-equity objective in 40.5 percent of scenarios, followed by the all-signal augmented hybrid in 26.7 percent and work-sample-only screening in 21.2 percent. The direct hybrid exceeded degree-plus-interview screening on the objective in 92.7 percent of scenarios and degree-plus-work-sample screening in 83.3 percent, but exceeded the all-signal augmented hybrid in only 60.2 percent. Sensitivity checks show that the direct-hybrid advantage largely disappears when direct and degree signals have matched validity, high dependence, and no modeled group penalties. The model supports a conditional claim: degree-free reform depends on structured, job-relevant, accessible evidence rather than informal networks or costly, biased, duplicative assessments.

KEYWORDS

AI education, assessment, scholarly integrity

1 INTRODUCTION

Learning technologies increasingly mediate the production, verification, and aggregation of evidence about what learners can do. Generative AI makes this evidence problem more consequential because polished academic artifacts may no longer cleanly support claims about individual competence, while abandoning degrees may amplify opaque referrals or discretionary selection. This paper treats credential reform as a screening-design problem rather than a binary choice between degrees and no degrees. Using a literature-informed simulation of non-licensed occupational hiring, it compares degree-based, direct-assessment, hybrid, all-signal, and network-fallback regimes under shared performance and equity criteria. The contribution is a transparent model showing when structured work samples and interviews can improve access and prediction, and when their advantage disappears because direct assessments duplicate degree information, become costly or biased, or are displaced by informal networks.

Degrees do more than record course completion. In many labor markets they organize access to interviews, licenses, professional tracks, salaries, migration routes, and social status. The degree is therefore a governance device as well as an educational award: it compresses years of learning, selection, assessment, institutional reputation, and social opportunity into a signal that employers and regulators can use at scale. Gorelik [19] recently argued that generative artificial intelligence (AI) makes this arrangement less stable because assessed academic work may no longer support the inference from submitted artifact to individual competence that degree credentialing requires. That claim raises a practical question for higher education and employers: if the degree signal weakens, what should replace it?

The answer cannot be that credentials should simply disappear. Employers already combine many signals, including degrees, interviews, work samples, portfolios, prior experience, referrals, probation, and informal reputation. Some of these are more closely tied to job tasks than academic credentials, but they are not automatically fair or valid. A work sample can be expensive to prepare, an interview can become discretionary, a portfolio can be coached, and referrals can reward network access. Reducing degree requirements may improve access for capable candidates who face high educational costs, yet it may also shift selection toward less visible and less contestable forms of advantage.

This paper treats degree-free credentialing as a screening-design problem. The policy question is not whether degrees are good or bad in the abstract. It is whether a particular signal portfolio improves prediction and access under plausible assumptions about AI contamination, selection-method validity, direct-assessment bias, referral inequality, and degree-acquisition costs. The question matters because a university can respond to generative AI in several ways. It can tighten surveillance and keep the degree as the central signal; it can move assessment into oral, supervised, or authentic demonstrations; it can create modular evidence of competence; or it can withdraw from mass

credentialing and let employers build their own selection systems. These choices have different implications for validity and equity.

Assessment theory clarifies the stakes. Validity concerns the interpretation and use of scores, not only the technical reliability of a marking procedure [22]. Generative AI complicates that interpretation when polished submitted work may reflect prompt engineering, external assistance, or model output as much as the student's internalized competence. Empirical studies of AI-authored university work and detection tools show that authorship is difficult to establish in unsupervised settings [34, 39, 42]. At the same time, highly controlled assessment can narrow the construct being assessed and can introduce new access and fairness concerns, as work on automated proctoring illustrates [43]. The problem is therefore not only misconduct. It is whether the assessment evidence warrants the credentialing inference in the first place [8, 10].

Labor-market theory adds a second constraint. Education can operate as a productive investment, a signal, a sorting mechanism, or a mixture of these functions [3, 23, 36]. Even if generative AI weakens one source of degree validity, the degree may still carry information through admissions, persistence, institutional sorting, socialization, and observed completion. Conversely, a direct assessment may be more job-related than a degree and still be biased or expensive. A useful reform must therefore compare complete screening regimes rather than a degree signal against an unspecified alternative.

We construct a literature-informed simulation model for that comparison. The model represents a pool of synthetic applicants to a non-licensed occupation. Each applicant has latent competence, group membership used only to measure inequality, a realized performance outcome, and four possible screening signals: a degree signal, a work-sample signal, a structured-interview signal, and a network signal. Across 600 design scenarios, employers choose under eight regimes: degree-only, interview-only, degree-plus-interview, degree-plus-work-sample, work-sample-only, direct-hybrid, all-signal augmented-hybrid, and network-fallback. The model evaluates hired performance, representation, false negatives among high-competence disadvantaged-group applicants, and a transparent performance-equity objective.

The results point to a conditional middle position. Degree-free screening is attractive in the model when degree removal is paired with structured direct evidence and when the degree signal is degraded by AI-era assessment risk. It is much less attractive when direct and degree signals are assigned similar validity, when direct signals are highly dependent, or when degree removal becomes a pathway to referrals or discretionary judgments. The degree-free question is therefore less a question of institutional identity than of infrastructure: can universities, employers, professional associations, or public agencies produce evidence of competence that is more interpretable, accessible, and externally checkable than the degree signal it replaces?

This paper makes four contributions:

- It provides a transparent screening model that compares degree-only, direct-assessment, degree-plus-direct-assessment, all-signal, and network-fallback regimes under a shared objective.
- It calibrates the model to broad literature-informed ranges for AI contamination, signal validity, unequal degree costs, direct-assessment bias, direct-signal dependence, referral inequality, and hire rates.
- It reports reproducible performance and access results showing when structured direct assessment is stronger than degree-based screening and when its advantage disappears.
- It reframes degree-free higher education as a problem of accountable evidence design rather than credential withdrawal.

2 RELATED WORK

2.1 Degrees, signaling, and employer learning

The degree's labor-market role is commonly analyzed through signaling and employer-learning frameworks. In Spence's model, education can convey information about productivity when employers cannot observe productivity directly [36]. Employer-learning studies add dynamics: schooling is especially informative early in employment, and firms update their beliefs as they observe performance [3, 23]. Sheepskin-effect research similarly shows that credential completion can carry value beyond accumulated years of schooling [20, 21]. These literatures do not imply that degrees are always valid measures of competence; they explain why a costly, institutionally recognized signal can persist even when direct productivity is initially hidden.

Recent work on non-degree candidates and skills-based hiring challenges the breadth of degree requirements. Blair and colleagues argue that many candidates without bachelor's degrees possess valuable skills that degree filters can obscure [5]. Job-posting studies show movement toward skills language in some sectors, particularly for technology-adjacent and green roles, but posting language is an incomplete measure of actual hiring [18, 28]. This literature motivates the access side of the model: a degree signal may be informative and still exclude high-competence candidates when opportunity costs are unequal.

2.2 Generative AI and assessment validity

Generative AI puts pressure on academic assessment because it changes the relationship between an assessed artifact and the person being certified. Alexander and Belloni describe this pressure as a challenge to both the human-capital and signaling roles of higher education [2]. Dawson and colleagues argue that validity matters more than policing misconduct because the central question is whether an assessment continues to support its intended use [10]. Corbin and colleagues emphasize structural assessment redesign rather than only statements of acceptable use [8].

The assessment challenge is dynamic because model capabilities are not tied to a single form of supervision or artifact. Work on uncertainty-guided learning from partly labeled data, novel-object captioning, language-grounded visual understanding, model reprogramming across tasks, and in-context semi-supervised adaptation illustrates how learned representations can be extended with sparse labels, new referents, or new demonstrations [7, 11–13, 24]. For assessment design, this capacity for adaptation makes the evidentiary conditions of performance—supervision,

oral defense, process visibility, and task authenticity—more durable anchors than any supposedly protected artifact format. That design principle complements validity-centered accounts of assessment, evidence of fairness problems in automated proctoring, and calls for structural redesign [8, 22, 43].

Empirical studies show why the issue is serious. AI-written university answers can receive high marks while escaping detection in blind marking [34]. AI-authored academic essays can be difficult to distinguish from human-authored work [42]. Detection tools change as models change and can produce unreliable classifications [39]. More restrictive alternatives can create their own validity and fairness problems: automated proctoring systems have shown disparities by skin tone, race, and sex [43]. These findings do not settle what institutions should do, but they justify modeling the degree signal as degraded under AI contamination.

2.3 Direct assessment, alternative credentials, and selection risk

Personnel-selection research gives reasons to take direct evidence seriously. Work samples and structured interviews have substantial predictive validity when job-related, standardized, and consistently scored [26, 31, 35]. Situational judgment tests and multimethod selection batteries can also improve prediction when designed carefully [25, 29, 32, 37]. These findings motivate the direct-hybrid regime in the model: a job-relevant work sample supplies task evidence, while a structured interview captures additional information in a standardized format.

The same literature warns against treating direct assessment as automatically equitable. Validity and bias jointly shape selection errors and adverse impact [1]. Discrimination can intensify after the callback stage, where interviews and discretionary judgments become more salient [30]. Algorithmic hiring evaluations also show that formal measurement systems require careful validation, not only technical sophistication [4, 33]. For this reason the model includes direct-assessment bias as a parameter rather than assuming that job-proximal evidence is neutral.

Alternative credentials face a related trade-off. Micro-credentials and digital badges can make specific skills visible, but reviews report uneven quality assurance, uncertain employer recognition, and risks of fragmenting the credential landscape [16, 27, 38, 41]. Competency-based education similarly raises a tension between measurable competences and broader forms of knowledge that are difficult to reduce to discrete tasks [15, 40]. The model therefore treats direct evidence as potentially valuable but institutionally demanding.

2.4 Attribution to the motivating proposal

This article takes from Gorelik's degree-free proposal the central institutional question: whether higher education should continue to rely on degrees as a mass credentialing device when generative AI weakens the assessment-to-competence inference [19]. It also uses the distinction between education as formation and degrees as occupational signals. The contribution here is a formal screening comparison with explicit levers, outcomes, and sensitivity checks. The model does not assume that degree-free systems are better; it identifies assumptions under which structured alternatives outperform degree-based screening and assumptions under which all-signal or degree-inclusive regimes remain competitive.

Referral and licensing literatures define two important edges of the comparison. Referrals can reduce search frictions and sometimes convey information, but they can also reproduce unequal network access [6, 9, 14]. Occupational licensing can protect consumers in safety-critical fields while creating barriers to entry and mobility [17]. The present model therefore focuses on non-licensed settings and includes network fallback as a risk scenario rather than a desirable design.

3 METHODS

3.1 Scope and design logic

We study screening for non-licensed, non-safety-critical occupations in which employers could plausibly replace broad degree requirements with structured evidence of competence. The model is not a causal estimate of any existing labor market, and the 5,000 synthetic applicants in each run should not be read as an empirical sample size. It is a transparent decision model designed to compare signal portfolios under assumptions drawn from assessment, personnel-selection, employer-learning, skills-based-hiring, referral, and alternative-credential literatures.

This scope is deliberate. A direct field evaluation would require linked data on applicant pools, degree requirements, assessment scores, hiring decisions, protected or socioeconomic characteristics, subsequent job performance, and implementation costs. Such data are rarely public and would raise privacy and access questions beyond this study. The simulation instead asks a narrower question: across specified ranges for signal validity, bias, signal dependence, and access costs, which screening regimes are attractive within the model and which depend on narrow assumptions?

The model has five design principles. First, the degree has baseline validity before AI contamination, because degrees may contain information through admissions, persistence, institutional sorting, and prior assessment. Second, AI contamination reduces the degree's effective validity, because unsupervised academic artifacts may become less informative about individual competence. Third, work samples and structured interviews are modeled separately because they differ in cost, structure, and validity. Fourth, direct-assessment signals are allowed to have correlated residual errors, so the direct hybrid is not automatically given two independent predictors. Fifth, direct-assessment bias and referral inequality are explicit, because removing degrees can shift selection toward other forms of exclusion.

3.2 Applicant pool and signals

Each simulated applicant has a latent competence score, a binary group indicator, and a realized job-performance outcome. Group B represents a disadvantaged group that may face higher degree-acquisition costs and weaker referral access. The binary label is a simplification used only to measure representation and false-negative patterns; it is not a recommendation for real hiring systems to use protected characteristics in selection.

Each scenario generates four signals. The degree signal is a noisy proxy for competence whose effective validity declines as

AI contamination increases. The work-sample signal represents a structured, job-relevant task demonstration. The structured-interview signal represents standardized behavioral or oral assessment. The network signal represents referrals and informal reputation. Work-sample and interview errors share a scenario-level residual-correlation parameter from 0.00 to 0.75, allowing the two direct assessments to overlap rather than contribute fully independent information.

Eight screening regimes combine these signals. Degree-only uses the degree signal alone. Interview-only and work-sample-only use one direct signal. Degree-plus-interview and degree-plus-work-sample combine a degree with one direct signal. Direct-hybrid combines work samples and structured interviews without a degree signal. Augmented-hybrid combines degree, work-sample, and interview evidence. Network-fallback combines work samples with referrals and represents the risk that degree removal leads employers toward informal access rather than structured public evidence.

3.3 Outcomes and policy objective

The employer selects the top-ranked applicants under each regime. We evaluate four outcomes: mean performance among hires, group-B share among hires, false-negative rate for high-competence group-B applicants, and selection loss relative to an oracle that observes realized performance. High-competence group-B applicants are those above the 75th percentile of latent competence.

The primary performance-equity objective is:

$$U = \overline{y}_{\text{hired}} - 0.50|s_B - s_B^{\text{pool}}| - 0.50\,\text{FNR}_{B,\text{high}},$$

where $\overline{y}_{\text{hired}}$ is mean realized performance, s_B is group B's share among hires, and $\text{FNR}_{B,\text{high}}$ is the false-negative rate among high-competence group-B applicants.

The objective is intentionally simple. It does not claim to be the single correct welfare function. Its purpose is to prevent the comparison from treating performance and access as separate narratives. A regime that hires strong performers while excluding many high-competence disadvantaged-group candidates should not be considered unambiguously superior, and a regime that improves representation while greatly weakening prediction also needs to be visible as a trade-off.

3.4 Parameter calibration

Table **??** summarizes the model ranges. The ranges are broad because the simulation is not fitted to one occupation. They are meant to create a transparent design distribution from the literature rather than estimate population parameters. The additional structural sensitivity checks below deliberately move beyond these ranges to test whether the main pattern is built into the assumed ordering of direct and degree signals.

3.5 Simulation protocol

The main run samples 600 parameter scenarios using reproducible pseudo-random stratified sampling and generates 5,000 applicants per scenario. The random seed is 2026040391. Signals are standardized within scenarios so that regime weights are interpretable. Degree-plus-interview, degree-plus-work-sample, direct-hybrid, and network-fallback weight their two component signals equally. Augmented-hybrid weights degree, work-sample, and interview evidence equally. Hire rates vary from 0.10 to 0.30.

We report averages across scenarios, paired contrasts within the same scenario, and 2.5th–97.5th percentiles across the scenario design distribution. These percentiles describe variation under the specified model; they are not population confidence intervals. Robustness checks repeat the simulation with 240 scenarios per configuration, applicant pool sizes of 2,000 and 10,000, and fairness weights of 0.25, 0.50, and 1.00. Structural checks further examine matched direct and degree validity, high direct-signal dependence, no AI contamination, no modeled group penalties, and a combined case with matched validity, high dependence, and no group penalties. All computations are regenerated by `code/run.sh` and written under `results/`.

4 RESULTS

4.1 No single regime dominates after stronger baselines

Figure **??** shows that direct-hybrid screening maximized the performance-equity objective in 40.5% of the 600 design scenarios. This is the largest win share, but it is not dominance. The all-signal augmented hybrid won in 26.7% of scenarios, work-sample-only screening in 21.2%, interview-only screening in 5.7%, degree-plus-work-sample screening in 3.5%, degree-plus-interview screening in 1.3%, and network fallback in 1.2%. Degree-only screening never won the main objective. The broader baseline set therefore changes the interpretation: structured direct evidence is often attractive, but a degree-inclusive all-signal regime remains competitive when the degree signal retains value.

Table **??** reports mean outcomes by regime. Direct-hybrid screening had the highest mean objective, 0.239, followed by augmented-hybrid screening at 0.225 and work-sample-only screening at 0.189. It did not have the highest mean hired performance: augmented-hybrid screening averaged 0.580, compared with 0.575 for direct-hybrid screening. The paired direct-hybrid objective advantage averaged 0.158 over degree-plus-interview screening and 0.108 over degree-plus-work-sample screening. Against the augmented hybrid, however, the direct-hybrid objective advantage averaged only 0.014 and its mean performance was lower by 0.005. Across the scenario distribution, the direct-minus-augmented objective ranged from −0.103 at the 2.5th percentile to 0.119 at the 97.5th percentile, confirming that the all-signal comparison is close.

4.2 Access improves most when direct evidence is structured and open

Degree-only screening had a mean false-negative rate of 0.767 for high-competence group-B applicants. Direct-hybrid screening reduced that rate to 0.626, work-sample-only screening to 0.645, and augmented-hybrid screening to 0.642. The comparison supports a specific interpretation of degree-free reform: access gains arise when degree removal is paired with structured direct evidence, not when the credential is simply removed.

Group-B representation also illustrates the difference between direct evidence and informal fallback. Interview-only screening produced the highest mean group-B hire share, 0.374, but with weaker mean performance than the direct hybrid. Direct-hybrid and work-sample-only screening produced mean group-B shares of 0.356 and 0.351. Network fallback produced the lowest group-B share, 0.311, despite removing the degree signal. The reason is built into the regime: referrals can carry information, but unequal network access penalizes disadvantaged-group applicants. A labor market can therefore become less degree-centered while remaining exclusionary.

4.3 AI contamination and direct-assessment bias define the main trade-off

Figure 2 plots the paired difference between direct-hybrid and degree-plus-interview objectives against AI contamination, with direct-assessment bias shown by color. The direct-hybrid advantage generally grows as AI contamination rises because the degree component of degree-plus-interview screening becomes less informative. Higher direct-assessment bias reduces the advantage, especially when AI contamination is low and degree validity remains high.

Stress cells make the same pattern concrete. In the low-AI-contamination, high-degree-validity subset, direct-hybrid screening beat degree-plus-interview screening in only 36.4% of 11 scenarios, and the mean direct-minus-degree-plus-interview objective difference was slightly negative, −0.006. In the high-AI-contamination, low-direct-bias subset, it did so in 100.0% of 21 scenarios. In the low-AI-contamination, high-direct-bias subset, it did so in 81.0% of 21 scenarios. These cells are model-generated stress tests, not estimates of sector frequencies, but they identify the assumptions under which degree-based screening remains most competitive.

4.4 Network fallback is the wrong lesson from degree-free reform

Network fallback won the main objective in only 1.2% of scenarios. Its mean hired performance, 0.481, exceeded degree-only screening but fell below direct-hybrid, augmented-hybrid, work-sample-only, and degree-plus-work-sample screening. Its mean objective, 0.089, also fell below direct-hybrid screening at 0.239 and work-sample-only screening at 0.189. In the high-direct-bias, weak-network-exclusion stress cell, network fallback beat direct-hybrid screening in 11.8% of 17 scenarios, which shows that network fallback can become less unattractive when direct tools are strongly biased. It remains a policy risk rather than a recommended design.

The implication is institutional rather than merely technical. If universities reduce degree credentialing, they still need to help produce portable evidence that employers can interpret. If employers remove degree requirements, they need structured selection tools that are open to candidates without privileged networks. Without those substitutes, degree-free language can mask a shift from formal credentials to informal gates.

4.5 Robustness and structural sensitivity

Sensitivity checks over applicant pool size and fairness weight support the paired direct-versus-degree comparison but not a claim of direct-hybrid dominance. With applicant pools of 2,000 or 10,000 and fairness weights of 0.25, 0.50, or 1.00, direct-hybrid win shares ranged from 34.6% to 41.7%. Its objective exceeded degree-plus-interview screening in 89.2% to 92.9% of scenarios and degree-plus-work-sample screening in 80.0% to 86.3%. Its objective exceeded augmented-hybrid screening in only 52.9% to 67.9% of scenarios. Work-sample-only win shares ranged from 18.8% to 24.6%.

Table 3 reports structural checks that address the main design fragilities. When direct and degree signals are assigned matched effective validity, the direct-hybrid win share falls to 5.4%, and the augmented hybrid wins 46.3% of scenarios. When direct-signal residual correlation is fixed at 0.90, work-sample-only screening wins most often, and the direct hybrid wins only 7.1%. When AI contamination is set to zero, the augmented hybrid wins 52.5% of scenarios. In the most conservative check—matched validity, high direct-signal dependence, no AI contamination, and no modeled group penalties—the direct hybrid does not win any scenario. These checks show that the direct-hybrid result is a conditional implication of the assumed AI-era degree-validity loss and the modeled advantages of structured direct evidence, not a general theorem about degree-free screening.

Figure 3 summarizes the performance-access trade-off. Direct-hybrid screening combines relatively strong mean performance with a lower false-negative rate for high-competence group-B applicants than most degree-inclusive alternatives. Augmented-hybrid screening has slightly higher mean performance but a higher false-negative rate. Work-sample-only screening improves access but gives up prediction relative to the direct hybrid. Degree-only screening performs poorly on both the objective and the high-competence false-negative measure under the modeled AI-era assumptions.

5 DISCUSSION

The model supports the concern that generative AI can weaken degree credentialing, but it also narrows the reform agenda. The strongest lesson is not that degrees should be removed everywhere. It is that degree removal is most defensible when institutions can provide structured, job-relevant, and accessible evidence of competence. This matters because one possible reading of degree-free higher education is that universities should stop producing occupational evidence. The simulation points in the opposite direction: if the degree signal becomes less credible, institutions need better evidence, not less evidence.

The result also clarifies the role of equity. Degree requirements can exclude high-competence candidates who face higher costs of formal education. Removing those requirements can improve access, but only if the substitute selection process is itself accessible and consistently scored. Network fallback shows the failure mode. It removes the formal credential but gives informal access a larger role, producing the lowest disadvantaged-group representation among the regimes. A degree-free system that depends on referrals may be less transparent than the degree system it replaces.

For higher education, the practical implication is to separate learning environments from high-stakes sorting more carefully. In fields where broad degrees are becoming weak signals, universities might still provide formation, community, intellectual development, and access to structured demonstrations of competence. These demonstrations could include externally interpretable work samples, oral defenses, supervised tasks, portfolio defenses, apprenticeships, or modular assessments tied to professional practice. The relevant feature is not whether the evidence is called a degree, certificate, badge, or portfolio; it is whether the evidence supports the inference being made from it.

For employers, the model argues against treating degree removal as an end in itself. Dropping a degree requirement without investing in selection design can increase dependence on referrals, prestige proxies, automated filters, or unstructured interviews. The direct-hybrid regime is stronger than many degree-based comparators when the degree signal is degraded, but the augmented hybrid remains competitive when degree information retains value. A work sample provides a closer view of job-relevant competence; a structured interview adds information while limiting arbitrary variation; and an explicit performance-equity objective keeps access visible.

The same logic applies to applicants. A fairer credentialing system should not require candidates to guess which hidden signals matter or to purchase access to informal gatekeepers. It should tell candidates what competence is being assessed, how evidence will be scored, how accommodations and appeals work, and how outcomes will be monitored after hiring. Degree-free governance is therefore a public information problem as much as an assessment problem: selection evidence has to be understandable to learners, educators, employers, and regulators.

The model also suggests where empirical work should go next. Degree-free reform should be most promising where job tasks can be sampled authentically, where scoring rubrics can be made transparent, where applicants do not need unpaid time or elite networks to demonstrate competence, and where employers can validate the signals against later performance. Caution is warranted where performance depends on long-term formation that cannot be observed in short assessments, where tasks are hard to simulate, where direct assessments are expensive, where direct and degree signals are measuring nearly the same thing, or where informal networks dominate entry-level access.

Finally, AI use should not always be treated as contamination. In many occupations, competent performance will include the ability to use AI systems well. The model's AI-contamination parameter represents a narrower problem: assessment artifacts become less informative when they are used to infer unaided individual competence. In fields where AI-mediated work is part of the construct, assessment should measure that competence directly. The broader principle is alignment between the evidence collected and the competence inferred.

6 LIMITATIONS

This study is a formal exploration, not a causal estimate. Synthetic applicants are not real employees, and the parameter ranges are calibrated to broad evidence streams rather than fitted to linked data on applicant pools, assessments, hiring decisions, subgroup access, and later performance. The scenario distribution is an author-designed grid of assumptions, so the reported percentages are frequencies inside that design rather than estimates of sector prevalence. The model reduces complex social categories to a binary group indicator and compresses many forms of disadvantage into degree costs, direct-assessment bias, and referral access. This simplification is useful for transparent comparison but inadequate for applied policy without field-specific evidence.

The model still gives direct assessment a favorable structure in the main design: work samples and interviews have broad validity ranges drawn from personnel-selection research, and implementation costs are not subtracted from the objective. In practice, implementation quality varies. A work sample can be authentic or performative, a structured interview can drift into a discretionary conversation, and an AI-scored portfolio can import proxy discrimination. The simulation includes bias and signal-dependence parameters, but it cannot capture every institutional failure mode.

The paper also does not estimate transition costs or optimize screening portfolios. Degree-free systems would need governance for assessment access, disability accommodation, appeal, privacy, portability, fraud prevention, employer interpretation, and legal compliance. Safety-critical and licensed fields require separate analysis because degrees may be embedded in accreditation, licensure, liability, and public-risk governance. The missing empirical validation remains the central limitation: real hiring-funnel data could overturn the model if degree signals remain strong, direct assessments are costly or biased, or applicants respond strategically.

7 CONCLUSION

Structured direct assessment can outperform degree screening in a credentialing model when degree validity is weakened by AI-era assessment risk and when direct evidence is implemented as structured, accessible measurement rather than informal discretion. Across 600 design scenarios, the direct-hybrid regime won the main performance-equity objective most often, but only in 40.5% of cases; the all-signal augmented hybrid and work-sample-only regimes were also frequent winners. The direct hybrid outperformed degree-plus-interview screening on the objective in 92.7% of scenarios and degree-plus-work-sample screening in 83.3%, but it outperformed the augmented hybrid in only 60.2%. The same model warns that degree-free reform can fail if employers default to informal networks or biased direct assessment, and that degree-inclusive regimes remain competitive when degree information retains value. A credible post-degree credentialing system should therefore be built around transparent, job-relevant, accessible, and consistently scored evidence of competence, not around degree removal alone.

REFERENCES

[1] Herman Aguinis and Marlene A. Smith. 2007. Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology* 60, 1 (2007), 165–199. https://doi.org/10.1111/j.1744-6570.2007.00069.x

[2] W. Robert J. Alexander and Raffaella Belloni. 2024. Artificial Intelligence and the Sustainability of the Signaling and Human Capital Roles of Higher Education. *Sustainability* 16, 20 (2024), 8802. https://doi.org/10.3390/su16208802

[3] J. G. Altonji and C. R. Pierret. 2001. Employer Learning and Statistical Discrimination. *The Quarterly Journal of Economics* 116, 1 (2001), 313–350. https://doi.org/10.1162/003355301556329

[4] Daniel Aobdia, Hao Ma, and Sheryl Zhang. 2025. Auditing Effects on Employment Hiring: Evidence from the New York City Algorithmic Bias Audit Law. *SSRN Electronic Journal* (2025). https://doi.org/10.2139/ssrn.5703842

[5] Peter Blair, Tomas Castagnino, Erica Groshen, Papia Debroy, Byron Auguste, Shad Ahmed, Fernando Garcia Diaz, and Cristian Bonavida. 2020. *Searching for STARs: Work Experience as a Job Market Signal for Workers without Bachelor's Degrees.* https://doi.org/10.3386/w26844

[6] Stephen V. Burks, Bo Cowgill, Mitchell Hoffman, and Michael Gene Housman. 2013. The Value of Hiring Through Referrals. *SSRN Electronic Journal* (2013). https://doi.org/10.2139/ssrn.2266809

[7] Ming-Yu Chung, Jiashuo Fan, Hancheng Ye, Qinsi Wang, Wei-Chen Shen, Chia-Mu Yu, Pin-Yu Chen, and Sy-Yen Kuo. 2025. Model Reprogramming Demystified: A Neural Tangent Kernel Perspective. *arXiv preprint arXiv:2506.00620* (2025). https://arxiv.org/abs/2506.00620

[8] Thomas Corbin, Phillip Dawson, and Danny Liu. 2025. Talk is cheap: why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education* 50, 7 (2025), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964

[9] Aurelie Dariel, Arno M. Riedl, and Simon Siegenthaler. 2019. Hiring Through Referrals in a Labor Market with Adverse Selection. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3390277

[10] Phillip Dawson, Margaret Bearman, Mollie Dollinger, and David Boud. 2024. Validity matters more than cheating. *Assessment & Evaluation in Higher Education* 49, 7 (2024), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662

[11] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. 2022. UCC: Uncertainty guided Cross-head Cotraining for Semi-Supervised Semantic Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 9937–9946. https://doi.org/10.1109/cvpr52688.2022.00971

[12] Jiashuo Fan, Yaoyuan Liang, Leyao Liu, Shaolun Huang, and Lei Zhang. 2023. RCA-NOC: Relative Contrastive Alignment for Novel Object Captioning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 15464–15474. https://doi.org/10.1109/iccv51070.2023.01422

[13] Jiashuo Fan, Paul Rosu, Aaron T. Wang, Zeyu Michael Li, Lawrence Carin, and Xiang Cheng. 2025. In-Context Semi-Supervised Learning. *arXiv preprint arXiv:2512.15934* (2025). https://arxiv.org/abs/2512.15934

[14] Roberto M. Fernandez. 2010. Creating Connections for the Disadvantaged: Networks and Labor Market Intermediaries at the Hiring Interface. *SSRN Electronic Journal* (2010). https://doi.org/10.2139/ssrn.1576608

[15] J. Gervais. 2016. The operational definition of competency-based education. *The Journal of Competency-Based Education* 1, 2 (2016), 98–106. https://doi.org/10.1002/cbe2.1011

[16] David Gibson, Nathaniel Ostashewski, Kim Flintoff, Sheryl Grant, and Erin Knight. 2013. Digital badges in education. *Education and Information Technologies* 20, 2 (2013), 403–410. https://doi.org/10.1007/s10639-013-9291-7

[17] Maury Gittleman, Mark Klee, and Morris Kleiner. 2015. *Analyzing the Labor Market Outcomes of Occupational Licensing.* https://doi.org/10.3386/w20961

[18] Eugenia Gonzalez Ehlinger and Fabian Stephany. 2024. Skills or Degree? The Rise of Skill-Based Hiring for AI and Green Jobs. *SSRN Electronic Journal* (2024). https://doi.org/10.2139/ssrn.4743968

[19] Boris Gorelik. 2026. An Old Answer to a New Problem: The Degree-Free Model for Higher Education in the Age of AI. *Preprints* (Oct 2026). https://doi.org/10.20944/preprints202604.0391.v2

[20] Thomas Hungerford and Gary Solon. 1987. Sheepskin Effects in the Returns to Education. *The Review of Economics and Statistics* 69, 1 (1987), 175. https://doi.org/10.2307/1937919

[21] David A. Jaeger and Marianne E. Page. 1996. Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education. *The Review of Economics and Statistics* 78, 4 (1996), 733. https://doi.org/10.2307/2109960

[22] Michael T. Kane. 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 50, 1 (2013), 1–73. https://doi.org/10.1111/jedm.12000

[23] Fabian Lange. 2007. The Speed of Employer Learning. *Journal of Labor Economics* 25, 1 (2007), 1–35. https://doi.org/10.1086/508730

[24] Yaoyuan Liang, Zhao Yang, Yansong Tang, Jiashuo Fan, Ziran Li, Jingang Wang, Philip H.S. Torr, and Shao-Lun Huang. 2023. LUNA: Language as Continuing Anchors for Referring Expression Comprehension. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. ACM, 5174–5184. https://doi.org/10.1145/3581783.3612584

[25] Filip Lievens and Pol Coetsier. 2002. Situational Tests in Student Selection: An Examination of Predictive Validity, Adverse Impact, and Construct Validity. *International Journal of Selection and Assessment* 10, 4 (2002), 245–257. https://doi.org/10.1111/1468-2389.00215

[26] Michael A. McDaniel, Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer. 1994. The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology* 79, 4 (1994), 599–616. https://doi.org/10.1037/0021-9010.79.4.599

[27] OECD. 2020. *The emergence of alternative credentials.* https://doi.org/10.1787/b741f39e-en

[28] Mariana Oseguera. 2025. From Degrees to Skills: How Job Requirements Shape Applicant Composition and Fit. https://doi.org/10.1257/rct.16070

[29] Robert E. Ployhart and Brian C. Holtz. 2008. The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology* 61, 1 (2008), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

[30] Lincoln Quillian, John J. Lee, and Mariana Oliver. 2020. Evidence from Field Experiments in Hiring Shows Substantial Additional Racial Discrimination after the Callback. *Social Forces* 99, 2 (2020), 732–759. https://doi.org/10.1093/sf/soaa026

[31] Philip L. Roth, Philip Bobko, and Lynn A. McFarland. 2005. A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology* 58, 4 (2005), 1009–1037. https://doi.org/10.1111/j.1744-6570.2005.00714.x

[32] Deborah E. Rupp, Q. Chelsea Song, and Nicole Strah. 2020. Addressing the so-called validity-diversity trade-off: Exploring the practicalities and legal defensibility of Pareto-optimization for reducing adverse impact within personnel selection. *Industrial and Organizational Psychology* 13, 2 (2020), 246–271. https://doi.org/10.1017/iop.2020.19

[33] Disa Sariola, Patrick Button, Aron Culotta, and Nicholas Mattei. 2026. The Illusion of Fairness: Auditing Fairness Interventions in Algorithmic Hiring with Audit Studies. *Proceedings of the AAAI Conference on Artificial Intelligence* 40, 46 (2026), 39191–39200. https://doi.org/10.1609/aaai.v40i46.41267

[34] Peter Scarfe, Kelly Watcham, Alasdair Clarke, and Etienne Roesch. 2024. A real-world test of artificial intelligence infiltration of a university examinations system: A Turing Test case study. *PLOS ONE* 19, 6 (2024), e0305354. https://doi.org/10.1371/journal.pone.0305354

[35] Frank L. Schmidt and John E. Hunter. 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 (1998), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

[36] Michael Spence. 1973. Job Market Signaling. *The Quarterly Journal of Economics* 87, 3 (1973), 355. https://doi.org/10.2307/1882010

[37] Chad H. Van Iddekinge, Filip Lievens, and Paul R. Sackett. 2023. Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology* 76, 2 (2023), 651–686. https://doi.org/10.1111/peps.12578

[38] Soovendran Varadarajan, Joyce Hwee Ling Koh, and Ben Kei Daniel. 2023. A systematic review of the opportunities and challenges of micro-credentials for multiple stakeholders: learners, employers, higher education institutions and government. *International Journal of Educational Technology in Higher Education* 20, 1 (2023). https://doi.org/10.1186/s41239-023-00381-x

[39] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomas Foltynek, Jean Guerrero-Dib, Olumide Popoola, Petr Sigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity* 19, 1 (2023). https://doi.org/10.1007/s40979-023-00146-z

[40] Leesa Wheelahan. 2007. How competency-based training locks the working class out of powerful knowledge: a modified Bernsteinian analysis. *British Journal of Sociology of Education* 28, 5 (2007), 637–651. https://doi.org/10.1080/01425690701505540

[41] Leesa Wheelahan and Gavin Moodie. 2021. Gig qualifications for the gig economy: micro-credentials and the hungry mile. *Higher Education* 83, 6 (2021), 1279–1295. https://doi.org/10.1007/s10734-021-00742-3

[42] Will Yeadon, Elise Agra, Oto-Obong Inyang, Paul Mackay, and Arin Mizouri. 2024. Evaluating AI and human authorship quality in academic writing through physics essays. *European Journal of Physics* 45, 5 (2024), 055703. https://doi.org/10.1088/1361-6404/ad669d

[43] Deborah R. Yoder-Himes, Alina Asif, Kaelin Kinney, Tiffany J. Brandt, Rhiannon E. Cecil, Paul R. Himes, Cara H. Cashon, Rachel M. P. Hopp, and Edna Ross. 2022. Racial, skin tone, and sex disparities in automated proctoring software. *Frontiers in Education* 7 (2022). https://doi.org/10.3389/feduc.2022.881449

DATA AVAILABILITY

No individual-level human-subject, proprietary administrative, or downloaded public microdata were used. The public sources used to set parameter ranges are published articles and reports cited with DOIs or URLs in `refs.bib` and recorded in `refs.jsonl`. Synthetic outputs used in the analysis are available under `results/`.

CODE AVAILABILITY

Code is available from the corresponding author and will be released at publication.

CREDIT AUTHOR CONTRIBUTION

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Conceptualization, methodology, writing–original draft, writing–review and editing. Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Methodology, software, formal analysis, validation, data curation, visualization, writing–review and editing. Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Investigation, validation, writing–review and editing. Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Investigation, project administration, writing–review and editing.

FUNDING

This research received no external funding.

DECLARATION OF COMPETING INTEREST

The author declares no competing interests.

DECLARATION OF GENERATIVE AI USE

During the preparation of this work the authors used AI-assisted tools for literature search, code development, and drafting. The authors reviewed and edited the content and take full responsibility for it.

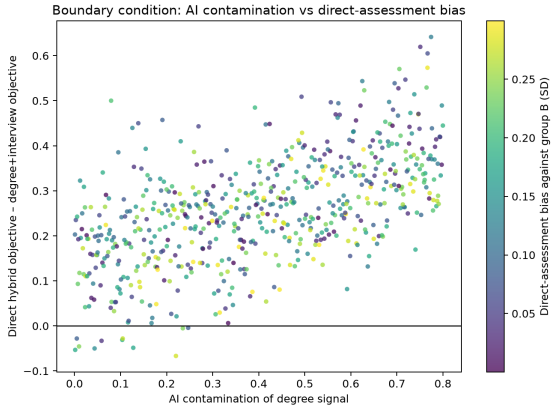

Figure 2: AI contamination, direct-assessment bias, and the objective difference between direct-hybrid and degree-plus-interview screening. Each point is a simulated scenario; positive values favor direct-hybrid screening.

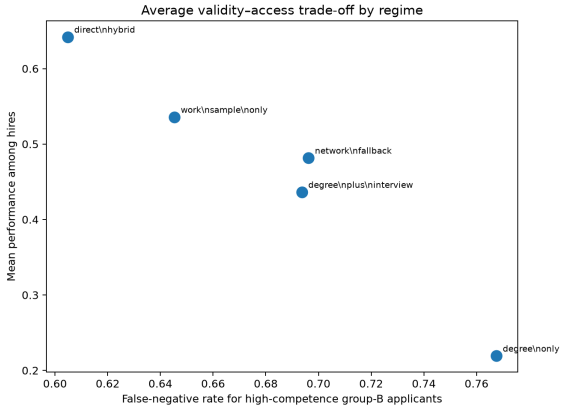

Figure 3: Average validity-access trade-off by screening regime. The figure compares mean performance and access outcomes across regimes; error bars show one standard deviation across design scenarios generated by the simulation.

Table 3: Structural sensitivity checks. Shares are proportions of scenarios in each 240-scenario check. D+I denotes degree plus interview; D+W denotes degree plus work sample; Aug. denotes augmented hybrid.

Check	Direct wins	Aug. wins	Direct > D+I	Direct > D+W	Direct > Aug.
Reference design	0.371	0.250	0.908	0.854	0.613
Matched effective validity	0.054	0.463	0.308	0.383	0.133
High direct-signal correlation	0.071	0.142	0.750	0.621	0.458
No AI contamination	0.129	0.525	0.700	0.533	0.163
No modeled group penalties	0.375	0.342	0.913	0.750	0.483
Matched validity, high correlation, no penalties	0.000	0.163	0.004	0.000	0.000