

# Do Academic Credentials Still Predict Professional Competence After Generative AI?

Doaa M. Bamasoud

University of Bisha

Mohammed Ayid Alqahtani

University of Bisha

Rasheed Mohammad

rasheed.mohammad@bcu.ac.uk

Birmingham City University

---

## Research Article

**Keywords:** generative artificial intelligence, signalling theory, predictive validity, academic credentials, professional licensure, equivalence testing

**Posted Date:** October 21st, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11163181/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

Generative artificial intelligence is widely expected to have hollowed out academic credentials by lowering the cost of producing assessed work. In signalling terms this is a cost shock to signal production, which destroys a separating equilibrium only if the cost reduction is regressive in ability. Framed through argument-based validity, it attacks the scoring and generalisation inferences, and matters for credentialing only if that contamination propagates to extrapolation. We derive five falsifiable predictions from these premises and test them against national licensure outcomes, an external proctored criterion. Linking each United States law school's entering-cohort credentials to that cohort's first-time bar passage three years later (2,350 school-year observations, outcome years 2014 to 2025), we first establish that the criterion is responsive to antecedent academic ability, then find no attenuation. The multiple correlation between entering credentials and licensure performance lies between 0.74 and 0.87 in every year and is higher after 2022 than before, and equivalence testing rejects attenuation beyond twenty per cent. A between-field contrast spanning the widest available exposure gap (2,152 program-year observations in law and nursing) shows no differential erosion, and no compression in the lower tail where the pooling account predicts it. The single discontinuity in the series precedes generative AI. The evidence favours a proportional or ability-complementary cost shock, and relocates the integrity risk upstream.

## 1. Introduction

Academic assessment is not merely a procedure for allocating marks; it is an inferential practice. From a limited sample of observable performance, assessors make claims about less directly observable capabilities: what a student knows, understands, can do independently, and can transfer beyond the immediate assessment task. In validity theory, this inference is defensible only when the interpretation and use of assessment outcomes are supported by evidence concerning the construct being measured, the conditions under which performance is produced, and the consequences attached to the resulting score (American Educational Research Association [AERA] et al., 2014; Cronbach & Meehl, 1955; Kane, 2013; Messick, 1995). Credentials amplify this inference, because they convert many local judgements into a portable signal of competence that is read by employers, professional bodies, postgraduate programs, and the public. Their legitimacy therefore depends on the continuing connection between the credential and the capability it is presumed to certify (Spence, 1973).

The rapid public diffusion of generative artificial intelligence (AI), particularly following the release of ChatGPT in November 2022, has disturbed one of the practical assumptions on which this inferential system depends: that submitted work is, within tolerable limits, an authentic representation of the student's own capability (OpenAI, 2022). Large language models now produce discipline-specific responses that human assessors often cannot distinguish from student-authored work, and in a consequential real-world test, AI-generated submissions inserted into an actual university examination system were largely undetected and on average received higher marks than student submissions (Scarfe et al., 2024). Detection-based responses have proved fragile under paraphrasing and inconsistent across

contexts (Sadasivan et al., 2023; Weber-Wulff et al., 2023), and institutional responses have consequently shifted towards assessment redesign and explicit use policies (Lodge et al., 2023; Perkins et al., 2024). This policy debate has moved considerably faster than the empirical evidence about credential-level consequences, and it is that gap the present study addresses.

The central distinction motivating this study is between technological capability and institutional consequence. That a language model can complete an essay, pass an examination item, or produce plausible coursework does not establish that grades, programs, or degrees have lost their signal value. Nor is the question merely whether such loss occurred: it is under what conditions it would be expected to occur, and those conditions are specifiable in advance. The concern that generative AI has hollowed out credentials is, when made precise, a conjunction of separable claims about the distributional incidence of a technological cost shock, about where in the curriculum that shock bites, and about which inference in the validity argument it breaks. Each conjunct is independently falsifiable, and setting them out separately is what permits a null result to carry information rather than merely to record an absence.

Section 2 derives the erosion hypothesis from three literatures, namely the economics of educational signalling, the task-based analysis of technological change, and argument-based validity, and states five hypotheses that follow from them. Professional licensure examinations then provide a rare opportunity to test those hypotheses, because they are external to the degree-awarding institution, high-stakes, standardised, and proctored, and because they are administered under conditions in which candidates cannot outsource the task to a language model. They therefore offer an unusually credible downstream criterion against which academic credentials can be assessed.

Two analyses follow. The primary analysis exploits the fact that the American Bar Association requires every accredited law school to disclose both the entering credentials of each cohort and the first-time bar passage of the cohort graduating that year. Because the Juris Doctor is a fixed three-year degree, these disclosures link into a genuine antecedent-to-criterion chain running from measured academic ability at matriculation, recorded under proctored conditions before any AI use during the degree, to independently certified competence three years later, likewise under proctored conditions. Everything between the two endpoints, meaning coursework, assessment, and grade formation, is exactly the interval in which substitution is alleged to occur, and nine outcome years precede the public release of generative AI while three follow it. The secondary analysis adds a between-field contrast, comparing programs in law and nursing, professions at opposite ends of the published language-model exposure gradient, in order to test the prediction that erosion should concentrate where assessed tasks are most substitutable.

The study makes three contributions. First, it derives the credential-erosion hypothesis from theory rather than from anxiety, decomposing it into a distributional claim about the incidence of a cost shock, a claim about where that shock should bite, and a claim about which inference in the validity argument it would break; it is this decomposition that allows the results to adjudicate rather than merely to reassure. Second, it tests those predictions against national, proctored licensure outcomes rather than against

laboratory demonstrations or self-reported use, and does so through an antecedent-to-criterion chain whose endpoints are both resistant to AI substitution. Third, it treats the resulting null with the methodological care that informative nulls require, combining equivalence testing with an explicit prior demonstration that the criterion is capable of registering the effect in question. That last step is routinely absent from null results in this literature, and without it a null is not evidence but silence. As reported below, the evidence does not support the erosion hypothesis in any of its testable forms, and read through the framework that generates it, the null identifies which version of the cost shock occurred and relocates the assessment-integrity risk to the part of the system this design cannot observe.

## **2. Literature Review and Theoretical Framework**

### **2.1 From capability to consequence**

The early empirical literature on generative AI in education was necessarily concerned with capability. Researchers asked whether large language models could answer examination questions, draft essays, write code, and perform discipline-specific reasoning at levels comparable to students, and the answer, across many domains, has been increasingly affirmative, though uneven across task types and model generations (Choi et al., 2022; Geerling et al., 2023; Katz et al., 2024; Kung et al., 2023). Capability evidence is reinforced by detection evidence, since if AI-generated work were reliably identifiable the threat to assessment validity would be containable through conventional misconduct procedures, and it is not. Broad evaluations of detection tools report inconsistent accuracy and sensitivity to obfuscation (Weber-Wulff et al., 2023), technical work shows that paraphrasing substantially undermines detector performance (Sadasivan et al., 2023), and the misclassification of non-native English writing creates a serious equity problem for punitive uses of detection (Liang et al., 2023).

The conclusion that AI can produce assessment-like performances does not, however, entail that credentials have changed their meaning. The erosion hypothesis requires a longer chain, in which students must use AI in ways that substitute for their own cognitive work, such use must remain undetected or institutionally tolerated, the affected tasks must carry sufficient weight in grade aggregation, and the resulting grades must be less aligned with the competencies that later matter. Each link is uncertain, and AI assistance may be complementary rather than substitutive, improving feedback, supporting revision, and enabling practice (Jensen et al., 2025; Perkins et al., 2024). Much of the public and institutional debate has treated the antecedent, that AI can complete the work, as sufficient evidence of the consequence, that the credential has lost validity. The remainder of this section separates those claims and specifies the conditions under which the second follows from the first.

### **2.2 Generative AI as a cost shock to signal production**

Academic credentials perform two analytically distinct functions, and generative AI threatens them by different routes. On the human-capital account, a degree is valuable because the education it certifies produces capability, and the credential predicts later performance because the capability is real (Becker, 1964). On the signalling account, a degree is valuable because acquiring it is differentially costly by

ability, and the credential is informative because only students above a threshold find it worth obtaining, whatever they learned along the way (Spence, 1973). Screening and filtering variants make the same structural point from the institution's side (Arrow, 1973; Stiglitz, 1975). The two accounts are not exclusive and their relative weight remains contested (Weiss, 1995), but they generate different predictions under a technology that lowers the cost of producing academic work without lowering the cost of acquiring competence.

This is precisely what generative AI does. Framed in Spence's terms, the public release of large language models constitutes an exogenous negative shock to the cost of signal production, because written coursework, case analysis, code, structured problem sets and literature summaries can now be generated at low marginal cost by a tool available to every enrolled student. Whether this destroys the signal depends entirely on how the cost reduction is distributed across the ability distribution, and it is at this point that the popular discourse and the formal theory diverge. A separating equilibrium in Spence's model is sustained by a negative correlation between ability and the cost of signal acquisition, and it collapses into pooling only if the cost shock is regressive in ability, meaning only if the tool reduces the cost of producing a given standard of work more for low-ability students than for high-ability ones.

Three outcomes are therefore possible rather than one. If the reduction is proportional, the ordering of students by realised attainment is preserved, levels shift but ranks do not, and the credential continues to separate. If the reduction is ability-complementary, with more capable students extracting more from the tool through better prompting, better evaluation of outputs, and better integration with their own reasoning, separation may even sharpen. Only under a regressive reduction does the signal degrade. The empirical literature on generative AI at work does not settle which obtains and reports both patterns, with substantial gains concentrated among lower performers in some settings (Brynjolfsson et al., 2025; Noy & Zhang, 2023) and sharply uneven returns contingent on task fit and user skill in others (Dell'Acqua et al., 2023). Evidence that unscaffolded access to generative tutors can reduce learning while raising immediate output quality (Bastani et al., 2024) is consistent with the regressive case for competence even where observed output converges.

The historical precedent is instructive and, we think, under-used in the present debate. American higher education absorbed a decades-long inflation of grades without a commensurate collapse in the predictive utility of academic records (Johnson, 2003; Rojstaczer & Healy, 2012), and the reason is structural, since inflation operated mainly on levels rather than on ranks, compressing the top of the distribution while largely preserving the ordering beneath it (Bar et al., 2009). Generative AI is a more powerful gaming technology than grade inflation, but it acts on the same object through the same channel, and the precedent counsels against inferring rank destruction from level movement. The general form of the concern is Campbell's law, under which an indicator recruited to high-stakes allocation attracts effort directed at the indicator rather than at what it measures (Campbell, 1979; Strathern, 1997), and contract cheating posed the same threat at smaller scale and higher cost a decade earlier (Curtis & Clare, 2017; Ellis et al., 2018; Newton, 2018), with the assessment-security literature that grew up around it anticipating much of the present debate (Dawson, 2021).

## 2.3 Where erosion should concentrate

Signalling theory says whether the credential could break, but it does not say where. For that we draw on the task-based framework of technological change, in which a technology displaces labour not by occupation but by task, and its aggregate effect on an occupation is a function of that occupation's task composition (Acemoglu & Restrepo, 2019; Autor, 2015; Autor et al., 2003). Applied to assessment rather than to work, the framework decomposes a field's assessment portfolio into tasks a language model can execute to rubric standard and tasks it cannot. The first category is textual, analytical, doctrinal, and rule-applying, and includes the essay, the case note, the problem set, the literature review, and the reflective commentary. The second is embodied, procedural, interpersonal, supervised or time-constrained, and includes the clinical placement, the simulated patient encounter, the practical examination, and the invigilated performance. A field's assessment exposure is the share of credential-determining assessment weight falling in the first category, and the erosion hypothesis, translated into this framework, predicts erosion increasing in assessment exposure.

Assessment exposure is not directly observable at national scale, but it is closely tracked by occupational exposure, because professional programs are accredited against the competencies their occupations require and assess accordingly. The language-modelling variant of the AI Occupational Exposure index (Felten et al., 2023), built on the ability taxonomy of Felten et al. (2021), therefore serves as an external and citable proxy for the construct the theory specifies, and its task-based analogues corroborate the resulting gradient (Eloundou et al., 2024). Law, with an index value of 1.45, and nursing, at 0.27, are not simply two convenient professions, since they lie close to the endpoints of the exposure distribution among occupations for which national program-level licensure outcomes exist. The contrast is theoretically motivated rather than merely available, and Section 3.2 sets out the occupational mapping in full.

## 2.4 Locating generative AI in the validity argument

Signalling theory identifies the mechanism and task composition identifies its incidence, but neither states what a credential's meaning consists in, or which part of it AI attacks. Argument-based validity supplies that precision (AERA et al., 2014; Kane, 2013; Messick, 1989, 1995). On this account, the use of an assessment result rests on a chain of inferences, each of which can fail independently: scoring, from the observed performance to a score reflecting that performance; generalisation, from the score on this task to the universe of similar tasks the student could perform; extrapolation, from that universe to the domain of professional practice the credential concerns; and implication, from the domain inference to the decision the credential licenses.

Generative AI attacks the chain at the first link. Where a submission is not the student's own work, the score does not reflect the student's performance, and generalisation fails as a consequence, because a score that does not reflect this performance cannot generalise across the student's other performances. Extrapolation and implication are not attacked directly, and they fail only if contamination of the first two links propagates. This inferential structure is what makes an external, proctored criterion the right

instrument, since a licensure examination administered under conditions the tool cannot reach re-establishes the scoring inference for a single standardised performance and tests extrapolation and implication against it. Observing whether the credential still predicts that performance is therefore a direct test of propagation.

If the relationship holds, the interpretation is not that scoring and generalisation were undamaged, but that the damage did not reach the inference the credential is actually used to license. A credentialing system containing multiple imperfect screens in series can sustain contamination of one screen without loss of terminal validity, and that structural redundancy, rather than the integrity of any single assessment, would be what protected it. Framing the test this way also delimits the claim, since it concerns one inference link rather than validity in general, addresses neither the construct breadth of licensure examinations nor the consequential dimension of validity (Sireci, 2016), and cannot speak to learning that was never assessed at the criterion either.

## 2.5 Hypotheses

The three layers combine into a set of falsifiable predictions, which we state as the erosion hypothesis would have them so that each can be rejected. The first is a precondition rather than a hypothesis. **P1 (criterion responsiveness)** holds that the licensure criterion covaries with antecedent academic ability. Unless P1 holds, the remaining hypotheses cannot be tested, because a criterion insensitive to what precedes it cannot register a change in what precedes it, and it must therefore be established before any null on the hypotheses below is interpretable.

The four substantive hypotheses follow from Sections 2.2 to 2.4. **H1 (differential erosion)** holds that post-2022 licensure outcomes declined more in high-exposure than in low-exposure fields, net of common shocks, and follows from the task-composition argument. **H2 (dose-response)** holds that erosion increases with the share of a cohort's program of study exposed to generative AI, and distinguishes a genuine exposure effect from a period effect. **H3 (pooling and compression)** holds that any differential gain is concentrated in the lower tail of the outcome distribution, consistent with a cost shock regressive in ability; this is the specific prediction of separating-to-pooling collapse, and no estimate of a conditional mean tests it. **H4 (predictive attenuation)** holds that the relationship between measured entering ability and subsequent proctored performance weakened for AI-exposed cohorts, and is the propagation test proper. **H5 (human-capital loss)** holds that conditional on entering ability, AI-exposed cohorts reached the criterion less able, indicating that the degree added less capability rather than merely reordering it less well, and distinguishes the human-capital from the signalling channel. Table 9 in Section 5 maps each prediction to the test that addresses it and to the result obtained.

## 3. Methodology

### 3.1 Data

The primary data source is the American Bar Association's Standard 509 Required Disclosures, which every accredited law school must file annually. Each disclosure reports the median Law School Admission Test (LSAT) score and median undergraduate grade point average of the entering cohort, alongside the first-time bar-passage rate of the cohort that graduated that year. Because the Juris Doctor is a three-year degree, the entering credentials of the cohort sitting the bar in year  $T$  were recorded in year  $T$  minus 3, and the two disclosures link into an antecedent-to-criterion chain spanning the full period of professional education. Merging on the three-year lag, dropping cohorts with fewer than ten first-time takers and a small number of records with artifactual zero rates, yields 2,350 school-year observations across 206 law schools, covering outcome years 2014 to 2025 and entering cohorts 2011 to 2022. Nine outcome years precede the public release of generative AI and three follow it. Where jurisdictions cancelled sittings or granted diploma privilege during 2020, the published state comparison rate is recorded as zero, and we treat these as missing rather than as a zero per cent jurisdictional pass rate.

We use this link deliberately and with a stated limitation. The LSAT is administered under proctored conditions before matriculation and is therefore uncontaminated by AI use during the degree, and the bar examination is likewise proctored, so what the chain measures is whether a school's intake of measured academic ability continues to convert into independently certified competence at the same rate. It is a proxy for, rather than an identical construct to, the coursework-to-competence relationship that generative AI is hypothesised to disturb, and we treat it as such throughout. Its virtue is that both endpoints resist AI substitution while everything between them is exactly the interval in which substitution is alleged to occur. If AI-assisted coursework has decoupled academic progression from underlying capability, a cohort of given measured entering ability should emerge less able to satisfy the proctored criterion, or the orderliness of that mapping should degrade, and both are observable here.

The secondary data source supports the between-field contrast, for which the outcome is the first-time pass rate of a degree-granting program's graduates on the relevant professional licensure examination, observed annually at the program level. For nursing we compile first-time NCLEX-RN pass rates for pre-licensure registered-nursing programs from United States state board of nursing reports, linked to institutional characteristics drawn from the federal Integrated Postsecondary Education Data System (IPEDS). For law we use school-level first-time bar-passage rates compiled from the same Standard 509 disclosures and distributed through the AccessLex Analytix repository, which assembles these reports into a longitudinal panel spanning 210 accredited law schools from 2008 onward. Throughout, we use first-time rather than eventual pass rates, because the first-time rate is most directly tied to the competence a graduating cohort carries out of its program, and because using it consistently across both professions keeps the criterion comparable.

Pooling the two fields over an analysis window running to 2025, and removing records with artifactual zero pass rates, yields 2,152 program-year observations across 326 programs. Table 1 summarises the two panels, and the contrast the design exploits is visible in the raw descriptive statistics, since law programs pass first-time candidates at a markedly lower and far more variable rate than nursing

programs. The greater dispersion of the law panel is analytically valuable, because it furnishes substantial outcome variance in precisely the high-exposure field where the hypothesised erosion should be easiest to detect. The two panels are not observed over the same span, however, since law programs enter in 2018 and nursing programs only in 2021, so the between-field design rests on a two-year pre-AI baseline for the low-exposure field, and both of those years fall within the period of pandemic disruption to nursing instruction and clinical placement. The consequences for inference are material and are addressed directly in Sections 3.5 and 4.9 rather than deferred to the limitations, and they are the reason the within-field analysis, which carries nine pre-AI outcome years and does not use the nursing data at all, is the design on which we rest the paper’s conclusion.

Table 1  
Descriptive statistics by field. The analysis window runs to 2025; law programs are observed from 2018, nursing programs from 2021.

| Field   | Programs | Obs.  | Years     | Mean pass % | SD    | Mean cohort | LM-AIOE |
|---------|----------|-------|-----------|-------------|-------|-------------|---------|
| Law     | 205      | 1,563 | 2018–2025 | 77.96       | 13.75 | 172         | 1.454   |
| Nursing | 121      | 589   | 2021–2025 | 90.73       | 8.28  | 105         | 0.272   |

### 3.2 Measures

The outcome throughout is the first-time pass rate expressed in percentage points, and, where the bounded nature of that outcome matters, its logit transform. The antecedent in the within-field analysis is the entering cohort’s median LSAT score, with median undergraduate grade point average entered alongside it as a second credential. A state-relative outcome, constructed as the school’s pass rate minus the pass rate of all candidates in the same jurisdiction, is used in one specification to net out jurisdictional difficulty.

Field exposure is measured by assigning each field the language-modelling AI Occupational Exposure (LM-AIOE) score of the occupation its credential licenses, namely registered nurses (Standard Occupational Classification 29-1141) for nursing and lawyers (23-1011) for law. On this published index nursing scores 0.272 and law 1.454, a gap of 1.18 units that represents the widest clean contrast available among professions for which national program-level licensure outcomes can be assembled. We use the raw, citable index value as the analytic exposure variable, so that a field’s exposure does not depend on the particular set of fields included in the study, and the construct, the occupational mapping, and the supporting validity evidence are detailed in the Appendix. Exposure to generative AI within a cohort’s program of study is, finally, graded rather than binary, because the cohort sitting the bar in 2023 encountered generative AI only in its final semester whereas the 2025 cohort had it available for the greater part of the degree. We therefore construct a continuous dose equal to the share of the thirty-six-month degree falling after November 2022, which takes values of 0.167, 0.500 and 0.833 for the 2023, 2024 and 2025 outcome years respectively, and we report results for both this dose and a binary post-2022 indicator.

### 3.3 Empirical strategy for the within-field test

The within-field analysis estimates whether the relationship between entering credentials and first-time bar passage changed for cohorts whose legal education overlapped the availability of generative AI. A school's outcome in a given year is modelled as a function of its entering cohort's median LSAT and grade point average, an indicator for the AI-exposed period, and the interaction of the two, with year fixed effects and, in the more demanding specification, school fixed effects, with standard errors clustered at the school level. Identification of the interaction rests on within-school change in how entering ability maps onto licensure performance, which removes by construction the objection that jurisdictions or institutions differ in unobserved ways. Because treatment varies across 206 clusters, conventional cluster-robust inference is appropriate here, and we corroborate every headline estimate with a wild cluster bootstrap under the null using Webb six-point weights (Cameron et al., 2008; Cameron & Miller, 2015; MacKinnon & Webb, 2017). Before estimating any change in the relationship we establish P1, that the criterion is capable of registering one, and this ordering is not presentational, since a null on H4 estimated against an unresponsive criterion is uninformative and the literature's nulls are routinely reported without the corresponding check.

### 3.4 Scale, dispersion, and the slope-correlation identity

The outcome is a bounded proportion, and between-school dispersion in pass rates fell markedly across the study window. This creates a measurement problem that must be confronted directly, because the two statistics commonly used to express a predictive relationship respond differently to changes in dispersion, and neither is unconditionally invariant to it. Ignoring the problem would allow a purely arithmetic movement to be read as a substantive one, which is the specific error the erosion hypothesis is most likely to induce in a bounded outcome approaching a ceiling. The two statistics are linked by the identity

$$b = \beta \times (SDy / SDx),$$

where  $b$  is the unstandardised regression slope of the criterion on the predictor,  $\beta$  the standardised coefficient, and  $SDy$  and  $SDx$  the standard deviations of criterion and predictor. Where the predictor is entered alone,  $\beta$  is simply the correlation between them, and where the model conditions on further covariates,  $\beta$  is the standardised partial coefficient. The identity is exact within any single estimation sample, so it partitions any observed change in the slope into a component attributable to the association and a component attributable to relative dispersion, without requiring an assumption about which of the two moved. The classical psychometric result concerns restriction of range on the predictor, and holds that where a population is selected on  $x$ , the unstandardised slope is approximately invariant while the correlation attenuates, which is why corrections for range restriction are applied to correlations and not to slopes (Sackett & Yang, 2000; Thorndike, 1949). Where, by contrast, the criterion distribution compresses for reasons unrelated to the predictor, as a bounded proportion approaching a ceiling will, the slope contracts arithmetically while the correlation need not. Which statistic moves therefore locates where the dispersion change occurred, and neither can be designated the invariant primitive in advance.

We accordingly privilege neither. We estimate all models on the raw percentage scale and on the logit scale, which is unbounded and immune to the ceiling mechanism, we report the correlation alongside them, and we report the decomposition of any observed change in  $b$  into its association and dispersion components. Where the scales disagree we say so and use the decomposition to explain why, rather than selecting the scale that favours a conclusion, and as a further check we report correlations corrected for restriction of range in the entering-credential distribution under Thorndike's Case II.

### **3.5 Empirical strategy for the between-field test**

The between-field analysis estimates the differential association of the generative-AI period with licensure pass rates as a function of field exposure. A program's first-time pass rate in a given year is modelled as a function of the interaction between its field's mean-centred LM-AIOE exposure and an indicator for the post-AI period, defined as years 2023 and later, together with field and year fixed effects, with standard errors clustered at the program level. The year fixed effects absorb any shock common to all fields in a given year, including pandemic-era disruptions, the Next Generation NCLEX transition, and contemporaneous bar-examination changes, while the field fixed effects absorb stable differences between the two professions. We report the coefficient scaled to the full nursing-to-law exposure gap, so that it reads as the implied differential erosion in percentage points between the least- and most-exposed professions.

Two features of this design constrain what it can support, and we state them here rather than in the limitations. First, exposure takes only two values, so the effective number of treated clusters is one, program-level clustering does not deliver valid inference on a field-level treatment, and conventional cluster-robust standard errors will be optimistic (Conley & Taber, 2011; Ferman & Pinto, 2019; MacKinnon & Webb, 2017). We therefore accompany every between-field estimate with randomisation inference over the exposure assignment and with a wild cluster bootstrap using Webb weights, and we treat the cluster-robust intervals as illustrative rather than as the basis for inference. Second, the two-year nursing pre-period makes the parallel-trends assumption unverifiable, and pre-trend tests conducted on so short a window have very low power against the alternatives that matter (Roth, 2022). Rather than assert the assumption, we report the specification that gives each field its own linear trend, which relaxes precisely the assumption the short pre-period makes hardest to defend, and we recommend sensitivity bounds under stated relaxations of parallel trends (Rambachan & Roth, 2023) for any future version of this design with a longer pre-period. The between-field analysis is reported as corroborating evidence on the exposure gradient and does not carry the paper's conclusion.

### **3.6 Equivalence testing and the smallest effect of interest**

Because the central findings are absences, we supplement conventional hypothesis tests with equivalence testing throughout, using the two one-sided tests procedure (Lakens, 2017; Lakens et al., 2018). A significant equivalence test licenses the positive claim that any effect is smaller than a stated bound, whereas a non-significant conventional test does not. The bounds require justification rather than assertion. For the between-field analysis we set the smallest effect size of interest at five percentage

points across the full exposure gap, which is approximately half the magnitude of the national decline in first-time bar passage over 2014 to 2016, the most recent change in licensure outcomes large enough to prompt sustained regulatory attention. A divergence between the least- and most-exposed professions of half that size would be the smallest that could plausibly carry practical meaning for the credentialing system. For the within-field analysis the bound is defined relative to the pre-AI predictive slope rather than in absolute units, since the practically meaningful question is what proportion of the credential's predictive power was lost, and we report the tightest bound at which equivalence can be established together with the bound at which it cannot. In both cases we report the minimum effect detectable at eighty per cent power, so that the precision of each null is explicit.

## 4. Results

### 4.1 The criterion is responsive (P1)

Before asking whether the predictive relationship weakened, we establish that the criterion can register it. Table 2 reports the association between entering credentials and first-time bar passage three years later, and it is large and precisely estimated. Each additional point of entering median LSAT is associated with 1.565 percentage points of first-time bar passage (SE 0.059,  $t = 26.73$ ), and the model explains 64 per cent of the variance in school-level outcomes. Adding school fixed effects, so that the estimate rests only on within-school variation over time, strengthens rather than weakens it, giving 1.839 points (SE 0.171,  $t = 10.73$ ) with an  $R^2$  of 0.847. The association also survives measurement against the state-relative outcome, which nets out jurisdictional difficulty, at 0.672 points (SE 0.238,  $p = .005$ ).

Table 2  
Entering-cohort credentials and first-time bar passage: the criterion-responsiveness test.

| Specification                                                                                                                                                                                           | Coef  | SE    | <i>t</i> | <i>p</i> | 95% CI       | <i>R</i> <sup>2</sup> | <i>n</i> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-------|----------|----------|--------------|-----------------------|----------|
| LSAT only, year FE                                                                                                                                                                                      | 1.565 | 0.059 | 26.73    | < .001   | [1.45, 1.68] | 0.643                 | 2,350    |
| LSAT + UGPA, year FE                                                                                                                                                                                    | 1.124 | 0.233 | 4.82     | < .001   | [0.66, 1.58] | 0.662                 | 2,350    |
| LSAT + UGPA, year + school FE                                                                                                                                                                           | 1.839 | 0.171 | 10.73    | < .001   | [1.50, 2.18] | 0.847                 | 2,350    |
| State-relative outcome, year FE                                                                                                                                                                         | 0.672 | 0.238 | 2.83     | .005     | [0.20, 1.14] | 0.605                 | 2,306    |
| Logit outcome, year FE                                                                                                                                                                                  | 0.084 | 0.009 | 8.89     | < .001   | [0.07, 0.10] | 0.699                 | 2,341    |
| Note. Coefficients are percentage points of first-time bar passage per one point of entering median LSAT, except the logit row, which is in logits. Standard errors clustered by school (206 clusters). |       |       |          |          |              |                       |          |

One feature of this result requires explicit statement, since it bears on how the magnitude should be read. These are associations between school-level aggregates, namely a median entering LSAT and a

cohort pass rate, and correlations between aggregates are systematically larger than the individual-level relationships they are sometimes taken to represent, because aggregation removes within-group variance (Robinson, 1950). The individual-level predictive validity of the LSAT against bar passage, as reported in the LSAC validity literature, is substantially smaller (Wightman, 1998). We make no inference from these estimates to individual candidates, and none is required, because the design asks whether the criterion as analysed, at the school-year level, is responsive to the antecedent as analysed at the same level, and it demonstrably is. The argumentative weight of this result is greater than its simplicity suggests, since the most serious objection to a null estimated on licensure data is that these examinations sit far downstream of coursework and are buffered by intensive commercial preparation, rendering them insensitive to upstream change. That objection is testable and it fails, because bar passage tracks the measured academic ability of the entering cohort three years earlier closely and consistently. P1 is established, the nulls that follow are therefore interpretable, and Section 5.1 returns to what this forecloses.

## 4.2 The validity coefficient did not decline (H4)

Figure 1 plots the validity coefficient, the multiple correlation between entering credentials and licensure performance, for each outcome year. It lies between 0.744 and 0.865 in every year of the panel, its pre-AI mean is 0.802 and its post-AI mean 0.846, all three AI-exposed years exceed the pre-AI mean, and the lowest post-AI value of 0.831 in 2023 exceeds seven of the nine pre-AI values. A school-clustered bootstrap of the pre-post difference, resampling schools rather than treating annual correlations as independent, returns an increase of 0.044 in correlation units with a 95 per cent interval of [+ 0.011, + 0.071] and a  $p$ -value of .005, which is distinguishable from zero and in the direction opposite to the erosion hypothesis. Correlations corrected for restriction of range in the entering-credential distribution under Thorndike's Case II show the same pattern. We do not interpret the increase substantively, because it is small and nothing in the erosion literature predicts improvement, but what the estimate establishes is that the credential's ordering of graduates by subsequent independent competence is at least as orderly after generative AI as before. H4 is not supported.

## 4.3 The predictive slope and the dispersion decomposition (H2, H4)

Table 3 reports the change in the predictive slope across specifications and both scales, and the two scales disagree in a way that must be confronted rather than resolved by selection. Estimated in percentage points with year fixed effects, the slope appears to attenuate by 36.8 per cent ( $p = .063$ ), and under the continuous-dose specification by 56.2 per cent ( $p = .025$ ). Estimated on the logit scale, the same comparisons give 18.7 per cent ( $p = .280$ ) and 24.5 per cent ( $p = .307$ ), and with school fixed effects added they fall to 3.9 and 4.3 per cent ( $p = .609$  and  $.678$ ). Wild cluster bootstrap  $p$ -values corroborate the logit results, at .728 and .715, and are markedly more conservative than the cluster-robust values throughout. Figure 2 shows the annual series behind these estimates on both scales.

Table 3  
Change in the credential-to-criterion predictive slope, by specification and scale.

| Scale                                                                                                                                                                         | Specification                 | Base slope | Change | % of base | SE    | <i>p</i> (CR1) | <i>p</i> (boot) |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------|------------|--------|-----------|-------|----------------|-----------------|
| Raw pp                                                                                                                                                                        | Binary post, year FE          | 1.230      | -0.452 | -36.8     | 0.242 | .063           | .237            |
| Raw pp                                                                                                                                                                        | Binary post, year + school FE | 1.932      | -0.235 | -12.2     | 0.170 | .168           | .274            |
| Raw pp                                                                                                                                                                        | Dose, year FE                 | 1.214      | -0.683 | -56.2     | 0.303 | .025           | .148            |
| Raw pp                                                                                                                                                                        | Dose, year + school FE        | 1.966      | -0.338 | -17.2     | 0.208 | .106           | .168            |
| Logit                                                                                                                                                                         | Binary post, year FE          | 0.087      | -0.016 | -18.7     | 0.015 | .280           | .644            |
| Logit                                                                                                                                                                         | Binary post, year + school FE | 0.120      | -0.005 | -3.9      | 0.009 | .609           | .733            |
| Logit                                                                                                                                                                         | Dose, year FE                 | 0.086      | -0.021 | -24.5     | 0.021 | .307           | .592            |
| Logit                                                                                                                                                                         | Dose, year + school FE        | 0.120      | -0.005 | -4.3      | 0.012 | .678           | .701            |
| <i>Note.</i> Standard errors clustered by school (206 clusters); <i>p</i> (boot) is a wild cluster bootstrap under the null with Webb six-point weights and 999 replications. |                               |            |        |           |       |                |                 |

The identity of Section 3.4 resolves the disagreement arithmetically rather than by argument, and it does so at the level of the annual estimate, where the identity is exact. Table 4 reports the three quantities for every outcome year, and they reconcile to machine precision, with a maximum absolute discrepancy of 2.2 times ten to the power of minus sixteen. The annual partial slope falls by 30.1 per cent between the pre-AI and AI-exposed periods, and that fall decomposes into a contraction of 19.3 per cent in the dispersion ratio and a contraction of 12.6 per cent in the standardised partial coefficient, whose product reproduces the observed change. Approximately three-fifths of the raw-scale attenuation is therefore arithmetic, produced by the narrowing of a bounded outcome distribution, and approximately two-fifths reflects a genuine, and small, weakening of the standardised association between entering credentials and licensure performance.

We report this division precisely because it is not the cleanest possible result. The dispersion mechanism does not account for all of the raw-scale movement, and a reader is entitled to know how much it leaves, with what remains bounded by the equivalence test of Section 4.4, which rejects attenuation greater than twenty per cent of the predictive slope on the specification least vulnerable to the scale problem. Two features of the data locate the arithmetic component. Between-school dispersion in the outcome fell from a pre-AI mean of 14.2 percentage points to 11.5 in the AI-exposed years, and to 10.7 by 2025, while dispersion in entering median LSAT was essentially flat, moving from 6.89 to 6.96 points. As Fig. 3 shows, the annual partial slope co-moves with outcome dispersion at a

correlation of + 0.62 across the pre-AI period, and with predictor dispersion at + 0.71, so a slope expressed in units of a bounded outcome must contract as that outcome's distribution narrows. On the unbounded logit scale, which is not subject to the ceiling mechanism, and with school fixed effects, the contraction falls to 3.9 per cent.

Table 4  
Annual decomposition of the predictive slope,  $b = \beta \times (\text{SDy} / \text{SDx})$ .

| Outcome year                                                                                                                                                                                                                                                                           | <i>R</i> | SDy (pp) | SDx (LSAT pts) | SDy/SDx | $\beta$ | <i>b</i> (pp per LSAT pt) |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|----------|----------------|---------|---------|---------------------------|
| 2014                                                                                                                                                                                                                                                                                   | 0.744    | 11.85    | 6.66           | 1.777   | 0.483   | 0.858                     |
| 2015                                                                                                                                                                                                                                                                                   | 0.783    | 13.35    | 6.92           | 1.930   | 0.432   | 0.834                     |
| 2016                                                                                                                                                                                                                                                                                   | 0.794    | 15.58    | 6.88           | 2.265   | 0.431   | 0.977                     |
| 2017                                                                                                                                                                                                                                                                                   | 0.799    | 15.08    | 7.03           | 2.144   | 0.659   | 1.413                     |
| 2018                                                                                                                                                                                                                                                                                   | 0.781    | 16.29    | 6.99           | 2.333   | 0.664   | 1.550                     |
| 2019                                                                                                                                                                                                                                                                                   | 0.797    | 14.14    | 6.97           | 2.029   | 0.589   | 1.195                     |
| 2020                                                                                                                                                                                                                                                                                   | 0.805    | 13.48    | 6.92           | 1.947   | 0.640   | 1.245                     |
| 2021                                                                                                                                                                                                                                                                                   | 0.853    | 14.10    | 6.88           | 2.048   | 0.669   | 1.370                     |
| 2022                                                                                                                                                                                                                                                                                   | 0.861    | 14.00    | 6.73           | 2.080   | 0.431   | 0.896                     |
| 2023                                                                                                                                                                                                                                                                                   | 0.831    | 12.84    | 6.79           | 1.890   | 0.432   | 0.817                     |
| 2024                                                                                                                                                                                                                                                                                   | 0.865    | 11.11    | 6.91           | 1.608   | 0.559   | 0.899                     |
| 2025                                                                                                                                                                                                                                                                                   | 0.840    | 10.68    | 7.17           | 1.491   | 0.465   | 0.693                     |
| <b>Pre-AI mean</b>                                                                                                                                                                                                                                                                     | 0.802    | 14.21    | 6.89           | 2.061   | 0.555   | 1.149                     |
| <b>AI-exposed mean</b>                                                                                                                                                                                                                                                                 | 0.846    | 11.54    | 6.96           | 1.663   | 0.485   | 0.803                     |
| <b>Change</b>                                                                                                                                                                                                                                                                          | + 5.5%   | -18.7%   | + 1.0%         | -19.3%  | -12.6%  | -30.1%                    |
| <i>Note.</i> <i>R</i> is the multiple correlation from the within-year regression of first-time bar passage on entering median LSAT and median undergraduate grade point average; $\beta$ is the standardised partial coefficient on LSAT and <i>b</i> its unstandardised counterpart. |          |          |                |         |         |                           |

The dose specification is the direct test of H2. On the raw scale it returns the largest apparent attenuation in Table 3, and on the logit scale with school fixed effects the smallest, and the decomposition explains why, since the dose variable is collinear with calendar time and therefore with the dispersion contraction, so that a raw-scale dose coefficient absorbs the arithmetic component almost entirely. H2 is not supported on any specification robust to the scale problem.

## 4.4 Equivalence

Conventional tests establish only that attenuation cannot be rejected, whereas equivalence testing supports the affirmative claim. Taking the more demanding specification, with school fixed effects on the logit scale, two one-sided tests reject attenuation greater than twenty per cent of the pre-AI predictive slope ( $p = .017$ ), equivalence is established more comfortably at the looser bounds of thirty-three and twenty-five per cent ( $p = .0001$  and  $.003$  respectively), and it cannot be established at fifteen per cent ( $p = .071$ ). The design retained eighty per cent power to detect attenuation of 21.3 per cent. The claim the data support is therefore specific and bounded, in that whatever generative AI did to legal education between 2022 and 2025, it did not cost the credential more than about a fifth of its power to predict independently certified competence. We do not claim more, and the fifteen-per-cent result marks where the evidence stops.

## 4.5 Dating the only discontinuity

The twelve-year series is not featureless, and the location of its one break is informative. On the logit scale the annual partial slope runs at 0.098 and 0.102 through 2020 and 2021 and then falls to 0.066 in 2022, where it remains, at 0.064, 0.078 and 0.071, across the three AI-exposed years, as the right panel of Fig. 2 shows. The break is dated to 2022, a cohort that matriculated in 2019 and completed its coursework before generative AI was publicly available. A second and independent analysis reaches the same date, since fitting the credential-to-criterion model with a linear time trend on pre-AI years alone and extrapolating forward gives the largest negative mean residual in the panel to 2022, at -3.59 percentage points, after which the AI-exposed cohorts recover monotonically toward the prediction at -2.62, -1.72 and -1.00 points. The trend term is required for the extrapolation to be meaningful, because without it the series has no anchor and 2016 rather than 2022 carries the largest negative residual, which is why we report the specification explicitly rather than describing the exercise generically.

Figure S1 makes the same point through boundary sensitivity. Re-estimating the identical model at every candidate boundary year from 2016 to 2025 produces changes ranging from +28.0 to -6.1 per cent of the base slope, the true 2023 boundary yields -3.9 per cent and ranks eighth of ten in magnitude, and several false boundaries produce larger and strongly significant apparent effects. We do not read this as a permutation test, since candidate boundaries in an ordered series are not exchangeable, but we do read it as demonstrating that the estimate at the AI boundary is unremarkable within the ordinary variation this design produces, and that boundary choice rather than generative AI drives any apparent effect.

## 4.6 The value-added test is not identified (H5)

Asking whether cohorts of given entering ability emerged less competent, which is a conditional level shift rather than a slope change, requires a time trend, because year fixed effects absorb the quantity of interest by construction. With only three post-AI years the trend specification does the work, and estimates range from -2.94 percentage points under a linear trend with a binary treatment to +7.82 under a quadratic trend with continuous dose, with sign reversal across defensible choices and strong significance on both sides of zero, as Supplementary Table S1 reports. H5 cannot be tested with the data

available, so we report the analysis as a sensitivity exercise rather than as a result and draw no conclusion from it. Its one stable feature is the per-year residual pattern described in Section 4.5, whose trough precedes generative AI.

## 4.7 The between-field contrast (H1)

Figure 4 plots mean first-time pass rates for the two fields across the study window, with the timing of generative AI’s public arrival marked. The fields differ substantially in level and volatility, as the descriptive statistics indicated, but neither series exhibits the divergence the erosion hypothesis predicts, since the law trajectory does not bend downward relative to nursing after 2022. Whatever year-to-year movement is present appears in both fields and predates the AI inflection as often as it follows it.

The principal specification returns a differential-erosion estimate of + 0.07 percentage points across the full nursing-to-law exposure gap, as row 1 of Table 5 reports. The point estimate is for practical purposes zero, and marginally positive, which is the opposite sign to that predicted. The cluster-robust interval of [-1.41, + 1.56] excludes any substantively meaningful differential erosion, though for the reasons set out in Section 3.5 it should be read as illustrative, and randomisation inference over the exposure assignment returns the same qualitative conclusion at a *p*-value of .942. H1 is not supported. Specification 5, which allows each field its own linear time trend and therefore relaxes the assumption the short nursing pre-period makes hardest to defend, returns a larger coefficient in the predicted direction but one still indistinguishable from zero and wholly within the region of practical equivalence.

Table 5

Differential-erosion estimates across specifications. Coefficients are scaled to the full nursing-to-law exposure gap (percentage points).

| Specification                                                                                                                                    | Coef (per gap) | SE   | <i>p</i> | 95% CI          |
|--------------------------------------------------------------------------------------------------------------------------------------------------|----------------|------|----------|-----------------|
| (1) Field + year FE [main]                                                                                                                       | + 0.07         | 0.76 | .921     | [-1.41, + 1.56] |
| (2) Boundary = 2022                                                                                                                              | -0.07          | 0.93 | .938     | [-1.89, + 1.75] |
| (3) Boundary = 2024                                                                                                                              | -0.44          | 0.77 | .568     | [-1.95, + 1.07] |
| (4) Program FE                                                                                                                                   | + 0.10         | 0.73 | .891     | [-1.33, + 1.53] |
| (5) Field-specific trends                                                                                                                        | -1.13          | 1.09 | .302     | [-3.28, + 1.02] |
| (6) Cohort <i>n</i> ≥ 25                                                                                                                         | + 0.33         | 0.69 | .627     | [-1.02, + 1.68] |
| Note. Standard errors clustered at the program level. See Section 3.5 on the limits of cluster-robust inference where exposure takes two values. |                |      |          |                 |

Figure 5 decomposes the result by year, and one constraint on this exhibit must be stated before it is read. Nursing programs enter the panel only in 2021, so exposure-related divergence is identified in four years alone, namely 2021 and the three AI-exposed years, and event-study coefficients for 2018 to 2020, which a naive estimation on the full window will produce, are identified from the law panel only and are not interpretable as divergence. Within the identified window the estimates do not describe an erosion

pattern, since the single pre-period point in 2021 is + 0.20 points and indistinguishable from zero, and the AI-exposed years are + 0.77, -2.28 and + 2.07 points, alternating in sign, with the negative 2024 estimate offset by a positive 2025 estimate of comparable magnitude. There is no step, no drift, and no accumulating wedge of the kind the erosion account requires. What there also is not, and we say so plainly, is a pre-trend, because one pre-period year cannot establish one, and the exhibit should be read as a decomposition of the post-period rather than as a test of parallel trends.

The two one-sided tests reject the presence of any differential erosion larger than the five-point bound justified in Section 3.6, with a *p*-value below .001, as Table 6 reports and Figure S2 displays. The design retained eighty per cent power to detect a differential erosion as small as 2.12 percentage points under cluster-robust inference, against which the observed + 0.07 points is negligible. Randomisation inference returns a null distribution with a standard deviation of 1.04 points, appreciably wider than the cluster-robust standard error of 0.76, implying a minimum detectable effect nearer 2.92 points, and we report both while relying on the second.

Table 6  
Equivalence, power, and randomisation inference for the between-field contrast.

| Quantity                                                    | Value               |
|-------------------------------------------------------------|---------------------|
| Point estimate (across gap)                                 | + 0.07 pts          |
| 95% CI (cluster-robust)                                     | [-1.41, + 1.56] pts |
| SESOI (equivalence bound)                                   | ± 5 pts             |
| TOST <i>p</i> -value                                        | < .001              |
| Min. detectable effect (80% power), cluster-robust          | 2.12 pts            |
| Min. detectable effect (80% power), randomisation inference | 2.92 pts            |
| Randomisation-inference <i>p</i>                            | .942                |

### 4.8 Distributional effects (H3)

A concern frequently voiced about generative AI is not that it shifts the average performance but that it compresses the distribution, lifting weaker students and weaker programs toward a passing standard they would not otherwise reach while leaving the already-strong unaffected. This is the pooling prediction of Section 2.2, and a difference-in-differences on the mean is by construction blind to it, so we estimate the differential effect separately at the tenth, twenty-fifth, fiftieth, seventy-fifth, and ninetieth percentiles of the program pass-rate distribution. Were AI buoying the lower tail of high-exposure programs, the differential effect would be positive, largest at the low quantiles, and fading toward the top, and it is not. The estimated divergence is small, ranging from – 1.19 to + 1.55 points, and it does not decline monotonically across the distribution as the compression account requires, since the estimate at the tenth percentile is positive at + 1.55 but the estimate at the twenty-fifth is negative at -1.19 and the

profile thereafter is flat. A pooling collapse produces a monotone gradient, not an alternating one of this magnitude, so H3 is not supported.

Table 7  
Differential erosion by quantile of the pass-rate distribution.

| Quantile                                                                                                           | Differential erosion at quantile (pts) |
|--------------------------------------------------------------------------------------------------------------------|----------------------------------------|
| P10                                                                                                                | + 1.55                                 |
| P25                                                                                                                | -1.19                                  |
| P50                                                                                                                | + 1.27                                 |
| P75                                                                                                                | + 0.53                                 |
| P90                                                                                                                | -0.81                                  |
| Note. Quantile regressions include field and year fixed effects; coefficients are scaled to the full exposure gap. |                                        |

## 4.9 Robustness, corroboration, and placebo

The conclusion is insensitive to the analytic choices that might be thought to drive it, as Table 5 shows. Moving the AI boundary a year earlier or later leaves the estimate small and non-significant, replacing field fixed effects with the far more demanding program fixed effects, so that identification comes only from within-program change, scarcely moves it, and restricting attention to programs with at least twenty-five candidates, which removes the noisiest small-cohort observations, changes nothing of substance. Figure S4 reports a specification curve in which the estimate is recomputed across all thirty-six combinations of boundary year, fixed-effects structure, cohort-size floor, and field-specific trends, following Simonsohn et al. (2020). Every one of the thirty-six estimates falls within the region of practical equivalence, the median is a negligible – 0.56 points, and the most extreme estimate any defensible specification produces is -4.31 points, well short of the five-point threshold, so the null is a property of the data rather than of a fortunate modelling choice.

The between-field design invites a standing objection, namely that nursing and law differ in countless ways besides AI exposure and that some unobserved difference might mask an effect that is really there. We therefore test the same mechanism using variation entirely within nursing, where the confounds that distinguish professions are held constant by construction. Exploiting the considerable variation across programs in online-instruction intensity, on the premise that programs delivering more instruction at a distance afford students greater opportunity and fewer frictions to substitute AI for their own work, we estimate whether more-online programs experienced greater post-AI erosion, with program and year fixed effects, and the interaction is small and statistically indistinguishable from zero, as Table 8 reports. A parallel test within the law panel, using each school's own pre-AI average pass rate as a proxy for baseline quality, is reported in Supplementary Table S2, and because the regressor is constructed from the same schools' pre-period outcomes, ordinary regression to the mean produces a

positive coefficient whether or not anything happened in 2022, so we report it for its sign rather than its magnitude.

Table 8  
Within-field test of the erosion mechanism, nursing ( $n = 340$  program-years, program and year fixed effects).

| Within-field test                               | Coef  | SE   | $p$  | 95% CI          |
|-------------------------------------------------|-------|------|------|-----------------|
| Nursing: online-intensity gradient (per + 1 SD) | -0.13 | 0.77 | .864 | [-1.66, + 1.39] |

Finally, we ask whether the design would manufacture an effect where none could exist. Within the long law panel, which reaches back to 2008, we assign schools a pseudo-exposure equal to their standardised early-period mean pass rate over 2008 to 2013 and re-run the differential design at a sequence of pre-AI boundary years, where by construction generative AI cannot be responsible for any divergence. The result is more informative than a simple placebo null, and we report it as we find it. The placebo estimates are not negligible, since they drift monotonically from + 0.92 percentage points at a 2015 boundary to -1.44 at a 2021 boundary, at a near-constant - 0.395 points per year of boundary displacement, and several are strongly significant. This drift is exactly what the construction should produce, because the pseudo-exposure is built from each school’s own early-period outcomes, so mean reversion generates an increasingly negative apparent effect the further the boundary sits from the baseline window.

The diagnostic value lies in the comparison, which Figure S3 makes visually. Extrapolating the 2015 to 2021 placebo trend to a 2023 boundary predicts - 2.34 points, and the estimate at the genuine generative-AI boundary is -2.29 points, a departure of 0.05 points from what the mechanical drift alone anticipates. The design does generate spurious differentials from ordinary cross-sectional variation in school quality, and the estimate at the AI date is indistinguishable from the spurious series that variation produces, which is a stronger diagnostic than an assertion that placebo estimates are negligible.

We restate the constraints of Section 3.5 once rather than repeatedly. Exposure takes two values, so the effective number of treated clusters is one and the cluster-robust intervals in Tables 5 and 6 overstate precision, with randomisation inference the appropriate basis for inference, returning the same qualitative conclusion at lower precision. The nursing pre-period is two years long and both years are pandemic-affected, so parallel trends cannot be verified. The between-field analysis therefore corroborates the within-field analysis on the exposure dimension that the within-field design cannot address, and should be read as doing no more than that.

## 5. Discussion

The pattern of results adjudicates between the two accounts of the cost shock set out in Section 2.2. Had generative AI reduced the cost of signal production regressively in ability, the collapse from separation towards pooling would have left a specific and multiply observable trace, comprising

compression at the lower tail of the outcome distribution (H3), attenuation of the credential-to-criterion slope (H4), and a conditional shortfall in criterion performance for exposed cohorts (H5). The first two are testable here and neither appears. The differential effect across the outcome distribution is small and non-monotone rather than concentrated in the lower tail, and the multiple correlation between entering credentials and licensure performance is higher after 2022 than before. The predictive slope did contract, but roughly three-fifths of that contraction is the arithmetic consequence of a narrowing outcome distribution, and what remains lies inside the twenty-per-cent equivalence bound. The observations are consistent instead with a cost reduction that was proportional or ability-complementary, in that levels may have moved but the ordering the credential imposes on graduates by subsequent independent competence did not. This is not a claim that students used generative AI sparingly or well, but a claim about the distributional incidence of whatever use occurred, and it is the distributional incidence rather than the volume of use on which the signalling account turns.

Table 9  
Hypotheses, tests, and results.

| P1 | Prediction                         | Test                                                | Section    | Result                            |
|----|------------------------------------|-----------------------------------------------------|------------|-----------------------------------|
|    | Criterion is responsive            | Entering credentials to bar passage, three-year lag | 4.1        | Established                       |
| H1 | Differential erosion by exposure   | Between-field difference-in-differences             | 4.7        | Not supported                     |
| H2 | Erosion rises with dose            | Continuous dose specification                       | 4.3        | Not supported                     |
| H3 | Compression in the lower tail      | Quantile differential estimates                     | 4.8        | Not supported                     |
| H4 | Predictive relationship attenuated | Validity coefficient; slope change; TOST            | 4.2 to 4.4 | Not supported; equivalence at 20% |
| H5 | Conditional competence loss        | Value-added with trend                              | 4.6        | Not identified                    |

## 5.1 Reading the null through the validity argument

Read through the inferential chain of Section 2.4, the same result says that contamination at the scoring and generalisation inferences, if it occurred, did not propagate to extrapolation. That is a statement about the architecture of professional credentialing rather than about the integrity of any single assessment, because a system that places multiple imperfect screens in series is robust to the failure of one of them, and the terminal proctored screen is the one generative AI cannot presently reach. The finding therefore says nothing reassuring about whether individual pieces of coursework were completed honestly, and it should not be read as though it did. What it constrains is the downstream consequence, and that constraint is what the credentialing system exists to deliver. Two readings of the null nonetheless remain. On the first, licensure examinations continued to certify competence as before,

and the terminal checkpoint proved robust by design. On the second, they sit far downstream of the coursework whose integrity is in question, insulated by an intervening machinery of commercial preparation that may overwrite whatever variation in competence the degree imparted, so that the absence of a signal reflects an insensitive instrument rather than an undisturbed system. The second account requires that bar passage be only weakly related to what candidates bring to and take from their legal education, and Section 4.1 shows that it is not, since an instrument tracking its antecedent this closely would not fail to register a substantial degradation in the education occurring between them. The insensitivity reading is not eliminated, because preparation certainly compresses some variation and a small effect could still be absorbed, but it can no longer be offered as an equally plausible account. The resulting claim is bounded by Section 4.4, in that attenuation greater than about a fifth of the credential's predictive power is excluded while the evidence does not reach the fifteen-per-cent range at all.

## 5.2 What the timing implies

The strongest evidence against an AI-driven account is not the magnitude of the estimates but their dating. Two analyses that share no specification, namely the annual predictive slope and the residual against a pre-AI-fitted model, locate the single discontinuity in this twelve-year series in 2022. The cohort concerned matriculated in 2019 and had completed the substance of its legal education before ChatGPT was released, so whatever produced that break, generative AI cannot have. The candidates most plausible on the evidence available are the pandemic disruption to the 2020 and 2021 academic years and the coincident compression of between-school outcome dispersion documented in Section 4.3, which fell from 14.2 to 11.5 percentage points across the same window. The AI-exposed cohorts that follow do not extend the break; they recover from it. This matters for how the null should be read, because a flat series admits the objection that the design is simply insensitive to everything, whereas a series with one clearly located break that is not AI-timed demonstrates the opposite, namely that the design detects discontinuities when they occur and that the discontinuity it detects is not the one the erosion hypothesis predicts.

## 5.3 Implications: relocating the risk

Both readings converge on a single and practically important relocation of the problem. Whatever the mechanism, the locus of any genuine assessment-integrity risk from generative AI is not the terminal, proctored examination, which either resisted the tool or was buffered from it, but the upstream, unsupervised coursework that feeds the credential. The corollary is uncomfortable rather than reassuring, in that the system held at the point where institutions have invested most anxiety and most resource, while the exposure that remains is concentrated exactly where the redundancy does not operate, in the continuous, low-stakes, unsupervised assessment that constitutes most of a student's graded work and most of their learning.

This has direct implications for institutional response, because effort directed at securing high-stakes final examinations may be addressing the part of the system least at risk. The assessment-security literature has argued for some time that validity rather than detection is the right frame for institutional

policy (Dawson, 2021; Dawson et al., 2024), and the present result supports that position empirically, since the credential's predictive validity survived, which is what institutions should care about, while the integrity of individual unsupervised tasks remains unobserved and, on this evidence, unaddressed. One boundary on the claim follows from the design itself, in that a difference-in-differences identifies only the differential effect across exposure and is blind to erosion that struck all fields equally, so we can rule out that AI eroded high-exposure credentials more than low-exposure ones but not a uniform effect, though it is unclear what mechanism tied to language-model substitutability would produce one.

## 5.4 Limitations and future research

Several limitations qualify the result and mark the path forward. The two panels are unequally observed and the constraints on the between-field analysis are set out in Sections 3.5 and 4.9, so the within-field analysis carries the conclusion. The exposure measure, though external and validated, is a field-level proxy for what is ultimately an individual behaviour, namely students' actual use of AI on assessed work, which remains unobserved at scale. The chain from entering credentials to licensure performance is a proxy for the coursework-to-competence relationship rather than that relationship itself, and a design able to observe grades directly would be sharper. Three post-AI outcome years cannot identify a conditional level shift, as Section 4.6 makes plain, and the equivalence bound is twenty per cent rather than zero, so attenuation below roughly a fifth of the predictive slope lies beyond the resolution of these data and only a longer post-AI series will close that gap. All estimates are school- or program-level aggregates, with the interpretive constraints noted in Section 4.1.

Three extensions follow directly. Two fields, however maximally contrasting, define only the endpoints of the exposure gradient and cannot trace its shape, so extending the design to intermediate-exposure professions for which institution-level licensure outcomes are published nationally, among them teacher preparation, accountancy, pharmacy, and engineering, would convert the present two-point contrast into a dose-response curve and would simultaneously raise the number of treatment clusters to a level at which conventional inference becomes defensible. Second, the analysis is confined to a single national credentialing architecture, so replication in jurisdictions with comparable published outcomes, such as the England and Wales Solicitors Qualifying Examination or Canadian NCLEX administration, would test whether the finding is a property of credentialing systems in general or of this one in particular. Most fundamentally, the interpretive account that remains open points toward the decisive next study, namely a design that measures the credential-to-competence relationship closer to the coursework, for instance by comparing the same students' performance on matched AI-permissive and proctored assessments of the same learning outcomes. We note, finally, that the analysis is exploratory and built on secondary data, so its conclusions are best regarded as a well-powered, falsifiable benchmark for a question of growing consequence, and an invitation to the confirmatory and coursework-proximal work it frames.

## 6. Conclusion

A premise of academic assessment, that submitted work signals independent capability, has been called into question by the arrival of generative AI, and with it the value of the credentials that assessment

produces. Stated precisely, the concern is that a technology lowering the cost of signal production would collapse a separating equilibrium and break the extrapolation inference on which credential use depends. Testing the falsifiable forms of that claim, we find that within the more exposed profession the credential-to-competence relationship held, at a precision sufficient to exclude attenuation greater than a fifth, with the larger part of the apparent contraction in the raw predictive slope traceable to a narrowing outcome distribution rather than to any loss of association. We find that across the widest available contrast in AI exposure no differential erosion appeared, on average or at any quantile. And we find that the one discontinuity in a twelve-year series belongs to a cohort educated before generative AI existed. Critically, the criterion on which these nulls rest is demonstrably responsive rather than inert, which is what allows them to be read as evidence rather than as silence.

The evidence is consistent with a cost shock that was proportional or ability-complementary rather than regressive, and with a credentialing architecture whose redundancy absorbed contamination of a single inferential link. What remains genuinely open is the interval these data cannot reach, comprising small erosion below our equivalence bound, and any effect concentrated in the unsupervised coursework that lies between the credentials we observe and the examination we observe. That is where we would direct the next study, and where institutions concerned about assessment integrity may most usefully look.

## **Declarations**

## **Author Contribution**

Rasheed Mohammad: Conceptualization; Methodology; Software; Formal analysis; Data curation; Investigation; Validation; Visualization; Writing - original draft; Writing - review and editing; Project administration. Doaa M. Bamasoud: Conceptualization; Writing - review and editing. Mohammed Ayid Alqahtani: Conceptualization; Resources; Writing - review and editing. All authors read and approved the final manuscript.

## **Data Availability Statement**

The data that support the findings of this study were derived from the following sources, all available to other researchers under the access conditions noted. School-level first-time bar-passage rates, median LSAT scores, and median undergraduate grade point averages derive from the American Bar Association's Standard 509 Required Disclosures, compiled and distributed through AccessLex Analytix (<https://analytix.accesslex.org>). The underlying disclosures are public and the compiled longitudinal datasets are available without charge following free registration, while the original per-school disclosures are also openly available via the ABA Required Disclosures portal (<https://www.abarequireddisclosures.org>). Program-level first-time NCLEX-RN pass rates were compiled from publicly available reports published by United States state boards of nursing, openly accessible on the respective state board websites, with national and state-level figures published by the National

Council of State Boards of Nursing used for contextual cross-checking (<https://www.ncsbn.org>). Program and institution characteristics used in constructing and describing the nursing panel were obtained from the U.S. Department of Education's Integrated Postsecondary Education Data System (<https://nces.ed.gov/ipeds>). Field-level generative-AI exposure was measured using the Language-Modelling AI Occupational Exposure index of Felten, Raj, and Seamans (2023), openly available at <https://github.com/AIOE-Data/AIOE>. Analysis code sufficient to reproduce every table and figure is available from the corresponding author.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude (Anthropic) in order to assist with language editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## References

1. Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2), 3–30.
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
3. Arrow, K. J. (1973). Higher education as a filter. *Journal of Public Economics*, 2(3), 193–216.
4. Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3), 3–30.
5. Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279–1333.
6. Bar, T., Kadiyali, V., & Zussman, A. (2009). Grade information and grade inflation: The Cornell experiment. *Journal of Economic Perspectives*, 23(3), 93–108.
7. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). *Generative AI can harm learning* [Working paper].
8. Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.
9. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*.
10. Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.
11. Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372.

12. Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.
13. Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT goes to law school. *Journal of Legal Education*, 71(3), 387–400.
14. Conley, T. G., & Taber, C. R. (2011). Inference with “difference in differences” with a small number of policy changes. *Review of Economics and Statistics*, 93(1), 113–125.
15. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
16. Curtis, G. J., & Clare, J. (2017). How prevalent is contract cheating and to what extent are students repeat offenders? *Journal of Academic Ethics*, 15, 115–124.
17. Dawson, P. (2021). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge.
18. Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*.
19. Dell’Acqua, F., McFowland, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier* (Working Paper 24 – 013). Harvard Business School.
20. Ellis, C., Zucker, I. M., & Randall, D. (2018). The infernal business of contract cheating: Understanding the business processes and models of academic custom writing sites. *International Journal for Educational Integrity*, 14, 1.
21. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of large language models. *Science*, 384(6702), 1306–1308.
22. Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12), 2195–2217.
23. Felten, E., Raj, M., & Seamans, R. (2023). *How will language modelers like ChatGPT affect occupations and industries?* arXiv. <https://arxiv.org/abs/2303.01157>
24. Ferman, B., & Pinto, C. (2019). Inference in differences-in-differences with few treated groups and heteroskedasticity. *Review of Economics and Statistics*, 101(3), 452–467.
25. Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). ChatGPT has aced the test of understanding in college economics: Now what? *The American Economist*, 68(2), 233–245.
26. Jensen, L. X., Buhl, A., Sharma, A., & Bearman, M. (2025). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher Education*, 89, 1145–1161.
27. Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. Springer.
28. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

29. Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270), 20230254.
30. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198.
31. Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.
32. Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269.
33. Liang, W., Yuksekogonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779.
34. Lodge, J. M., Howard, S., Bearman, M., Dawson, P., & Associates. (2023). *Assessment reform for the age of artificial intelligence*. Tertiary Education Quality and Standards Agency.
35. MacKinnon, J. G., & Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2), 233–254.
36. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
37. Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.
38. Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Frontiers in Education*, 3, 67.
39. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192.
40. OpenAI. (2022, November 30). *Introducing ChatGPT*. <https://openai.com/index/chatgpt/>
41. Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(6).
42. Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5), 2555–2591.
43. Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357.
44. Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record*, 114(7), 1–23.
45. Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3), 305–322.
46. Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85(1), 112–118.

47. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected?* arXiv. <https://arxiv.org/abs/2303.11156>
48. Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A “Turing Test” case study. *PLOS ONE*, 19(6), e0305354.
49. Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4, 1208–1214.
50. Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, 23(2), 226–235.
51. Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.
52. Stiglitz, J. E. (1975). The theory of “screening,” education, and the distribution of income. *American Economic Review*, 65(3), 283–300.
53. Strathern, M. (1997). “Improving ratings”: Audit in the British university system. *European Review*, 5(3), 305–321.
54. Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. Wiley.
55. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, 26.
56. Weiss, A. (1995). Human capital vs. signalling explanations of wages. *Journal of Economic Perspectives*, 9(4), 133–154.
57. Wightman, L. F. (1998). *LSAC national longitudinal bar passage study*. Law School Admission Council.

## Figures

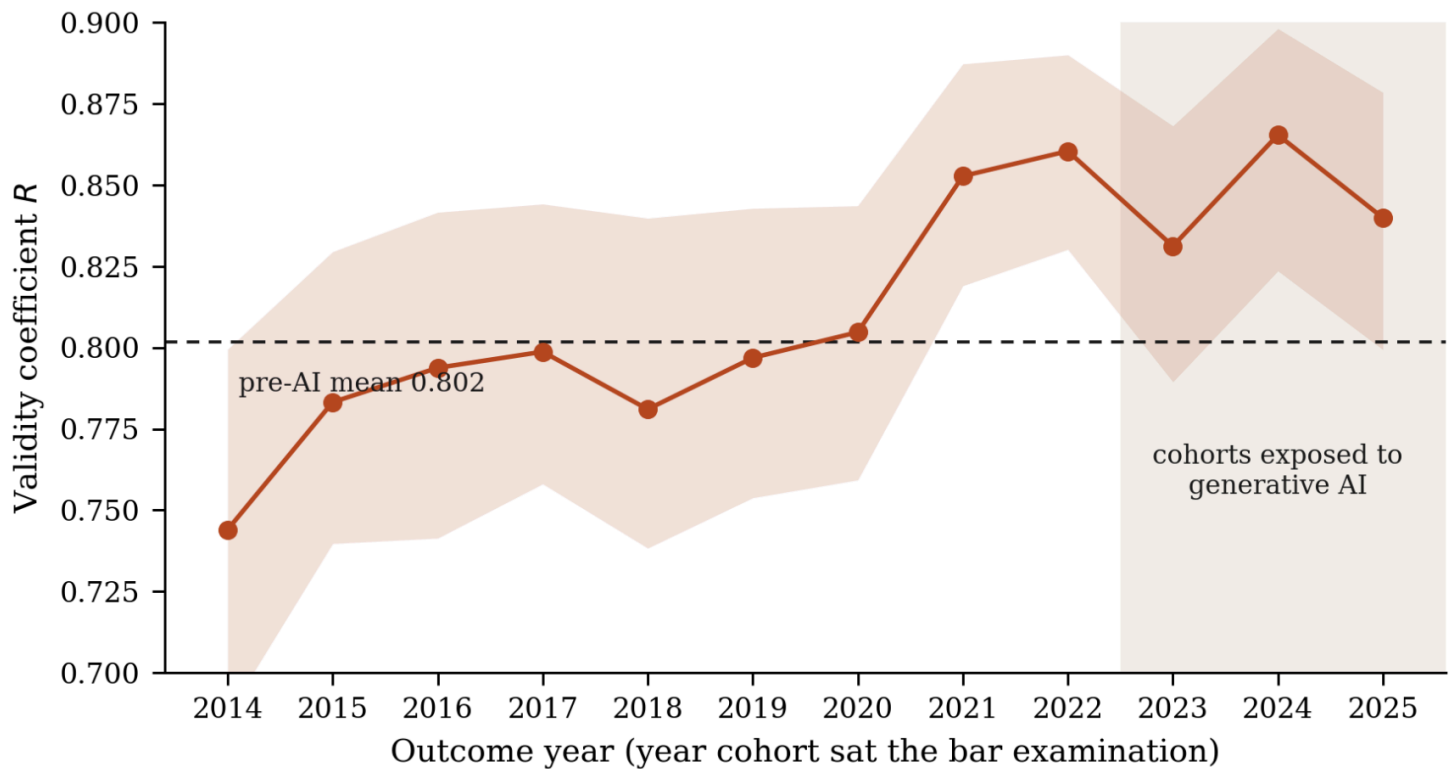


**Figure 1**

Validity coefficient by outcome year, 2014 to 2025. The coefficient is the multiple correlation between entering-cohort median LSAT and median undergraduate grade point average and that cohort's first-time bar passage three years later. The shaded band marks cohorts whose legal education overlapped the availability of generative AI, the ribbon shows school-clustered bootstrap 95 per cent intervals, and the dashed line marks the pre-AI mean of 0.802.

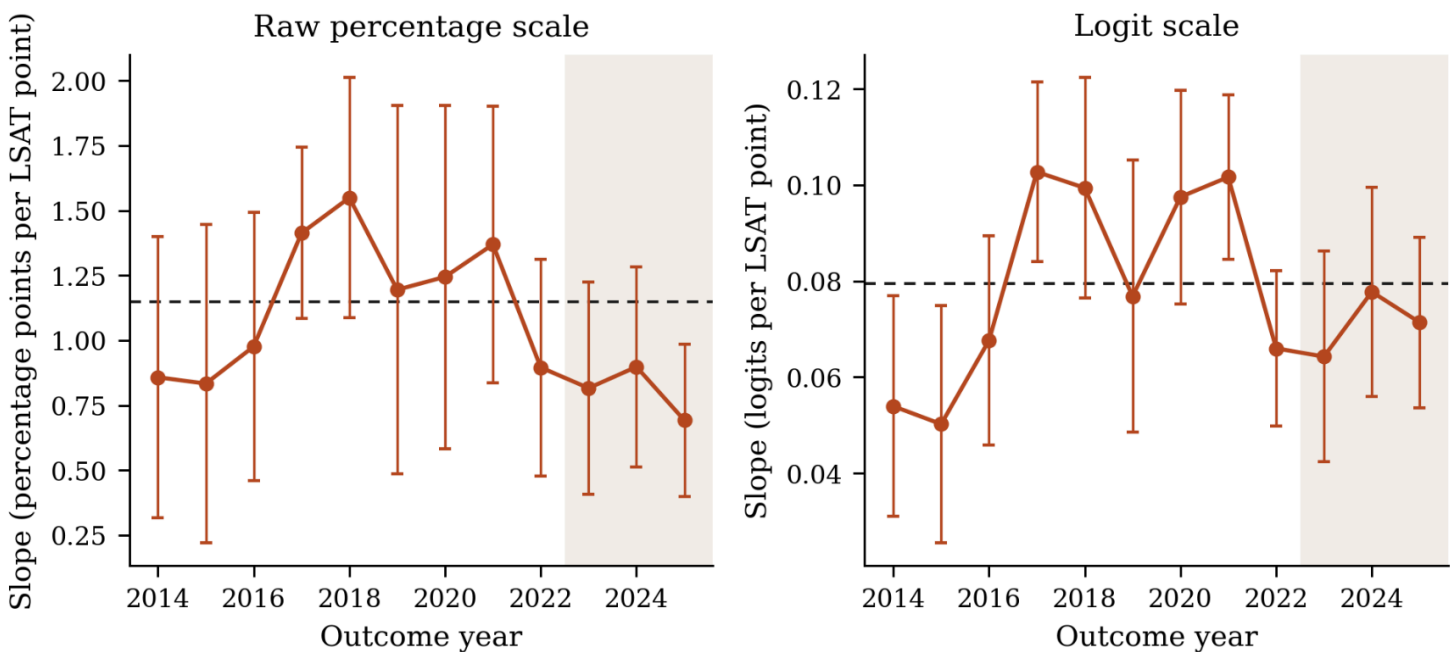


Figure 2

Annual predictive slope with 95 per cent confidence intervals, estimated on the raw percentage scale (left) and the logit scale (right). Each point is the partial coefficient on entering median LSAT from a within-year regression that also conditions on median undergraduate grade point average. The dashed line marks the pre-AI mean and the shaded band marks the AI-exposed cohorts.

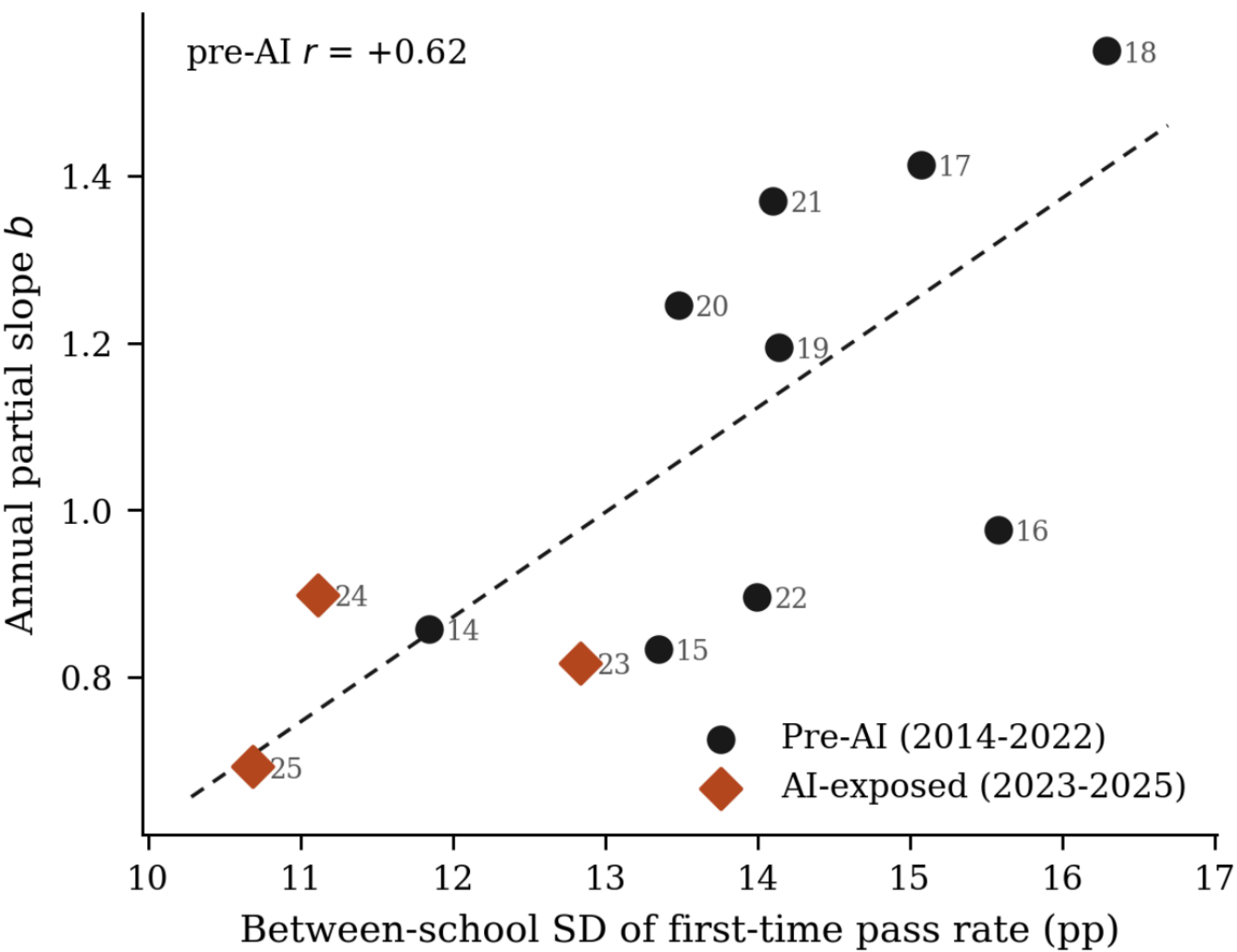


Figure 3

The annual partial predictive slope plotted against the between-school standard deviation of first-time pass rates, with each point labelled by outcome year. The two series co-move at a correlation of +0.62 across the pre-AI period, locating the larger part of the raw-scale attenuation in dispersion contraction rather than in loss of predictive power.

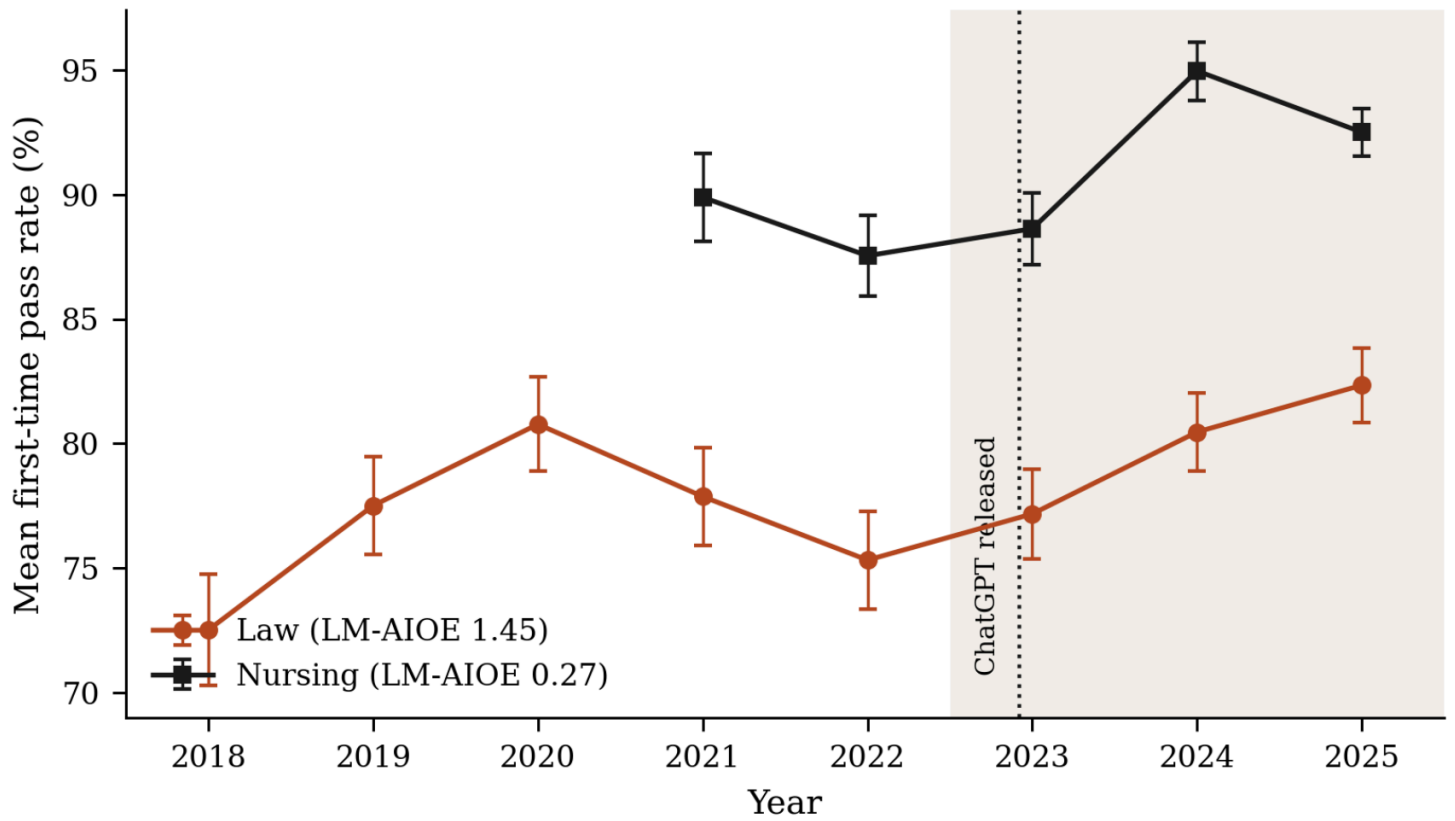


**Figure 4**

Mean first-time licensure pass rates by field, 2018 to 2025. Error bars denote 95 per cent confidence intervals, the dotted line marks the public release of ChatGPT in November 2022, and the shaded band marks the post-AI period. Law programs enter the panel in 2018 and nursing programs in 2021.

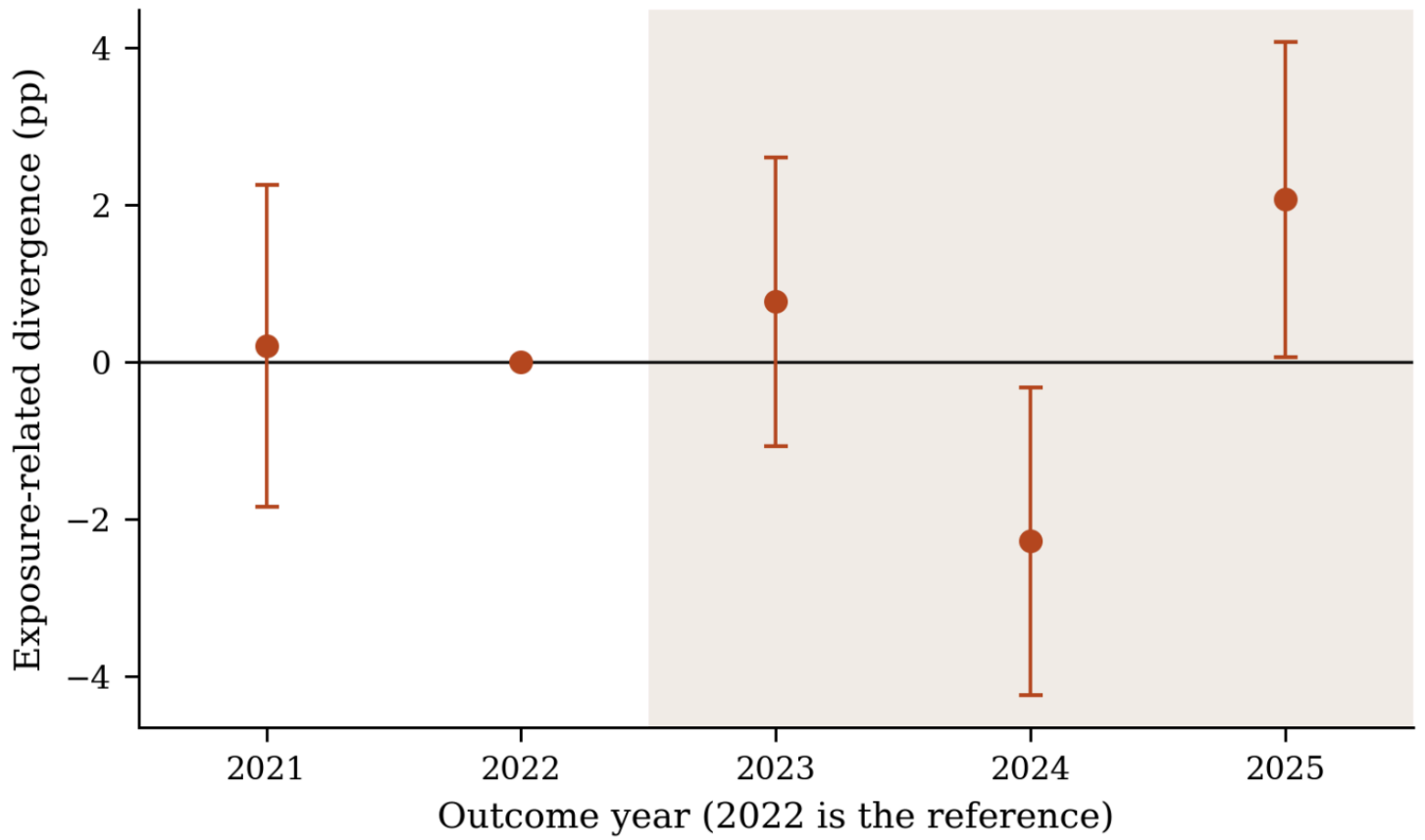


**Figure 5**

Event-study estimates of exposure-related divergence by outcome year, relative to 2022, with 95 per cent confidence intervals. Only years in which both fields are observed are shown, because exposure-related divergence is not identified before nursing enters the panel in 2021. The shaded band marks the AI-exposed period.

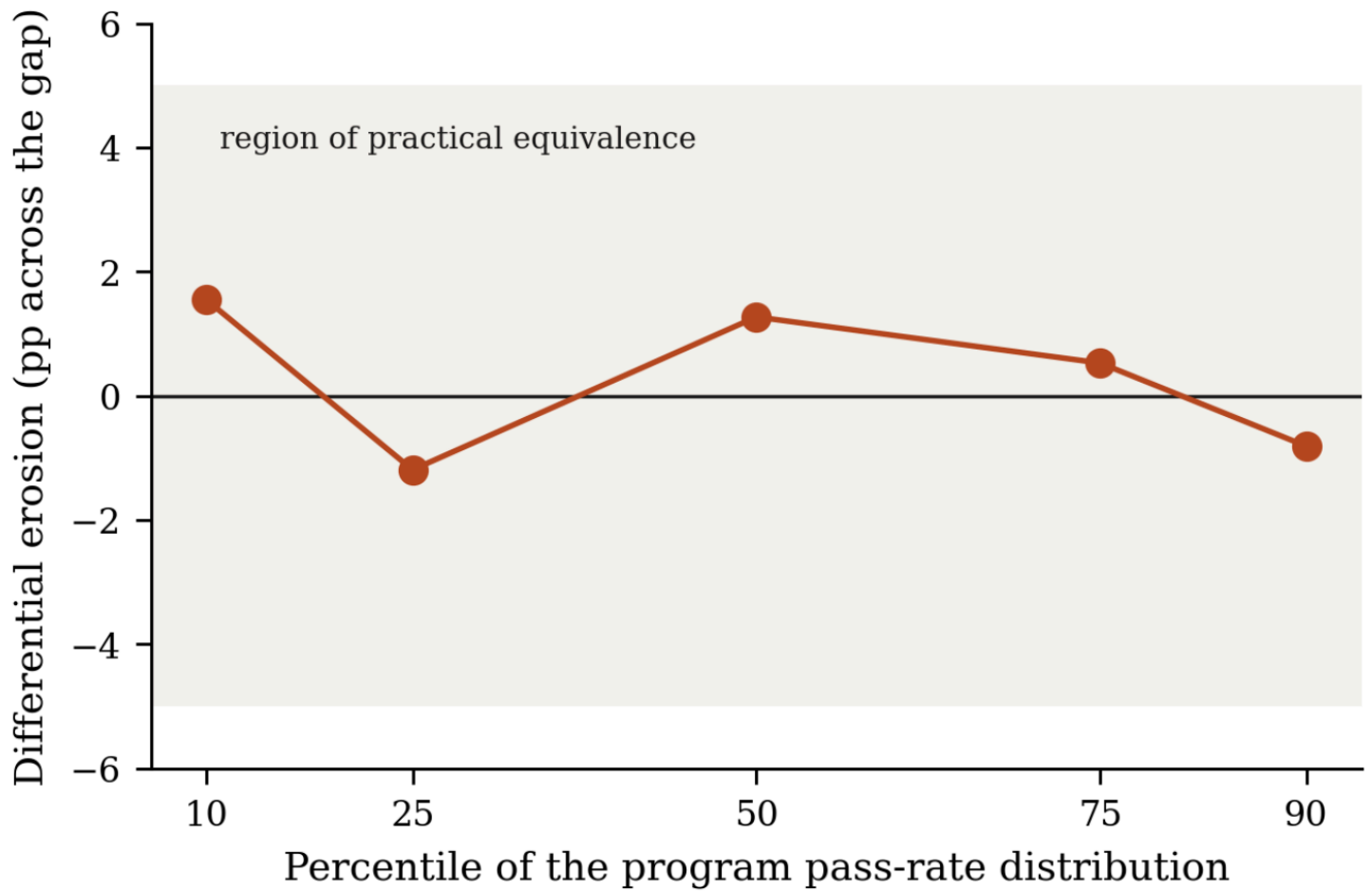


Figure 6

Distributional test: differential erosion estimated at each quantile of the program pass-rate distribution, with the region of practical equivalence shaded. A flat and non-monotone profile indicates no compression of the kind the pooling account predicts.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixandSupplementaryFigures.docx](#)