Informative Grading Requires Cross-Course Comparability*

Joshua S. Gans† ⓡ Scott Duke Kominers‡

March 13, 2026

Abstract

We study the information problem underlying academic grading. A grade collapses student ability and course difficulty into a single scalar signal, creating a cross-course comparability problem. For curricula rich enough to realise easy-course/hard-course reversals, no universally calibrated threshold grading rule can guarantee informative comparisons of student ability from a single grade alone. We examine three partial resolutions. First, under exogenous (random) course assignment and a monotone-likelihood-ratio condition on the difficulty-marginalised performance density, grades remain statistically informative about expected ability, although this property can fail when course selection is endogenous. Second, multi-course transcripts restore identification in an additive latent-performance model: with sufficient overlap in enrolments, the student–course outcome matrix identifies student ability and course difficulty up to normalisations, and in the complete-data benchmark an "eigengrade" construction gives a meaningful aggregate comparison statistic. Third, we analyse strategic grade inflation in a stylised game among instructors, drawing on an analogy to monetary inflation, and show that a Taylor-rule–style feedback mechanism can support a target symmetric equilibrium while preserving within-course variation.

Journal of Economic Literature Classification: D82, I23, C13, D47.

Keywords: Grade inflation, impossibility theorem, identification, eigengrades, spectral methods, mechanism design.

*We are grateful to Jason Furman, David Laibson, Jeff Miron, Jesse Shapiro, and especially Yannai Gonczarowski, Jack Hirsch and Shengwu Li (who, to our knowledge, originated the term "eigengrade"), for conversations that shaped this paper. We also thank Refine.ink, ChatGPT 5.{2, 4} Pro, Claude Opus 4.6, Claude Sonnet 4.5, and Gemini 3 for research assistance; Kominers primarily accessed ChatGPT, Claude, and Gemini through Poe with the support of Quora, where he is an advisor. Responsibility for all errors remains our own. Kominers is both a faculty member and an alum of Harvard University, whose policy conversation around grade inflation in part inspired this paper; and he has expressed views on this question within Harvard based on the analysis found herein.

†Rotman School of Management, University of Toronto, and NBER. Email: joshua.gans@utoronto.ca.

‡Harvard Business School; Department of Economics and Center of Mathematical Sciences and Applications, Harvard University; and a16z crypto. Email: kominers@fas.harvard.edu.

1 Introduction

In early 2026, Harvard University publicly debated a proposal to cap the fraction of "A" grades awarded in each course in response to concerns about grade inflation. The policy provoked immediate debate both within Harvard[1] and among the broader academic public. Blanchard drew an analogy between grade inflation and price inflation, arguing that instructors who attempt to raise their relative grades generate an externality analogous to firms raising relative prices.[2] More generally, the macroeconomic comparison is to rule-based inflation stabilisation in the spirit of Taylor (1993). Later in the paper we distinguish sharply between a literal distributional cap and a Taylor-rule–style feedback on course mean grades. On the other side, critics noted that a blunt cap destroys information: if a course is genuinely full of excellent students, forcing a grade distribution that does not reflect this is both unfair and informationally wasteful.

This paper takes that debate as a starting point for a deeper investigation. Using economic theory, we argue that the central difficulty in academic grading is not merely inflation, understood as the upward drift of average grades over time (Rojstaczer and Healy, 2012; Johnson, 2003), but an *identification problem*. A letter grade is a one-dimensional projection of a two-dimensional object: the interaction between student ability and course difficulty. An outside observer who sees only the grade cannot, in general, distinguish a highly able student in a difficult course from a less able student in an easy one.

The paper makes three distinct claims that should be kept separate. First, a single grade alone cannot generally identify student ability across heterogeneous courses. Second, under additional assumptions, grades can still be informative in a weaker Bayesian sense. Third, with transcript-level overlap and an additive latent-performance structure, one can recover difficulty-adjusted student effects from multi-course data. The force of each claim depends on a different set of assumptions.

We introduce a simple model of grading as a function of student ability and course difficulty, and examine several margins along which a grading system may remain informative once full pointwise comparability fails. We begin with an impossibility result for a single, universally calibrated scalar grade. The sharp statement requires a curriculum rich enough to realise easy-course/hard-course reversals and is most naturally stated for threshold grading rules, the canonical class of monotone scalar grading rules. On such curricula, a lower-ability student in an easy course can receive a higher grade than a higher-ability student in a hard

[1] *The Harvard Crimson*'s Opinion Section convened a series of editorials and columns on the proposal, some supportive and some opposed. Full citations should be provided in the bibliography rather than as raw URLs in footnotes.

[2] Public post on $\mathcal{X}$, 9 February 2026. A full citation should be provided in the bibliography.

course, so grades alone cannot guarantee correct cross-course ability comparisons.

This impossibility should be read as a limitation on *grade-only* inference. Once course labels or full transcripts are observed, the relevant question is no longer whether a single naked grade suffices, but how much course-level context is needed to restore informative comparisons.

Two weaker, decision-relevant notions survive the impossibility, at least partially. The first is *statistical sufficiency*: under exogenous course assignment and an MLRP condition, higher grades imply higher posterior expected ability on average, though this property can fail under endogenous selection. The second is *transcript sufficiency*: with enough course overlap, multi-course transcripts identify course difficulty and student ability in an additive latent-performance model.

Following the setup of the model in Section 2, Section 3 establishes the impossibility precisely. We show that no universally calibrated non-trivial threshold grading rule can guarantee ability-sufficient cross-course comparisons on curricula rich enough to generate reversal witnesses. The argument does not depend on parametric distributional assumptions; it is driven by the elementary fact that a scalar grade pools ability and difficulty.

Section 4 explores avenues for partially resolving the impossibility. First, we weaken the sufficiency requirement from pointwise to statistical, asking only that higher grades correspond to higher expected ability. We show that this weaker condition holds under exogenous (random) course assignment but can fail under endogenous selection, such as when ambitious students sort into harder courses (Section 4.1). We emphasise that this result concerns the difficulty-marginalised performance density; the paper provides sufficient conditions under which the required monotone-likelihood-ratio property is inherited from the underlying latent-performance structure.

We then turn to transcript-level information (Section 4.2). When an observer sees a student's full vector of grades across courses, the identification problem can be addressed in an additive latent-performance model by exploiting the cross-sectional structure of the student–course matrix. We formalise the notion of *eigengrades*, a spectral score extracted from the student affinity matrix after projecting out an unavoidable all-ones component, and show that in the complete-data noiseless benchmark the eigengrade construction is a spectral re-expression of centred ability. This benchmark result is best understood as an identification and representation theorem. For noisy or incomplete latent-score data we provide an oracle perturbation bound for the spectral recovery step, and for observed ordinal grades we show how an ordered-response first stage can, under additional assumptions, be translated into the eigengrade direction. We do not claim that these results by themselves constitute a full feasible estimation theory for sparse transcript data with raw ordinal grades.

We then consider the strategic environment that may support grade inflation (Section 5). We study a stylised game among self-interested instructors in which mean grades generate a competitive externality. Drawing on the Blanchard inflation analogy, we analyse a Taylor-rule–style feedback mechanism (Taylor, 1993). Our result is intentionally limited: the mechanism can support a target symmetric equilibrium while preserving within-course variation. The section is therefore best read as a tractable benchmark for thinking about feedback-based grading policy, rather than as a complete dynamic theory of institutional stabilisation.

Section 6 discusses the implications for policy design, including the interaction of our within-school grading mechanisms with the between-school matching-market framework of Ostrovsky and Schwarz (2010), and extensions to multi-dimensional ability profiles, where we separate scalar baseline ability from comparative-advantage profiles and emphasise that transcript data recover a skill subspace rather than a canonical universal ranking absent additional structure.

Directional guidance for grading policy. Even for institutions that do not adopt the full transcript-based machinery developed here, the analysis yields several design principles. First, distributional grade caps have a clear informational cost: they make the interpretation of a given grade depend more heavily on the enrolled cohort and can compress within-course variation. Second, if inflation is the policy concern, institutions should focus on course-level mean grades rather than rigid within-course grade shares, and should report contextual information about course difficulty alongside grades so that observers can condition on course identity when forming inferences. Third, curricular breadth requirements and shared core courses serve an informational function that is easy to overlook: they create the cross-course enrolment overlap needed for transcript-based difficulty adjustment, whether implemented through eigengrades or simpler methods. Finally, preserving within-course grade variation is essential; policies that compress the spread of grades within a course destroy signal both for direct observers and for transcript-based identification methods.

1.1 Related Work

The paper contributes to the literature on grade inflation and its consequences (Sabot and Wakeman-Linn, 1991; Bar et al., 2009; Popov and Bernhardt, 2013; Chan et al., 2007), to the information economics literature on the design of signals and certificates (Spence, 1973; Blackwell, 1953), and to the growing literature on spectral methods for ranking and identification (Negahban et al., 2017).

The eigengrade construction, in particular, belongs to a broader family of spectral methods for recovering latent rankings from matrices of pairwise observations. PageRank (Page et al.,

1999) identifies webpage importance from the principal eigenvector of the link matrix; rank centrality (Negahban et al., 2017) recovers competitor strength from the principal eigenvector of the comparison matrix in Bradley–Terry–Luce environments (Bradley and Terry, 1952; Luce, 1959); and the Pinski–Narin method (Pinski and Narin, 1976) ranks journals by a similar recursive logic. All exploit the same spectral ranking logic: a square matrix encoding relational comparisons can reveal a latent one-dimensional ranking through its leading eigenvector. In our setting the primitive data are bipartite student–course observations, and the relevant square object is the induced student affinity matrix.

A particularly close analogue arises in the paired-comparison literature on rating competitors in games and sports. The classic Elo system (Elo, 1978) and its Bayesian extensions (Glickman, 1999) estimate player strength from match outcomes, treating each match as a pairwise comparison. These systems face a version of our identification problem when draws are common: if the best players tie most of their games, match outcomes carry little information about relative strength. Glickman (2026) recently addressed this for chess by developing a model in which the probability of a draw depends on player strength. The analogy to our setting is in spirit rather than in architecture: in both cases, a richer observation model allows the analyst to extract more information from the same raw data, even though our transcript problem involves a two-way student–course latent structure rather than a direct paired-comparison likelihood.

2 Model

We consider a university with a finite set of students $S = \{1, \dots, n\}$ and a finite set of courses $C = \{1, \dots, m\}$. Each student i is characterised by an ability $a_i \in \mathbb{R}$ and each course c by a difficulty $d_c \in \mathbb{R}$. The ability and difficulty parameters are unobserved by the outside observer (an employer, a graduate school admissions committee, etc.) who must form inferences about student quality from grades.

Although we index students and courses by finite sets for notation, the axioms and the impossibility results are intended as describing *uniform* limits on the informativeness of grading at the institution level: the question is whether a single scalar grading standard can guarantee cross-course comparability across the class of curricula with sufficiently heterogeneous difficulty (see Remark 1). Instances with essentially no difficulty heterogeneity are therefore outside the intended scope.

2.1 Performance

A **performance function** $\pi : \mathbb{R}^2 \to \mathbb{R}$ maps ability–difficulty pairs to a real-valued performance outcome, which we can think of as an aggregate "score" in a course reflecting, for example, performance on assignments, exams, and in-class activities. Rather than treat monotonicity of performance as an independent assumption, we build it into the definition of the object:

Definition 1 (Performance Function)**.** A function $\pi : \mathbb{R}^2 \to \mathbb{R}$ is a *performance function* if $\pi(a, d)$ is strictly increasing in a and strictly decreasing in d.

Definition 1 imposes the minimal requirement that higher ability improves performance and greater difficulty reduces it, holding the other term fixed.

We impose one additional structural condition on π which captures the idea that high performance can sometimes be meaningfully easier to achieve in an easier course. While one could require a strong *full range* property under which, for every $a \in \mathbb{R}$ and every p in the range of π, there exists $d \in \mathbb{R}$ with $\pi(a, d) = p$ (as in additive or multiplicative models),[3] the information tension uncovered in our results can instead rely on the following much weaker condition:

Assumption 1 (Weak Overlap)**.** For any performance levels $p_H > p_L$ in the range of π, there exist ability levels $a_H > a_L$ and course difficulties $d_H, d_L \in \mathbb{R}$ such that $\pi(a_L, d_L) \geq p_H$ and $\pi(a_H, d_H) \leq p_L$.

Assumption 1 is a richness condition on the performance technology, not yet on any realised finite curriculum. It says that for any two performance levels in the range of π, there exist ability and difficulty values capable of generating an easy-course/hard-course reversal across them. Whether a particular institution's realised course set and student body actually contain such a witness is a separate, operational question taken up after Theorem 1.

2.2 Grading

A **grading system** is a collection of course-specific grading rules $\{g_c : \mathbb{R} \to G\}_{c \in C}$ mapping student performance to grades, where $G = \{g_1 < g_2 < \cdots < g_K\}$ is a finite ordered set of grade levels (e.g., $G = \{F, D, C, B, A\}$ with the usual ordering). When student i takes course c, the grade assigned is $g_c(\pi(a_i, d_c))$.

[3]The full range property is satisfied, for instance, by any additive model $\pi(a, d) = a - d$ or any multiplicative model $\pi(a, d) = a/d$ (for positive arguments), or more generally by any function that is surjective in each argument conditional on the other.

A natural baseline requirement is for the grading rule to be monotone in performance levels within a given course, i.e.:

Axiom 1 (Within-Course Monotonicity)**.** For all $c \in C$ and all $i, j \in S$: if $a_i > a_j$, then $g_c(\pi(a_i, d_c)) \geq g_c(\pi(a_j, d_c))$.

Within-course monotonicity means that in any given course, a more able student receives a weakly higher grade.

Under the point-wise monotonicity of the performance function assumed in Definition 1, Axiom 1 implies that, within any individual course, the grading map is weakly increasing in performance on the set of performance levels attainable in that course. In particular, whenever two performance levels $p \geq p'$ are both attainable *within the same course*, the higher performance must receive a weakly higher grade.

Because G is a finite ordered set, any weakly increasing mapping from performance to grades can be represented by performance-level cutoffs. We therefore focus on *threshold grading rules* as a canonical representation of monotone scalar grading:

Definition 2 (Threshold Grading Rule)**.** A grading rule $g : \mathbb{R} \to G$ is a *threshold grading rule* if there exist cutoffs $t_0 = -\infty < t_1 < t_2 < \cdots < t_{K-1} < t_K = \infty$ such that, for every attainable performance level $p \in \text{rng}(\pi)$,

$$g(p) = g_k \quad \text{if and only if} \quad p \in [t_{k-1}, t_k).$$

The threshold structure is a canonical right-continuous representation of a weakly increasing finite-valued grading map on the attainable performance set. Boundary tie-breaking conventions (for example, left- versus right-continuity exactly at a cutoff) are immaterial for our purposes. What matters is that higher attainable performance levels never receive lower grades. Accordingly, throughout the impossibility results we use "threshold rule" as convenient shorthand for a universally calibrated grading rule that is weakly increasing on $\text{rng}(\pi)$ and represented by performance cutoffs up to harmless boundary conventions.

2.3 Calibration

There is also often a strict structural design constraint on the grading rule itself, which captures the idea that grades are meant to be cross-course comparable: an "A" in one course represents the same performance standard as an A in another. This is what enables outside observers to compare grades across courses and aggregate them into aggregate statistics like grade-point averages.

Axiom 2 (Universal Calibration)**.** For all courses $c, c' \in C$ and all performance levels $p \in \text{rng}(\pi)$, $g_c(p) = g_{c'}(p)$. Equivalently, there exists a single function g such that $g_c \equiv g$ for all c; in the case of a threshold grading rule, this means that the performance thresholds $t_1 < t_2 < \cdots < t_{K-1} < t_K$ do not depend on c.

In practice, universal calibration corresponds to an institutional mandate of absolute grading standards—such as a rigid university-wide rule that a raw score of >90% always yields an 'A', or a standardized qualitative rubric; for example, the University of Toronto states that an 'A' grade is awarded when there is "[s]trong evidence of original thinking; good organization; capacity to analyze and synthesize; superior grasp of subject matter with sound critical evaluations; evidence of extensive knowledge base."[4] Crucially, the universal calibration requirement implicitly forbids instructors from unilaterally "grading on a curve"—i.e., adjusting their specific grade thresholds to compensate for the intrinsic difficulty of their assessments.

Universal calibration (Axiom 2) specialises our general environment to the calibrated case $g_c \equiv g$ for all c, allowing us to write a single function g without ambiguity. Without universal calibration, the *axiomatic conflict* behind Theorem 1 dissolves: each course can use its own standard g_c, turning grades into purely within-course ordinal rankings. But at the same time, of course, abandoning universal calibration also destroys cross-course comparability and typically makes it harder, not easier, for outside observers to infer ability without observing the course-specific grading standards.

2.4 Design Objective

Finally, we introduce an overarching (or institutional) objective for the grading system—namely, that grades should be meaningfully interpretable across courses as indicators of ability.

Axiom 3 (Ability Sufficiency)**.** A grading system $\{g_c : \mathbb{R} \to G\}_{c \in C}$ is *ability-sufficient* if for all students $i, j \in S$ and all courses $c, c' \in C$,

$$g_c(\pi(a_i, d_c)) > g_{c'}(\pi(a_j, d_{c'})) \quad \Longrightarrow \quad a_i \geq a_j.$$

Under universal calibration (Axiom 2), there exists a single function g with $g_c \equiv g$ for all c, and Axiom 3 definition reduces to $g(\pi(a_i, d_c)) > g(\pi(a_j, d_{c'})) \Rightarrow a_i \geq a_j$.

[4] https://advice.writing.utoronto.ca/general/grading-policy/. Universal calibration is also very common in the UK system, where a 70%+ is often a 'First Class' degree across all subjects (see, e.g., https://gostudyin.com/india/study-in-uk/study-guides/uk-university-grading-system-explained/ for discussion).

Ability sufficiency is a strong informational goal, but also one that matches a core comparability assumption underpinning school transcripts: Under ability sufficiency, an observer who sees that student i's reported grade in course c exceeds student j's reported grade in course c' can infer that i is at least as able as j. Formally, this inference is required to be valid for the realised grades $g_c(\pi(a_i, d_c))$ and $g_{c'}(\pi(a_j, d_{c'}))$, even when the underlying performance-to-grade mappings differ across courses.

Note that standard intuitions about how grading works may make universal calibration and ability sufficiency seem, at first glance, like they are closely related criteria. But in fact, they are quite distinct—and as we show soon, they are to some degree at odds with each other. The issue is that universal calibration imposes a uniform grading as a function of student performance, but achieving the same performance level in courses of different difficulty levels indicates very different levels of ability.

3 The Impossibility Theorem

Before deriving our main impossibility theorem, we must first exclude one degenerate case: the grading rule that assigns everyone the same grade, which satisfies our main axioms trivially by collapsing all distinctions in performance.

Axiom 4 (Non-Triviality)**.** The grading rule separates at least two attainable performance levels: there exist (a, d) and (a', d') such that $\pi(a, d) > \pi(a', d')$ and $g(\pi(a, d)) > g(\pi(a', d'))$. Equivalently, $|g(\mathrm{rng}(\pi))| \geq 2$.

With non-triviality in place, we can sharpen the central question: *Can a single calibrated scalar grade guarantee correct cross-course ability comparisons?* A raw performance reversal is not yet enough for a contradiction. If the reversed performances all lie in a range on which g is constant, then the grading rule reports the same grade and no violation of ability sufficiency follows. What matters is a reversal that crosses a grade boundary. We, therefore, separate the logical mechanism of the impossibility from the primitive conditions that make such a witness available.

Definition 3 (Grade-separated reversal)**.** Fix a universally calibrated grading rule g. The realised curriculum contains a *grade-separated reversal* for g if there exist performance levels $p_H > p_L$ with $g(p_H) > g(p_L)$, students $i_H, i_L \in S$, and courses $c_H, c_L \in C$ such that

$$a_{i_H} > a_{i_L}, \qquad \pi(a_{i_L}, d_{c_L}) \geq p_H, \qquad \pi(a_{i_H}, d_{c_H}) \leq p_L.$$

Lemma 1 (Grade-separated reversals violate ability sufficiency)**.** *Let $g : \mathbb{R} \to G$ be a universally calibrated threshold grading rule. If the realised curriculum contains a grade-separated reversal for g, then ability sufficiency fails.*

Proof. Let $p_H > p_L$, i_H, i_L, and c_H, c_L satisfy Definition 3. Since g is a threshold grading rule, it is weakly increasing on the attainable performance range $\mathrm{rng}(\pi)$. Therefore,

$$g(\pi(a_{i_L}, d_{c_L})) \geq g(p_H) > g(p_L) \geq g(\pi(a_{i_H}, d_{c_H})).$$

Thus the lower-ability student i_L receives a strictly higher grade than the higher-ability student i_H, violating ability sufficiency. □

Lemma 1 isolates the logical core of the impossibility. Whenever the grading system can be "confused" by an easy-course/hard-course reversal that spans distinct grade bins, one must give up either universal calibration or ability sufficiency. The remaining task is, therefore, to show that primitive conditions on the performance technology and the grading range make such confusion unavoidable.

Theorem 1 (Impossibility of Scalar Grading under Weak Overlap)**.** *Let $g : \mathbb{R} \to G$ be a universally calibrated non-trivial threshold grading rule, and suppose the performance technology satisfies Assumption 1. Then there exist attainable performance levels $p_H > p_L$ with $g(p_H) > g(p_L)$, ability levels $a_H > a_L$, and course difficulties d_H, d_L such that*

$$\pi(a_L, d_L) \geq p_H, \qquad \pi(a_H, d_H) \leq p_L.$$

Hence any realised curriculum that contains students of abilities a_H, a_L and courses of difficulties d_H, d_L contains a grade-separated reversal for g. Consequently, no universally calibrated non-trivial threshold rule can guarantee ability-sufficient cross-course comparisons on every realised curriculum whose student and course supports contain the witness quadruple.

Proof. By non-triviality, choose attainable performance levels $p_H > p_L$ such that $g(p_H) > g(p_L)$. By Assumption 1, there exist ability levels $a_H > a_L$ and course difficulties d_H, d_L satisfying

$$\pi(a_L, d_L) \geq p_H, \qquad \pi(a_H, d_H) \leq p_L.$$

Any realised curriculum containing students with abilities a_H, a_L and courses with difficulties d_H, d_L therefore contains a grade-separated reversal for g in the sense of Definition 3. By Lemma 1, ability sufficiency fails on any such curriculum. The concluding statement follows immediately. □

Theorem 1 decomposes the impossibility into two layers. The lemma is the mechanism: a one-dimensional calibrated grade cannot undo a two-dimensional performance environment once a reversal crosses a grade boundary. The theorem is the primitive bridge: weak overlap and non-triviality ensure that some such boundary-crossing reversal lies in a region of the performance range on which the grading rule is genuinely heterogeneous. The impossibility is, therefore, not that every curriculum generates reversals, but that once the domain is rich enough to produce one in a grade-relevant region, no single scalar calibration can rule it out uniformly. Equivalently, weak overlap plus non-triviality make universal calibration and ability sufficiency mutually incompatible as uniform design criteria.

The argument is robust in the sense that it does not depend on any special algebra of π or on distributional assumptions. Three ingredients are jointly load-bearing: universal calibration forces the same performance cutoffs to apply across courses; non-triviality together with the finite grade scale ensures that at least one such cutoff genuinely separates attainable performance levels; and weak overlap in the ability–difficulty technology ensures that easy-course/hard-course reversals can straddle that separating region. The theorem binds only when those ingredients operate together. The underlying logic is elementary: a scalar function of a two-dimensional argument cannot be inverted, so any fixed mapping from performance to grades will confound ability with difficulty. Formalising the impossibility in this two-step way clarifies exactly where additional structure is needed.[5]

Theorem 1 can be interpreted as a failure of single crossing. In mechanism design, single-crossing conditions ensure that a one-dimensional instrument (a price, an allocation rule) can screen agents along a one-dimensional type (Myerson, 1981). Here, the "instrument" is the grade and the "type" is two-dimensional (ability and difficulty). Each fixed performance threshold t_k corresponds in (a, d)-space to the iso-performance boundary $\{(a, d) : \pi(a, d) = t_k\}$, which is upward-sloping because π increases in ability and decreases in difficulty. Moving along that boundary trades off ability against difficulty, so a single grade bin necessarily spans multiple ability levels once difficulty varies. Lemma 1 identifies the realised manifestation of this failure: once even one such threshold is crossed in opposite directions by an easy-course/hard-course reversal, the scalar grade misorders ability.

Remark 1 (Primitive versus realised overlap). Lemma 1 is the realised-curriculum statement: one boundary-crossing reversal is enough. Assumption 1 is the primitive condition on the ability–difficulty technology that makes such witnesses available once the grading rule is

[5]Related "impossibility under calibration" phenomena appear in algorithmic fairness: when base rates differ across groups, one generally cannot satisfy calibration and certain balance/error-rate constraints simultaneously (Kleinberg et al., 2017; Chouldechova, 2017). Our setting differs substantively, but the shared lesson is that forcing a single calibrated scalar signal across heterogeneous environments can destroy identification.

non-trivial.

Operationally, fix a finite realised student set S and course set C, with observed ability support $A^{\text{obs}} := \{a_i : i \in S\}$ and realised difficulty set $D^{\text{obs}} := \{d_c : c \in C\}$. For a given grading rule g, call the realised curriculum *g-compatible* if it contains a witness quadruple $(a_L, a_H, d_L, d_H) \in A^{\text{obs}} \times A^{\text{obs}} \times D^{\text{obs}} \times D^{\text{obs}}$ with $a_H > a_L$, together with attainable levels $p_H > p_L$ satisfying $g(p_H) > g(p_L)$, $\pi(a_L, d_L) \geq p_H$, and $\pi(a_H, d_H) \leq p_L$. The theorem says that no universally calibrated non-trivial threshold rule can guarantee ability sufficiency uniformly over the class of g-compatible curricula.

If a realised curriculum has very narrow difficulty dispersion relative to the relevant grade-bin widths on the realised ability support, the theorem may simply fail to bind. The point is not that every institution lies in the impossible region, but that once the administrative range of courses and students is wide enough to realise one witness, no single calibrated scalar grade can rule out cross-course reversals uniformly.

Remark 2. For a simple illustration, consider the additive model $\pi(a, d) = a - d$. In that model, a sufficient condition for the underlying technology to *admit* a grade-separated reversal for a given threshold rule, and hence for any realised curriculum whose observed ability support contains the required ability pair to contain one, is that feasible course difficulties span more than one grade bin in performance units. Concretely, if g has adjacent cutoffs $t_{k-1} < t_k$, and the curriculum contains $d_L < d_H$ with $d_H - d_L > t_k - t_{k-1}$, then the interval $[\, t_k + d_L,\ t_{k-1} + d_H \,)$ is non-empty. One can therefore choose a_L in that interval and then choose a_H with $a_L < a_H < t_{k-1} + d_H$, which yields $a_L - d_L \geq t_k$ while $a_H - d_H < t_{k-1}$. Thus the additive model makes the reversal feasible whenever difficulty dispersion exceeds a grade-bin width; a realised curriculum contains such a reversal once its observed ability support includes such an $a_L < a_H$. The contradiction is therefore driven by the interaction of difficulty dispersion and grade coarseness, not by exotic functional forms.

Remark 3 (Course-Specific vs. Baseline Ability). For the impossibility results here, we model student ability as a scalar fixed effect, a_i, that is invariant across courses. In reality, a student's ability is likely to have both a portable baseline component and a course-specific match component. However, treating ability as a global scalar makes our impossibility theorem strictly stronger: it demonstrates that even if students possessed a single, perfectly portable level of 'ability,' a universally calibrated grading scale still could not identify it due to variations in course difficulty. If ability were explicitly course-specific, the identification problem would be strictly harder, as grades would confound baseline ability, course difficulty, and course-specific ability match. We explicitly incorporate such match effects in the additive transcript model (Section 4.2) and in the Multi-Dimensional Additive Interaction model (Section 6.2). The additive model continues to recover a scalar baseline ability measure, while

the multi-dimensional extension recovers a latent skill subspace and comparative-advantage structure rather than a canonical universal ranking unless an additional benchmark is imposed.

Relation to Grade Caps

A policy of capping the number of A grades constrains the *distribution* of grades within a course (for example, by limiting the fraction of students who may receive the top grade). When a grading cap binds, it is typically implemented by explicitly adjusting course-specific performance thresholds to earn a given grade, by imposing a quota (which implicitly adjusts thresholds whenever the quota binds), or by using a forced curve that maps within-course percentile ranks into letter grades. In each case, the cap induces an *effective* performance-to-grade map $g_{c,t}$ that is both course- and cohort-dependent: the mapping used in course c in term t depends on the realised performance distribution of the enrolled cohort.

Either implementation of a grade cap abandons universal calibration (Axiom 2) in the literal sense that there is no longer a single fixed mapping from performance to grades across courses (or even within a course across cohorts). Importantly, a cap need not violate within-course monotonicity: it can preserve monotone grading within a course by assigning the top fraction the top grade.

What is sacrificed is the interpretation of a given letter grade as a fixed performance standard across courses and time. In effect, caps *reverse-calibrate* the grade labels: instead of calibrating grades to an absolute performance scale, the institution calibrates the performance thresholds to hit pre-specified grade shares. A cap does not create the underlying difficulty-driven confounding proved in Theorem 1; that problem is already present under universal calibration. What the cap adds is a second pathology: once it binds, the same letter grade corresponds to different effective performance thresholds across courses and cohorts. Reversals can therefore arise not only because course difficulty differs, but because the meaning of a given grade label itself has become cohort- and context-dependent.

Put another way, the practical implication of Theorem 1 is that a grading system that aspires to enable meaningful cross-course comparisons requires some form of course-specific information in order to benchmark relative course difficulties' impact on performance (and hence, on grading outcomes). Yet without cross-course comparability, aggregate grade statistics like grade-point averages do not meaningfully identify student performance. An institution (or an outside observer) must therefore either (i) accept the identification failure and interpret grades as only partially comparable across courses, or (ii) systematically incorporate information about course difficulty (or related contextual variables) into the reported signal or the inference problem.

4 Resolving the Impossibility

Theorem 1 identifies a fundamental limitation of purely scalar grading. In this section and the next, we explore three avenues for recovering useful ranking information from grades despite this limitation. Each avenue operates by relaxing a different aspect of the impossibility: weakening the sufficiency requirement, enriching the information available to the observer, or addressing the strategic incentives that drive grade inflation.

4.1 Statistical Sufficiency Under Random Course Allocation

The ability sufficiency axiom (Axiom 3) requires that grade comparisons *always* correctly rank ability, for every pair of students in every pair of courses. This is a worst-case requirement. A natural weakening asks only that grades correctly rank ability *on average*.

Microfoundation. To ground this weakening economically, consider the inference problem of an outside observer. Let student ability be $a \in \mathbb{R}$ and course difficulty be $d \in \mathcal{D}$. A student taking a course with difficulty d generates a (latent) performance

$$p \;=\; \pi(a, d, \varepsilon), \qquad \text{e.g., } \pi(a, d, \varepsilon) = a - d + \varepsilon, \tag{1}$$

where ε is idiosyncratic noise. The student receives a discrete grade $g = \tau(p) \in \{g_1, g_2, \ldots, g_K\}$ via thresholds: $g = g_k$ if and only if $t_{k-1} \leq p < t_k$. (In this subsection the observer is assumed to see only the reported grade, so the object of interest is $w(g) = \mathbb{E}[a \mid g]$. Universal calibration makes a common threshold map τ natural, but it does *not* make course identity uninformative: if the observer also sees the course, the natural Bayesian ranking problem is posed in terms of $w(g, c) = \mathbb{E}[a \mid g, c]$, even when τ is common across courses.)

An outside observer—perhaps an employer considering whether to hire a student post-graduation, or a university governing assessing what level of honors to award—values ability and, observing only a grade g, sets priority (or wages/award level) using the Bayes posterior mean:

$$w(g) = \mathbb{E}[a \mid g]. \tag{2}$$

Thus, ranking applicants by grades alone is optimal if and only if $w(g)$ is weakly increasing in g; strict increase is needed only if one wants a unique strict ranking with no ties across grade levels. This maps directly to our statistical weakening, defining what we might reasonably hope for: that a higher grade is a noisy but unbiased signal of higher ability.

Axiom 5 (Weak Statistical Ability Sufficiency)**.** For any grades $g_H > g_L$ in G,

$$\mathbb{E}[a \mid g = g_H] \geq \mathbb{E}[a \mid g = g_L].$$

If the induced discrete grade experiment is strictly MLRP, the inequality is strict.

Whether this condition holds depends crucially on the precise assignment mechanism by which students sort into courses.

The Benchmark of Random Allocation

To see when this statistical ordering works, we can write the grade likelihood for a student of ability a as an integral over the course difficulties they might face:

$$\Pr(g = g_k \mid a) = \int_{\mathcal{D}} \Pr(g = g_k \mid a, d)\, dF(d \mid a), \tag{3}$$

$$\text{where} \quad \Pr(g = g_k \mid a, d) = \int_{t_{k-1}}^{t_k} f(p \mid a, d)\, dp. \tag{4}$$

Equation (3) is just the mixture representation for grade probabilities. Exogenous assignment, $F(d \mid a) = F(d)$, removes ability-dependent mixing weights from that expression. Under that restriction, the additional MLRP condition imposed in Proposition 1 yields the desired monotone Bayesian ordering of posterior mean ability by grade. Under strict independence, the course difficulty a student faces is unrelated to their underlying ability.

Proposition 1 (Weak statistical sufficiency under exogenous course assignment)**.** *Fix a grading rule satisfying Axioms 1, 2 and 4. Suppose ability a and course difficulty d are independently distributed and $\mathbb{E}|a| < \infty$. Let $f(p \mid a)$ denote the* unconditional *density of $p = \pi(a, d, \varepsilon)$ given a, i.e.*

$$f(p \mid a) = \int_{\mathcal{D}} f(p \mid a, d)\, dF(d),$$

where $f(p \mid a, d)$ is as in (3) *(integrating out ε). Assume that the family $\{f(\cdot \mid a)\}_{a \in \mathrm{supp}(F_a)}$ has a common support $\mathcal{P} \subseteq \mathbb{R}$ and satisfies the monotone likelihood ratio property (MLRP) in (p, a) on $\mathcal{P}$. Then the induced discrete grade experiment $\{\Pr(g = \cdot \mid a)\}_a$ has MLRP on the* realised grade support*: for every $a_H > a_L$, the ratios*

$$q_k(a_H, a_L) := \frac{\Pr(g = g_k \mid a_H)}{\Pr(g = g_k \mid a_L)}$$

are weakly increasing in k whenever the denominator is positive. Consequently, for any

realised grades $g_H > g_L$ with positive ex ante probability,

$$\Pr(g = g_H) > 0,\ \Pr(g = g_L) > 0 \qquad \Longrightarrow \qquad \mathbb{E}[a \mid g = g_H] \geq \mathbb{E}[a \mid g = g_L].$$

If, in addition, the induced discrete experiment is strictly MLRP on its realised support, then the posterior mean is strictly increasing in the realised grade.

Proof. Under independence, the grade probabilities can be written directly in terms of the unconditional performance density:

$$\Pr(g = g_k \mid a) = \int_{[t_{k-1}, t_k) \cap \mathcal{P}} f(p \mid a)\, dp.$$

Fix $a_H > a_L$ and consider a grade bin k with $\Pr(g = g_k \mid a_L) > 0$. Because the family has common support, $f(\cdot \mid a_H)$ is absolutely continuous with respect to $f(\cdot \mid a_L)$ on $\mathcal{P}$, so the likelihood ratio

$$r(p) := \frac{f(p \mid a_H)}{f(p \mid a_L)}$$

is well defined $f(\cdot \mid a_L)$-almost surely on $\mathcal{P}$. By MLRP, $r(p)$ is weakly increasing in p. Therefore

$$q_k(a_H, a_L) = \frac{\Pr(g = g_k \mid a_H)}{\Pr(g = g_k \mid a_L)} = \frac{\int_{[t_{k-1}, t_k) \cap \mathcal{P}} r(p)\, f(p \mid a_L)\, dp}{\int_{[t_{k-1}, t_k) \cap \mathcal{P}} f(p \mid a_L)\, dp} = \mathbb{E}\left[r(P) \mid a_L,\ t_{k-1} \leq P < t_k\right].$$

Because r is weakly increasing and the bins $[t_{k-1}, t_k)$ are ordered intervals, these conditional expectations are weakly increasing in k over bins with positive denominator. Hence the induced discrete experiment has MLRP on its realised grade support.

Via Bayesian monotone comparative statics (Milgrom, 1981), MLRP implies that the posterior distribution of a given a realised grade is weakly increasing in the monotone-likelihood-ratio order and therefore in the first-order stochastic dominance order. Since $\mathbb{E}|a| < \infty$, posterior means are well-defined for grades with positive ex ante probability, and first-order stochastic dominance implies

$$g_H > g_L,\ \Pr(g = g_H) > 0,\ \Pr(g = g_L) > 0 \implies \mathbb{E}[a \mid g = g_H] \geq \mathbb{E}[a \mid g = g_L].$$

This proves Axiom 5 in the weak sense on the realised grade support. If one strengthens the assumptions so that the induced discrete experiment is strictly MLRP there, the same argument yields strict inequality of posterior means. □

Proposition 1 establishes monotone Bayesian ranking in the weak sense. Strict ordering of posterior mean ability by grade requires a stronger discrete strict-MLRP condition. The

distinction matters because ties in posterior means can arise after coarsening a continuous performance signal into discrete grades.

Remark 4 (One primitive route to the marginal MLRP condition). Proposition 1 is stated in terms of the difficulty-marginalised density $f(p \mid a)$ because the coarsening step acts on realised performance after difficulty has been integrated out. Independence matters upstream by keeping the mixing measure over course difficulties fixed across ability types. A convenient sufficient condition for the resulting marginal MLRP hypothesis is a log-concave *location family* for the unconditional performance distribution. If $p = a + u$ where u is independent of a and has a log-concave density, then $f(p \mid a) = f_u(p - a)$ has common support and satisfies MLRP in (p, a).

In particular, in the additive model $\pi(a, d, \varepsilon) = a - d + \varepsilon$ with $d \perp a$ and $\varepsilon \perp (a, d)$, if d and ε are independent and log-concave then $u = \varepsilon - d$ is log-concave (log-concavity is preserved under convolution and reflection), so the marginal family inherits MLRP from the primitive additive structure. The proposition itself does not require this specific route; the remark simply makes transparent one non-knife-edge benchmark in which the marginal condition arises from familiar primitive assumptions.

Under Proposition 1, ranking by the posterior mean $w(g) = \mathbb{E}[a \mid g]$ is monotone in the observed grade when one conditions only on g. This is weaker than saying that the grade is a sufficient statistic once course identity is observed: even under exogenous assignment, $w(g, c) = \mathbb{E}[a \mid g, c]$ can refine $w(g)$. The proposition establishes monotone Bayesian ranking from grades alone, not irrelevance of course labels.

The Challenge Introduced by Endogenous Course Choice

However, in practice, universities typically do not randomly assign students to courses—students choose their own schedules.[6] Once course choice is endogenous, $F(d \mid a)$ depends on a, and the grade likelihood becomes the *ability-dependent mixture* in (3).

Endogenous course choice creates a classic selection problem: the distribution of observed grades is conditional on an unobserved (and ability-correlated) enrollment decision. In this sense, comparing grades across courses without conditioning on course choice resembles sample-selection bias in econometric settings where outcomes are observed only under endogenous participation (Heckman, 1979). More broadly, our results underscore an identification

[6]There are sometimes exceptions in which course assignment is indeed independent of ability, such as first-year intro curricula that must be taken by every student in the school. In those cases, Proposition 1 implies that even a single course grade can be statistically informative about expected ability in the posterior-mean sense. A larger number of common intro courses strengthens the signal and helps transcript-based aggregation, but that is a precision issue rather than a condition for the proposition itself.

point: without exogenous variation in difficulty or assignment, separating "ability" from "environment" is intrinsically underdetermined, a theme that appears in other identification problems in the social sciences (Manski, 1993).

Even if $f(p \mid a, d)$ is MLRP for every fixed d, the mixture over d with *weights that move with* a need not preserve MLRP in (g, a). Intuitively: higher a may come bundled with systematically higher d (positive sorting), so a high grade can easily be "explained" by an easy course taken by a lower-ability student rather than by high-ability student taking a hard course.

Proposition 2 (Failure of statistical sufficiency under endogenous course selection)**.** *Assume Assumption 1, and let $g : \mathbb{R} \to G$ be a universally calibrated non-trivial threshold grading rule. Then there exists a joint distribution $F(a, d)$ exhibiting positive dependence between ability and difficulty (i.e. higher-ability students take harder courses) such that statistical ability sufficiency (Axiom 5) is violated. This failure arises already in the noiseless benchmark $\varepsilon \equiv 0$.*

Proof. Work in the noiseless benchmark $\varepsilon \equiv 0$, so performance is $p = \pi(a, d)$.

By non-triviality (Axiom 4), there exist *attainable* performance levels $p_H > p_L$ in the range of π such that $g(p_H) > g(p_L)$. Define $g_H := g(p_H)$ and $g_L := g(p_L)$. By Weak Overlap (Assumption 1), for these $p_H > p_L$ there exist abilities $a_H > a_L$ and difficulties d_H, d_L such that

$$\pi(a_L, d_L) \geq p_H \qquad \text{and} \qquad \pi(a_H, d_H) \leq p_L.$$

Because g is a threshold rule (equivalently, weakly increasing in performance on attainable outcomes), these inequalities imply

$$g(\pi(a_L, d_L)) \geq g(p_H) = g_H \qquad \text{and} \qquad g(\pi(a_H, d_H)) \leq g(p_L) = g_L.$$

Since $g_H > g_L$, it follows that

$$g(\pi(a_L, d_L)) > g(\pi(a_H, d_H)).$$

Now construct an endogenous assignment environment with two equally common ability types:

$$a \in \{a_L, a_H\}, \qquad \Pr(a = a_L) = \Pr(a = a_H) = \tfrac{1}{2},$$

and two course difficulties $\{d_L, d_H\}$, with *positive sorting* given by

$$d = \begin{cases} d_L & a = a_L \\ d_H & a = a_H. \end{cases}$$

(Equivalently: lower-ability students select the easier course, and higher-ability students select the harder course.)

In this environment, the observed grade takes only two values, and we have deterministically

$$g = g(\pi(a_L, d_L)) \text{ when } a = a_L, \qquad g = g(\pi(a_H, d_H)) \text{ when } a = a_H,$$

with $g(\pi(a_L, d_L)) > g(\pi(a_H, d_H))$. Hence higher observed grades correspond to lower ability:

$$\mathbb{E}[a \mid g = g(\pi(a_L, d_L))] = a_L \ < \ a_H = \mathbb{E}[a \mid g = g(\pi(a_H, d_H))].$$

Therefore Axiom 5 is violated. □

Proposition 2 is not a contradiction to Proposition 1; it isolates where the exogenous-assignment result gets its force. Coarsening a continuous performance signal preserves monotone Bayesian ranking once the *difficulty-marginalised* density $f(p \mid a)$ satisfies MLRP. Independence matters upstream because it fixes the mixing measure over difficulties, making that marginal structure plausible; endogenous selection replaces it with ability-dependent mixing that can destroy the marginal MLRP property. The proposition is therefore an existence result about what can go wrong under sufficiently strong selection, not a claim that any arbitrarily small dependence must overturn monotone posteriors.

Synthesis. We can now map the conceptual landscape of the im/possibility results in our framework:

- **Theorem 1**: Worst-case impossibility of universal cross-course comparability from a scalar grade.
- **Proposition 1**: Best-case possibility of monotone Bayesian inference from grades when course difficulty is independent of student ability and the MLRP holds.
- **Proposition 2**: Selection (endogenous $F(d \mid a)$) can make grades anti-informative about ability, *even on average*.

Distortions from Student and Professor Incentives

Empirically, course selection is strongly endogenous: students respond to learning objectives, signalling motives, and grading standards, generating non-random matching between student characteristics and course environments (Sabot and Wakeman-Linn, 1991). Depending on the setting, this can manifest as "flight from strict graders," positive sorting into more demanding tracks, or both. For our purposes, the key implication is that the joint match distribution

$F(a, d)$ need not factorise. Proposition 2 exhibits one natural form of dependence under which statistical sufficiency fails sharply; other forms of selection can also break the grade–ability ordering.

The assumption of an exogenous, fixed course difficulty d is also routinely violated on the *supply side* of the classroom. Instructors respond to institutional incentives and student preferences, and those incentives can induce grade inflation—a dynamic we model formally in Section 5. For now, the key observation is that instructors adjusting their courses as a function of student ability, like endogenous student sorting, can break the independence property that is needed for statistical sufficiency.

The Information Destruction of Grade Caps

As we discussed above, to combat grade inflation, Harvard has recently proposed (and other institutions have in the past tried) capping the proportion of top grades awarded in every course. While such a cap mechanically halts the aggregate rise in GPAs, its effect on the grade–ability inference problem is disastrous.

Mathematically, a fixed distribution cap explicitly abandons universal calibration (Axiom 2). The absolute performance threshold for an "A" in course c in term t, $t_{A,c,t}$, is no longer a fixed standard; it becomes endogenous to the course's realised performance distribution, which depends on both the enrolled ability distribution $F_t(a \mid c)$ and the course difficulty d_c. Equivalently, the cap induces a cohort-dependent effective rule $g_{c,t}$ (or threshold map $\tau_{c,t}$) mapping performance into the common grade labels G.

Because course choice is endogenous (Proposition 2), cohorts can differ systematically in ability across courses. Under a cap, the implied "A" threshold $t_{A,c,t}$ is therefore cohort-specific. Suppressing the term index only when harmless, it need not be monotone in structural course difficulty or aligned with any administrative ordering of courses: higher enrolled ability tends to push $t_{A,c,t}$ up, while higher difficulty d_c tends to push it down. The key point is that $t_{A,c,t}$ is no longer a fixed, universal performance standard, but an endogenous object that varies across courses and cohorts. Under a cap, those thresholds can therefore differ sharply across courses and cohorts. An employer observing a transcript and computing $w(g = \text{A})$ is now forced to average together the top 20% of an elite cohort with the top 20% of a weaker cohort.

Worse, the cap hard-codes the reversal into institutional policy. A student of ability $a = 90$ in a course taken by exceptionally strong students might place in the 75th percentile and receive a "B," while a student of ability $a = 60$ in a weaker classroom cohort could place in the 85th percentile and receive an "A." This produces a particularly perverse strategic incentive for students: to maximise their GPAs, students must not only avoid hard syllabi, but actively hunt for courses populated by *lower-ability peers*. By transforming the grade

from a measure of absolute performance into a measure of strictly local, within-course rank, caps destroy the signal of absolute ability that employers, graduate programs, and other parties would hope to use grades to infer.

What a Rational Employer Does When Selection Exists

If grades alone fail to be informative under endogenous course selection—and are even less informative under institutional grade caps—how should an employer form inferences? If employers observe course identities c but not latent difficulty or cohort strength, they *must* condition on course identity when making their assessments:

$$w(g, c) = \mathbb{E}[a \mid g, c], \tag{5}$$

which can be more informative than just considering $w(g)$. Grades-only rankings are optimal only in the special case where conditioning on c does not change posterior estimates of ability (e.g., effectively exogenous assignment, negligible difficulty dispersion, or no reliable information to learn d_c).

With enough market history (past transcripts and subsequent job performance), a micro-founded approach is to treat difficulty as a latent parameter d_c and estimate it explicitly. For example, fitting a two-way fixed effect or item-response theory (IRT) style model

$$p_{ic} = a_i - d_c + \varepsilon_{ic}, \qquad g_{ic} = \tau_c(p_{ic}), \tag{6}$$

allows the employer to use the full transcript posterior $\mathbb{E}[a_i \mid \{(g_{ic}, c)\}_c]$ for hiring. This insight—that restoring identification requires extracting relational information from the cross-sectional structure of multi-course transcripts—motivates our formal resolution of the problem using spectral methods in the following section.

4.2 Multi-Course Transcripts and Eigengrades

The impossibility in Theorem 1 arises because the observer sees only a single grade. In practice, transcripts contain many grades: a student's performance across a portfolio of courses. This richer information structure can, under appropriate conditions, resolve the identification problem entirely.

4.2.1 The Additive Model

For analytical tractability, we henceforth adopt a version of the additive performance function used in illustrative examples earlier:

$$\pi(a_i, d_c) = a_i - d_c + \varepsilon_{ic},$$

where ε_{ic} is a mean-zero noise term representing idiosyncratic factors that are independent across student–course pairs with distribution F_ε and $\mathbb{E}[\varepsilon] = 0$ (such as luck on exam days, or idiosyncratic dimensions of topic match). The observer sees the grade matrix $\mathbf{G}$ with entries $G_{ic} = g(a_i - d_c + \varepsilon_{ic})$. For analytical purposes, we work with the latent performance matrix $\mathbf{P}$ with entries $P_{ic} = a_i - d_c + \varepsilon_{ic}$. In applications, numerical grade points (e.g., $A = 4, B = 3$) can be viewed as a coarse proxy for this latent index, but formally the observed outcomes are grades $G_{ic} = g(P_{ic})$ generated by discretisation via cutpoints.[7]

Remark 5 (Ordinal grades versus cardinal performance). Transcripts typically report either grade points (e.g., $A = 4, B = 3, \ldots$) or *ordinal* letter grades. For clarity our identification arguments are stated for a latent *cardinal* performance variable P_{ic}. A standard way to connect the two is an ordered-response model: there exist cutpoints $-\infty = \tau_0 < \tau_1 < \cdots < \tau_K = +\infty$ and a continuous strictly increasing CDF F such that

$$P_{ic} = a_i - d_c + \varepsilon_{ic}, \qquad \mathbb{P}(G_{ic} \le k \mid a_i, d_c) = F\Big(\tau_k - (a_i - d_c)\Big),$$

where $G_{ic} \in \{1, \ldots, K\}$ is the observed letter grade and ε_{ic} is a location-family noise with CDF F (e.g., probit or logit).

Under this interpretation, the "grade points" encoding used in practice (and in toy examples) should be viewed as a convenient monotone re-scaling of an ordinal signal. When only ordinal information is available and one wants a fully structural estimator, one can estimate $(\mathbf{a}, \mathbf{d})$ via ordered-logit/probit fixed effects (plus the cutpoints τ_k), which identifies the latent index $a_i - d_c$ up to the usual location (and, if F's scale is not fixed, scale) normalisations.

Proposition 3 (Identification with ordinal grades)**.** *Assume the ordered-response model in Remark 5 with known cutpoints (τ_k) and strictly increasing CDF F. Suppose the observation graph $\mathcal{G} \subseteq S \times C$ is connected. If two parameter vectors $(\mathbf{a}, \mathbf{d})$ and $(\mathbf{a}', \mathbf{d}')$ induce the same*

[7] In the additive benchmark, the spectral construction does not identify a new latent object beyond the centred student effect; its value is to express that object in a form that generalises naturally to perturbation analysis.

conditional distribution of $(G_{ic})_{(i,c)\in\mathcal{G}}$*, then there exists a constant* $\kappa \in \mathbb{R}$ *such that*

$$a_i' = a_i + \kappa \quad \forall i, \qquad d_c' = d_c + \kappa \quad \forall c.$$

Proof. Fix $(i, c) \in \mathcal{G}$. For any k, the cumulative probability

$$\mathbb{P}(G_{ic} \le k \mid a_i, d_c) = F\Big(\tau_k - (a_i - d_c)\Big)$$

is strictly monotone in the index $(a_i - d_c)$ because F is strictly increasing. Hence equality of conditional grade distributions implies $a_i - d_c = a_i' - d_c'$ for every observed edge $(i, c) \in \mathcal{G}$. Connectivity then implies $\mathbf{a}' = \mathbf{a} + \kappa\mathbf{1}_n$ and $\mathbf{d}' = \mathbf{d} + \kappa\mathbf{1}_m$ for some constant κ, exactly as in the cardinal case. □

Remark 6 (Normalisations when cutpoints or scale are unknown)*.* If the cutpoints are not fixed in advance, or if F is specified only up to a location–scale normalisation, then the ordered-response model carries the familiar additional location/scale indeterminacies. Because those standard normalisations play no separate role in the spectral arguments below, we do not use them as a standalone formal result here.

4.2.2 Identification via Double Centring

We begin with the complete-data, noiseless case to build intuition.

Proposition 4 (Identification with complete data)**.** *In the noiseless additive model* ($\varepsilon_{ic} = 0$) *with complete data (every student takes every course), abilities and difficulties are identified up to a common additive constant from the performance matrix.*

Proof. With $P_{ic} = a_i - d_c$, define the row means and column means:

$$\bar{P}_{i\cdot} = \frac{1}{m}\sum_{c=1}^{m} P_{ic} = a_i - \bar{d}, \qquad \bar{P}_{\cdot c} = \frac{1}{n}\sum_{i=1}^{n} P_{ic} = \bar{a} - d_c,$$

where $\bar{a} = \frac{1}{n}\sum_i a_i$ and $\bar{d} = \frac{1}{m}\sum_c d_c$. It follows immediately that:

$$a_i - \bar{a} = \bar{P}_{i\cdot} - (\bar{a} - \bar{d}) = \bar{P}_{i\cdot} - \bar{P}_{\cdot\cdot},$$

where $\bar{P}_{\cdot\cdot}$ is the grand mean. The centred abilities $a_i - \bar{a}$ and centred difficulties $d_c - \bar{d}$ are exactly identified. Since the ranking of $a_i - \bar{a}$ is the same as the ranking of a_i, the ability ranking is fully recovered. □

The result is simple but important: with complete data, the identification problem disappears because the row and column structure of the matrix disentangles ability from difficulty. A student who scores highly across many courses of varying difficulty is revealed as high-ability, regardless of the absolute level of grades.

4.2.3 Eigengrades

Double-centring already solves the complete-data benchmark. The spectral formulation below is useful because it isolates the same centred-ability object in a way that extends naturally to noisy or partially observed data.

Definition 4 (Centred grade matrix)**.** Let $\tilde{\mathbf{P}}$ denote the *grand-mean–centred* performance matrix with entries $\tilde{P}_{ic} = P_{ic} - \bar{P}_{\cdot\cdot}$.

Definition 5 (Student affinity matrix)**.** The student affinity matrix is $\mathbf{A} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^\top$, with entries $A_{ij} = \sum_c \tilde{P}_{ic}\tilde{P}_{jc}$.

The affinity matrix measures the similarity of students across courses: A_{ij} is large and positive when students i and j have similar (centred) performance profiles, and large and negative when their profiles are anti-correlated. The eigengrade vector is the leading eigenvector of the affinity matrix after projecting out the constant component (equivalently, the leading eigenvector in the subspace orthogonal to $\mathbf{1}_n$).

Definition 6 (Eigengrades)**.** Let $\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ denote the centring matrix that projects onto the subspace orthogonal to $\mathbf{1}_n$. The **eigengrade vector $\boldsymbol{\alpha}$** is the unit-norm eigenvector corresponding to the largest eigenvalue of the projected affinity matrix $\mathbf{MAM}$ (equivalently, the leading eigenvector of $\mathbf{A}$ restricted to $\mathbf{1}_n^\perp$).[8]

The term "eigengrades" reflects the spectral origin of the construction. The idea is analogous to PageRank (Page et al., 1999), where the importance of a webpage is determined by the principal eigenvector of the link matrix, or to the Pinski–Narin method for ranking journals (Pinski and Narin, 1976). In our setting, a student is interpreted as high-ability if that student performs well in courses where other high-ability students also perform well—a recursive, self-referential definition that the eigenvector resolves into a fixed point.

Theorem 2 (Eigengrade identification)**.** *In the noiseless additive model with complete data, assume $\tilde{\mathbf{a}} \neq 0$. Then the eigengrade vector $\boldsymbol{\alpha}$ is proportional to the centred ability vector $\tilde{\mathbf{a}} = (a_1 - \bar{a}, \ldots, a_n - \bar{a})^\top$. Eigengrades thus recover the true ability ranking.*

[8]Note that any eigenvector x of $\mathbf{MAM}$ with nonzero eigenvalue satisfies $x \in \mathbf{1}_n^\perp$ automatically (since $\mathbf{M}\mathbf{1}_n = \mathbf{0}$ implies $\mathbf{MAM}\mathbf{1}_n = \mathbf{0}$), so when the leading eigenvalue of $\mathbf{MAM}$ is positive, this is equivalent to simply taking its leading eigenvector (up to sign).

Proof. In the noiseless additive model, $\tilde{P}_{ic} = \tilde{a}_i - \tilde{d}_c$ where $\tilde{a}_i = a_i - \bar{a}$ and $\tilde{d}_c = d_c - \bar{d}$. In matrix notation,

$$\tilde{\mathbf{P}} = \tilde{\mathbf{a}}\mathbf{1}_m^\top - \mathbf{1}_n\tilde{\mathbf{d}}^\top,$$

where $\mathbf{1}_n$ and $\mathbf{1}_m$ are the n-dimensional and m-dimensional vectors of ones. The student affinity matrix satisfies

$$\begin{aligned}\mathbf{A} &= \tilde{\mathbf{P}}\tilde{\mathbf{P}}^\top \\ &= \left(\tilde{\mathbf{a}}\mathbf{1}_m^\top - \mathbf{1}_n\tilde{\mathbf{d}}^\top\right)\left(\mathbf{1}_m\tilde{\mathbf{a}}^\top - \tilde{\mathbf{d}}\mathbf{1}_n^\top\right) \\ &= m\,\tilde{\mathbf{a}}\tilde{\mathbf{a}}^\top + \|\tilde{\mathbf{d}}\|^2\,\mathbf{1}_n\mathbf{1}_n^\top,\end{aligned}$$

where the cross terms vanish because $\mathbf{1}_m^\top\tilde{\mathbf{d}} = \sum_c \tilde{d}_c = 0$.

By Definition 6, eigengrades are computed from the projected affinity matrix $\mathbf{MAM}$, where $\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ satisfies $\mathbf{M}\mathbf{1}_n = \mathbf{0}$ and $\mathbf{M}\tilde{\mathbf{a}} = \tilde{\mathbf{a}}$ (since $\tilde{\mathbf{a}}$ is centred). Therefore,

$$\mathbf{MAM} = m\,\tilde{\mathbf{a}}\tilde{\mathbf{a}}^\top,$$

which is rank one with principal eigenvector proportional to $\tilde{\mathbf{a}}$. Hence $\boldsymbol{\alpha} \propto \tilde{\mathbf{a}}$, and eigengrades recover the true ability ranking. □

Remark 7 (Complete-data eigengrades are a spectral re-expression of row means). In the benchmark of Proposition 4, centred row means already recover $a_i - \bar{a}$. Theorem 2 therefore does not add new identification in the complete-data noiseless case; its value is that the same centred-ability object appears as the leading spectral direction, which is the formulation that extends naturally to perturbation analysis under noise and missingness.

Remark 8 (The unavoidable "all-ones" component). Even after grand-mean centring, heterogeneity in course difficulty induces a rank-one component in the affinity matrix $\mathbf{A}$ proportional to $\mathbf{1}_n\mathbf{1}_n^\top$. This component reflects cross-course level effects rather than differences in student ability. The projection $\mathbf{MAM}$ removes it mechanically. In the complete-data noiseless additive model, the projected matrix is exactly rank one, so ability is recovered without requiring any comparison between $\mathrm{Var}(a)$ and $\mathrm{Var}(d)$. In finite samples, the same projection plays the familiar role of removing the "mean" principal component before applying PCA.

Proposition 5 (Identification with incomplete data)**.** *Consider the noiseless additive model $P_{ic} = a_i - d_c$, but suppose P_{ic} is observed only for student–course pairs (i, c) belonging to an observed bipartite graph $\mathcal{G} \subseteq S \times C$. Then $(a_1, \ldots, a_n)$ and $(d_1, \ldots, d_m)$ are identified up to a common additive constant if and only if $\mathcal{G}$ is connected.*

Proof. Suppose two parameter vectors $(\mathbf{a}, \mathbf{d})$ and $(\mathbf{a}', \mathbf{d}')$ generate the same observed outcomes on $\mathcal{G}$:

$$a_i - d_c = a'_i - d'_c \quad \text{for all } (i,c) \in \mathcal{G}.$$

Rearranging yields

$$a_i - a'_i = d_c - d'_c \quad \text{for all } (i,c) \in \mathcal{G}.$$

Fix any student i_0. For any course c adjacent to i_0, we have $d_c - d'_c = a_{i_0} - a'_{i_0}$. For any student i adjacent to such a course c, we then have $a_i - a'_i = d_c - d'_c = a_{i_0} - a'_{i_0}$. By connectivity of $\mathcal{G}$, this argument propagates along paths to all vertices, implying that there exists a constant κ such that $a_i - a'_i = \kappa$ for all i and $d_c - d'_c = \kappa$ for all c. Hence, the parameters are identified up to a common additive constant.

Conversely, if $\mathcal{G}$ is disconnected, then one can shift $(\mathbf{a}, \mathbf{d})$ by different constants on different connected components without changing any observed difference $a_i - d_c$, so identification fails. □

Remark 9 (Estimation with noise and large, sparse transcripts). With i.i.d. mean-zero noise ε_{ic} and incomplete transcripts, a natural approach is a two-way fixed-effects (least-squares) estimator,

$$(\hat{\mathbf{a}}, \hat{\mathbf{d}}) \in \arg\min_{\mathbf{a},\mathbf{d}} \sum_{(i,c)\in\mathcal{G}} \Big(P_{ic} - a_i + d_c\Big)^2,$$

under a normalisation such as $\sum_i a_i = 0$ (or $\sum_c d_c = 0$). Under standard regularity conditions and increasing graph density, this estimator is consistent for $(\mathbf{a}, \mathbf{d})$ up to the normalisation; see, e.g., Abowd et al. (1999) for the canonical connected-graph identification argument in a related bipartite setting.

Spectral methods such as eigengrades provide a computationally simple approximation. In our notation, left-multiplication by M student-centres the matrix (it subtracts, from each course column, its student mean), equivalently restricting attention to the subspace orthogonal to $\mathbf{1}_n$, so the estimator can be written directly as $\widehat{A} = M\,\widehat{P}\,\widehat{P}^\top M$. Consistency then follows from operator-norm control of $\widehat{A} - A$ and standard eigenvector perturbation bounds, once appropriate sampling/degree conditions (MCAR versus MNAR, degree growth, etc.) are imposed.

The inverse-probability weighting logic is standard in missing-data and survey-sampling settings (Horvitz and Thompson, 1952; Rubin, 1976; Little and Rubin, 2002).

Following the standard Davis–Kahan/Wedin $\sin\Theta$ perturbation theory, for nonzero vectors u, v we define

$$\sin\Theta(u,v) := \sqrt{1 - \frac{(u^\top v)^2}{\|u\|^2\|v\|^2}};$$

when u and v have unit norm, this reduces to $\sin\Theta(u,v) = \sqrt{1-(u^\top v)^2} = \sqrt{1-(|u^\top v|)^2}$, which is the sine of the principal angle between the one-dimensional subspaces $\mathrm{span}(u)$ and $\mathrm{span}(v)$.[9] This scalar notion is the one-dimensional specialisation of the usual subspace $\sin\Theta(\cdot,\cdot)$ matrix (as in Section B).

Our next result is an oracle theorem for the latent-score matrix: it isolates the spectral step after the analyst has already obtained an unbiased proxy for the latent index on the observed entries and knows the observation propensities $\{p_{ic}\}$. This should not be read as a direct consistency theorem for raw ordinal transcript data G_{ic}; that additional step requires a first-stage ordered-response model and is handled separately in Proposition 6. Accordingly, the result below is best interpreted as a perturbation bound for the centred rank-one signal under latent ignorability and oracle inverse-probability weighting.

Theorem 3 (Oracle perturbation bound for dense latent-score panels with known propensities)**.** *Treat $(a_i)_{i=1}^n$, $(d_c)_{c=1}^m$, and $(p_{ic})_{i,c}$ as fixed arrays satisfying*

$$|a_i| \le A, \qquad |d_c| \le D, \qquad p_{ic} \ge p_{\min} > 0.$$

Assume the additive latent-score model

$$P_{ic} = a_i - d_c + \varepsilon_{ic},$$

where (ε_{ic}) are independent mean-zero σ-subgaussian random variables.[10]

Let $\Omega_{ic} \in \{0,1\}$ be observation indicators such that, conditional on the fixed effects (a_i, d_c), the random variables Ω_{ic} are independent with

$$\Omega_{ic} \mid (a_i, d_c) \sim \mathrm{Bernoulli}(p_{ic}),$$

and assume $\Omega_{ic} \perp \varepsilon_{ic} \mid (a_i, d_c)$. This allows missingness to depend on the latent fixed effects through p_{ic}, but rules out selection on the realised idiosyncratic shock.

Suppose the analyst has access to the oracle inverse-probability-weighted latent-score proxy

$$\widehat{P}_{ic} := \frac{\Omega_{ic}}{p_{ic}} P_{ic}.$$

Let $M := I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and define

$$\widehat{A} := M\,\widehat{P}\,\widehat{P}^\top M.$$

[9] The squaring (equivalently, the absolute value) automatically resolves the eigenvector sign ambiguity.
[10] For example, $\mathbb{E}[\exp(t\varepsilon_{ic})] \le \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$.

Let $\widehat{\alpha}$ be the top unit eigenvector of $\widehat{A}$ restricted to $\mathbf{1}_n^{\perp}$ (equivalently, the top left singular vector of $M\widehat{P}$). Write $\widetilde{\mathbf{a}} := M\mathbf{a}$ and assume $\widetilde{\mathbf{a}} \neq 0$. Then there exist universal constants $C, c > 0$ such that, conditional on the fixed arrays $(\mathbf{a}, \mathbf{d}, (p_{ic}))$,

$$\Pr\left(\sin\Theta\Big(\widehat{\alpha},\ \frac{\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{a}}\|}\Big) \le \min\left\{1,\ C\,\frac{A+D+\sigma}{p_{\min}} \cdot \frac{\sqrt{n}+\sqrt{m}}{\sqrt{m}\,\|\widetilde{\mathbf{a}}\|}\right\}\right) \ge 1 - 2\exp(-c(n+m)).$$

In particular, if $\|\widetilde{\mathbf{a}}\| = \Theta(\sqrt{n})$, $m = \Theta(n)$, and $p_{\min}$ is bounded away from zero, then

$$\sin\Theta\Big(\widehat{\alpha}, \frac{\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{a}}\|}\Big) = O_{\mathbb{P}}(n^{-1/2}).$$

If m does not grow with n, the bound need not vanish, so consistency of the eigengrade direction is not guaranteed without additional structure.

Proof. All probability statements below are conditional on the fixed arrays $(\mathbf{a}, \mathbf{d}, (p_{ic}))$. Let

$$P^0 := \mathbf{a}\mathbf{1}_m^{\top} - \mathbf{1}_n\mathbf{d}^{\top}$$

denote the noiseless matrix and write

$$\widehat{P} = P^0 + W, \qquad W_{ic} := \Big(\frac{\Omega_{ic}}{p_{ic}} - 1\Big)(a_i - d_c) + \frac{\Omega_{ic}}{p_{ic}}\,\varepsilon_{ic}.$$

Conditional on $(\mathbf{a}, \mathbf{d}, (p_{ic}))$, the entries (W_{ic}) are independent and mean-zero.

To bound the perturbation scale, write

$$X_{ic} := \Big(\frac{\Omega_{ic}}{p_{ic}} - 1\Big)(a_i - d_c), \qquad Z_{ic} := \frac{\Omega_{ic}}{p_{ic}}\varepsilon_{ic}.$$

Since $\left|\frac{\Omega_{ic}}{p_{ic}} - 1\right| \le 1/p_{\min}$ and $|a_i - d_c| \le A + D$, we have $|X_{ic}| \le (A+D)/p_{\min}$, so X_{ic} is subgaussian with scale $\lesssim (A+D)/p_{\min}$. Moreover, $Z_{ic} \mid \Omega_{ic} = 0 = 0$ and $Z_{ic} \mid \Omega_{ic} = 1 = \varepsilon_{ic}/p_{ic}$, so Z_{ic} is subgaussian with scale at most $\sigma/p_{\min}$. By standard closure properties of subgaussian norms, $W_{ic} = X_{ic} + Z_{ic}$ is subgaussian with scale on the order of $(A + D + \sigma)/p_{\min}$.

Now remove course difficulties by student-side centring:

$$MP^0 = M(\mathbf{a}\mathbf{1}_m^{\top} - \mathbf{1}_n\mathbf{d}^{\top}) = (M\mathbf{a})\mathbf{1}_m^{\top} = \widetilde{\mathbf{a}}\mathbf{1}_m^{\top}.$$

Hence

$$M\widehat{P} = \widetilde{\mathbf{a}}\mathbf{1}_m^{\top} + MW.$$

The signal matrix $S := \widetilde{\mathbf{a}}\mathbf{1}_m^\top$ is rank one, with top singular value

$$\sigma_1(S) = \|\widetilde{\mathbf{a}}\|\,\|\mathbf{1}_m\| = \|\widetilde{\mathbf{a}}\|\sqrt{m},$$

and all remaining singular values equal to 0.

By a standard operator-norm bound for rectangular subgaussian matrices with independent entries, there exist universal constants $C, c > 0$ such that with conditional probability at least $1 - 2\exp(-c(n+m))$,

$$\|MW\|_{\mathrm{op}} \le \|W\|_{\mathrm{op}} \le C\,\frac{A + D + \sigma}{p_{\min}}\,(\sqrt{n} + \sqrt{m}).$$

Applying Wedin's (equivalently, Davis–Kahan's) $\sin\Theta$ theorem to the top left singular vector of the rank-one signal S perturbed by MW gives

$$\sin\Theta\Big(\widehat{\alpha},\ \frac{\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{a}}\|}\Big) \le \min\left\{1, \frac{\|MW\|_{\mathrm{op}}}{\sigma_1(S)}\right\} \le \min\left\{1,\ C\,\frac{A + D + \sigma}{p_{\min}} \cdot \frac{\sqrt{n} + \sqrt{m}}{\sqrt{m}\,\|\widetilde{\mathbf{a}}\|}\right\},$$

which is the stated bound. □

Remark 10 (Scope of the perturbation bound: dense versus sparse transcripts). Theorem 3 is intentionally a dense-oracle perturbation result. The lower bound $p_{ic} \ge p_{\min} > 0$ implies student degrees of order m and course degrees of order n, so the theorem is not a formal guarantee for transcript regimes in which each student takes only $O(1)$ or $O(\log n)$ courses. In sparse bipartite graphs, connectivity still suffices for *identification* in the noiseless sense of Proposition 5, but *estimation* requires additional graph-dependent machinery involving minimum or average degree, spectral gap, expansion, and the dependence structure of the observation pattern. The policy lesson about overlap should therefore be read in two layers: overlap is necessary for identification, while precise spectral recovery in sparse data is a separate quantitative problem.

The next proposition formalises the econometric interface between observed ordinal transcripts and our latent-index eigengrade geometry. It shows that if the first-stage ordered-response fit consistently recovers centred student effects, then the resulting spectral score is asymptotically aligned with centred ability.

Proposition 6 (A rigorous ordinal-to-eigengrade bridge)**.** *Assume the ordered-response model in Remark 5. Suppose a first-stage estimator based on the observed ordinal transcript returns fitted latent effects* $(\hat{\mathbf{a}}, \hat{\mathbf{d}})$ *such that*

$$\|M\hat{\mathbf{a}} - M\mathbf{a}\| = o_{\mathbb{P}}(\|M\mathbf{a}\|).$$

Define the fitted latent-index matrix $\hat{P}^0_{ic} := \hat{a}_i - \hat{d}_c$*, the fitted affinity matrix* $\hat{A}^{\mathrm{ord}} := M\hat{P}^0(\hat{P}^0)^\top M$*, and let* $\hat{\alpha}^{\mathrm{ord}}$ *be its top unit eigenvector on* $\mathbf{1}_n^\perp$*. If* $M\mathbf{a} \neq 0$*, then*

$$\sin\Theta\Big(\hat{\alpha}^{\mathrm{ord}}, \frac{M\mathbf{a}}{\|M\mathbf{a}\|}\Big) \to 0.$$

Proof. Because $M\mathbf{1}_n = \mathbf{0}$,

$$M\hat{P}^0 = M(\hat{\mathbf{a}}\mathbf{1}_m^\top - \mathbf{1}_n\hat{\mathbf{d}}^\top) = (M\hat{\mathbf{a}})\mathbf{1}_m^\top.$$

Hence

$$\hat{A}^{\mathrm{ord}} = M\hat{P}^0(\hat{P}^0)^\top M = m\,(M\hat{\mathbf{a}})(M\hat{\mathbf{a}})^\top,$$

so its top unit eigenvector is $\hat{\alpha}^{\mathrm{ord}} = M\hat{\mathbf{a}}/\|M\hat{\mathbf{a}}\|$ up to sign. The assumed relative first-stage consistency implies

$$\left\| \frac{M\hat{\mathbf{a}}}{\|M\hat{\mathbf{a}}\|} - \frac{M\mathbf{a}}{\|M\mathbf{a}\|} \right\| = o_{\mathbb{P}}(1),$$

and therefore the principal angle between the two one-dimensional subspaces converges to zero. □

The hard problem is the first-stage recovery of centred student effects from sparse ordinal panel data. Proposition 6 shows only that, once such a first stage is available, the spectral step preserves the same centred-ability direction.

Thus, for observed letter grades, the rigorous route is two-step: estimate the latent ordered-response index first, then apply the eigengrade geometry to the fitted latent matrix. Theorem 3 handles the spectral step itself; Proposition 6 shows that once a first stage consistently recovers centred student effects, the spectral step preserves that direction. The proposition is therefore an algebraic bridge rather than a standalone identification theorem for sparse many-effects ordered-response panels; primitive conditions guaranteeing the first-stage rate are a separate econometric issue.

A natural sparse-panel asymptotic regime would let $n, m \to \infty$ with the observation graph remaining connected and student/course degrees growing sufficiently for both the incidental-parameter problem and the ordered thresholds to be estimable. Under such a regime, one would need a first-stage estimator that delivers $\|M\hat{\mathbf{a}} - M\mathbf{a}\| = o_{\mathbb{P}}(\|M\mathbf{a}\|)$ despite ordinal coarsening and graph sparsity. Establishing those primitive conditions for a specific many-effects ordered logit/probit procedure is beyond the present paper, but this is the precise econometric seam on which feasible sparse-transcript recovery turns.

Remark 11 (Centring bridge). Recalling that $\mathbf{A} = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^\top$ in Section 4.2, note that $M\tilde{\mathbf{P}} = M\mathbf{P}$ because $M\mathbf{1}_n = \mathbf{0}$, and therefore $M\mathbf{A}M = M\tilde{\mathbf{P}}\tilde{\mathbf{P}}^\top M = (M\mathbf{P})(M\mathbf{P})^\top$. Thus the estimator

in Theorem 3 targets the same centred-ability object as the eigengrade definition.

Remark 12 (Selection on idiosyncratic fit). If enrolment depends on realised match shocks (e.g., students drop courses when ε_{ic} is low), then $\Omega_{ic} \not\perp \varepsilon_{ic} \mid (a_i, d_c)$ and IPW based on (a_i, d_c) is insufficient. Addressing such "selection on match" requires an explicit participation model or instruments that shift enrolment without directly shifting performance, so we do not pursue it further here. Similarly, we analyse an independent Bernoulli observation model for (Ω_{ic}); allowing within-student dependence (e.g., fixed course loads) or within-course constraints (e.g., capacity limits) would require different concentration tools for dependent observation graphs.

Remark 13 (A feasible MNAR estimator: propensity-weighted eigengrades). Theorem 3 is stated in an oracle form, treating the propensities p_{ic} as known. A feasible implementation replaces p_{ic} with an estimated propensity $\hat{p}_{ic}$ computed from the observed enrolment matrix $(\Omega_{ic})_{i,c}$ (and any observed covariates x_{ic} capturing prerequisites, timetable feasibility, programme requirements, and so forth). One convenient specification is a two-way logit model:

$$\Pr(\Omega_{ic} = 1 \mid x_{ic}) = \Lambda\big(\mu + u_i + v_c + x_{ic}^\top \theta\big), \qquad \Lambda(z) = \frac{1}{1 + e^{-z}},$$

with normalisations $\sum_i u_i = 0$ and $\sum_c v_c = 0$. Let $\hat{p}_{ic}$ denote the fitted probabilities and use *stabilised* weights $w_{ic} := 1/\max\{\hat{p}_{ic}, p_{\min}\}$ for some trimming constant $p_{\min} > 0$. Let

$$Y_{ic} := \Omega_{ic} P_{ic}$$

denote the observed masked latent score, so unobserved pairs contribute 0 before reweighting.

The resulting MNAR-corrected eigengrade estimator is:

$$\widehat{P}_{ic}^{\mathrm{IPW}} := w_{ic} Y_{ic} \qquad \text{and} \qquad \widehat{\alpha}^{\mathrm{IPW}} := \text{top eigenvector of } M\widehat{P}^{\mathrm{IPW}}(\widehat{P}^{\mathrm{IPW}})^\top M \text{ on } \mathbf{1}_n^\perp.$$

For ordinal grades rather than latent scores, one would analogously replace P_{ic} by the relevant first-stage latent-index proxy. Alternatively, one can estimate $(\mathbf{a}, \mathbf{d})$ via weighted two-way fixed effects and use $M\hat{\mathbf{a}}$ as an eigengrade-type score. This targets the same centred ability object under the additive model, but it is not generally identical sample-by-sample to the principal eigenvector of $M\widehat{P}^{\mathrm{IPW}}(\widehat{P}^{\mathrm{IPW}})^\top M$. Under standard regularity conditions ensuring $\max_{i,c} |\hat{p}_{ic} - p_{ic}| \to 0$ and p_{ic} bounded away from 0, the plug-in version inherits Theorem 3's rate up to an additional term of smaller order coming from propensity estimation error.

Remark 14 (Endogenous selection revisited). Transcript-based estimators react differently to endogenous selection. If enrolment depends only on the additive fixed effects (a_i, d_c) and still satisfies $\Omega_{ic} \perp \varepsilon_{ic} \mid (a_i, d_c)$, then a two-way fixed-effects or ordered-response fit on the

observed edges remains a natural primary estimator of the latent index $a_i - d_c$ on the connected observation graph. The sharper concern is with matrix-based constructions that implicitly treat missing entries as informative zeros or rely on unweighted affinity calculations—for such estimators, the target changes with the observation design unless one reweights by propensities.

Bias for fixed-effect estimators becomes a genuine issue once selection is non-ignorable relative to the maintained additive specification. If enrolment is ignorable conditional on observables or latent fixed effects, so that $\Omega_{ic} \perp \varepsilon_{ic} \mid (a_i, d_c, x_{ic})$, then inverse-probability weighting or a correctly specified participation model can restore the relevant moment conditions. By contrast, if selection depends on realised match shocks themselves, standard IPW based only on observables or fixed effects is not enough; one then needs additional structure, such as an explicit participation equation with exclusion restrictions or valid instruments. More generally, the formal guarantees in Section 4.2 are best read as applying when overlap is institutionally mandated or when selection is ignorable conditional on the maintained latent structure; the strategic enrolment response modelled in Section 5.1 goes beyond that benchmark and would require a joint analysis of sorting and identification. Nevertheless, when there is enough overlap, transcript-based methods still extract substantially more relational information than simple GPA comparisons, which entirely fail under the selection pattern in Proposition 2.

4.2.4 An Illustrative Example: Distortions and Difficulty-Adjusted Scores

To make the mechanics of the identification failure and its transcript-based resolution concrete, consider a hypothetical university with ten students ($s_1, \ldots, s_{10}$, indexed by true descending ability) and three courses of varying intrinsic difficulty: c_{Hard}, c_{Med}, and c_{Easy}.

Each student takes exactly two courses. The five strongest students ($s_1, \ldots, s_5$) take c_{Hard} and c_{Med}; the five weakest ($s_6, \ldots, s_{10}$) take c_{Med} and c_{Easy}. All ten students therefore appear in c_{Med}, which serves as the bridge linking the two halves of the enrolment graph. The bipartite observation graph $\mathcal{G}$ is connected, even though no student takes both c_{Hard} and c_{Easy} directly. Thus, in the additive latent-performance model of Proposition 5, abilities and difficulties are identified up to a common additive constant.

To combat grade inflation, the university imposes a strict grading cap: in a class of N students, at most $\lceil 0.2N \rceil$ may receive an A (4.0) and at most $\lceil 0.4N \rceil$ a B (3.0); the remainder receive C's (2.0).[11] Within each course, grades are assigned strictly by true ability rank; there are no idiosyncratic shocks in this example, so the *only* source of distortion is the grading

[11]This captures a quota-style rule in which grades are largely determined by within-course rank rather than an absolute standard.

cap itself.

The caps allocate grades to courses as follows: c_{Hard} ($N = 5$) awards 1 A, 2 Bs, and 2 Cs; c_{Med} ($N = 10$) awards 2 As, 4 Bs, and 4 Cs; c_{Easy} ($N = 5$) awards 1 A, 2 Bs, and 2 Cs. Table 1 reports the resulting grades, raw GPAs, and a difficulty-adjusted score constructed from transcript overlap.

Student	True Rank	Course Grades			Raw GPA	Difficulty-Adjusted Score
		c_{Hard}	c_{Med}	c_{Easy}		
s_1	1	4.0 (A)	4.0 (A)	—	4.00	1.50
s_2	2	3.0 (B)	4.0 (A)	—	**3.50**	1.00
s_3	3	3.0 (B)	3.0 (B)	—	3.00	0.50
s_4	4	2.0 (C)	3.0 (B)	—	2.50	0.00
s_5	5	2.0 (C)	3.0 (B)	—	2.50	0.00
s_6	6	—	3.0 (B)	4.0 (A)	**3.50**	0.40
s_7	7	—	2.0 (C)	3.0 (B)	2.50	−0.60
s_8	8	—	2.0 (C)	3.0 (B)	2.50	−0.60
s_9	9	—	2.0 (C)	2.0 (C)	2.00	−1.10
s_{10}	10	—	2.0 (C)	2.0 (C)	2.00	−1.10

Table 1: Grades under a quota cap, with raw GPAs and a transcript-based difficulty-adjusted score. Bolded entries illustrate the GPA false equivalence: s_6 (rank 6) ties s_2 (rank 2) at 3.50. In this illustrative example, the final column is obtained from a two-way fixed-effects completion of the additive latent-performance model; in the exact complete-data benchmark, this completion recovers the same centred-ability direction as the eigengrade construction.

GPA distortions. The raw GPA column exhibits three familiar pathologies of capped, within-class grading:

1. *Reversals.* s_6 (true rank 6) posts a GPA of 3.50, tying s_2 (rank 2) and strictly beating s_3 (rank 3, GPA 3.00), s_4, and s_5 (ranks 4–5, GPA 2.50).

2. *False equivalence.* s_2 and s_6 share a 3.50 despite radically different transcript difficulty. Four students—s_4, s_5, s_7, s_8—are pooled at 2.50.

3. *Penalising ambition.* s_5 (rank 5) has the same GPA as s_7 and s_8 (ranks 7–8), even though s_5 competed in the hardest course. A rational student maximising GPA would avoid c_{Hard}.

Of the 45 ordered student pairs, raw GPA correctly ranks 34, with 3 reversed and 8 tied.

Constructing the difficulty adjustment. We treat each reported grade point as a coarse measurement of an additive latent index $P_{ic} \approx a_i - d_c$ and estimate relative course difficulties from the overlap cohorts. For transparency, the reported difficulty-adjusted scores in this example are computed via a two-way fixed-effects completion of this additive structure. In the exact additive benchmark, this completion yields the same centred-ability *direction* as the eigengrade construction of Theorem 2; however, we do not apply the spectral estimator directly to the sparse raw grade matrix here.

Hard versus Medium. The five overlap students s_1–s_5 earn $(4, 4, 3, 3, 3)$ in c_{Med} and $(4, 3, 3, 2, 2)$ in c_{Hard}. The pairwise grade differences $G_{i,\text{Med}} - G_{i,\text{Hard}}$ are $(0, 1, 0, 1, 1)$, with mean $\frac{3}{5}$. Under $P_{ic} = a_i - d_c$, this implies $\hat{d}_{\text{Hard}} - \hat{d}_{\text{Med}} = \frac{3}{5}$.

Medium versus Easy. The five overlap students s_6–s_{10} earn $(3, 2, 2, 2, 2)$ in c_{Med} and $(4, 3, 3, 2, 2)$ in c_{Easy}. The differences $G_{i,\text{Easy}} - G_{i,\text{Med}}$ are $(1, 1, 1, 0, 0)$, with mean $\frac{3}{5}$, so $\hat{d}_{\text{Med}} - \hat{d}_{\text{Easy}} = \frac{3}{5}$.

No student takes both c_{Hard} and c_{Easy}; the difficulty gap $\hat{d}_{\text{Hard}} - \hat{d}_{\text{Easy}}$ is therefore estimated by *chaining through* c_{Med}. This is a concrete instance of the identification logic in Proposition 5: connectivity of $\mathcal{G}$ suffices for identification, though in noisy settings chained comparisons may accumulate error across links.

Normalising $\hat{d}_{\text{Med}} = 0$ yields

$$\hat{d}_{\text{Hard}} = \tfrac{3}{5}, \qquad \hat{d}_{\text{Med}} = 0, \qquad \hat{d}_{\text{Easy}} = -\tfrac{3}{5}.$$

Each observed grade point G_{ic} then implies an adjusted ability signal $G_{ic} + \hat{d}_c$ (since $a_i \approx G_{ic} + d_c$). For example, s_4's C (2.0) in c_{Hard} adjusts to $2.0 + \frac{3}{5} = 2.6$, while s_6's A (4.0) in c_{Easy} adjusts to $4.0 - \frac{3}{5} = 3.4$. Define each student's difficulty-adjusted score as the centred average of adjusted signals across the transcript:

$$\hat{a}_i \;:=\; \frac{1}{|\{c : (i,c) \in \mathcal{G}\}|} \sum_{c:(i,c)\in\mathcal{G}} \big(G_{ic} + \hat{d}_c\big), \qquad \tilde{\alpha}_i \;:=\; \hat{a}_i - \frac{1}{n}\sum_{j=1}^{n} \hat{a}_j.$$

The values in the final column of Table 1 are exactly these centred adjusted averages, with grand mean $\bar{\hat{a}} = 2.80$.

Relation to the formal eigengrade definition. In matrix form, the construction above coincides with the least-squares two-way fixed-effects fit of $G_{ic} \approx a_i - d_c$ on the observed bipartite graph. Let $\widehat{P}_{ic} := \hat{a}_i - \hat{d}_c$ be the fitted completed matrix. Because $\widehat{P}$ has the additive structure assumed in Theorem 2 (cf. Proposition 5), applying the eigengrade operator to $\widehat{P}$ recovers the same centred-ability direction as $\tilde{\alpha}$, up to the usual eigenvector normalisation

and sign. The point of the example is therefore illustrative: it shows the difficulty-adjusted target that the eigengrade construction recovers in the exact additive benchmark, rather than claiming that the sparse ordinal matrix here has itself been analysed spectrally.

What the difficulty adjustment corrects. The transcript-based adjustment removes several of the most damaging GPA distortions:

- The s_2/s_6 false equivalence is broken decisively. Raw GPA pools them at 3.50; the difficulty-adjusted scores separate them to $\tilde{\alpha}_2 = 1.00$ versus $\tilde{\alpha}_6 = 0.40$.
- The s_3/s_6 reversal is corrected: $\tilde{\alpha}_3 = 0.50 > 0.40 = \tilde{\alpha}_6$, whereas raw GPA has s_6 a full half-point ahead.
- The four-way tie at GPA 2.50 (s_4, s_5, s_7, s_8) is split into two correctly ordered groups: $\tilde{\alpha}_4 = \tilde{\alpha}_5 = 0.00 > -0.60 = \tilde{\alpha}_7 = \tilde{\alpha}_8$. Students who competed in c_{Hard} are now credited for their more demanding transcripts.

Across all 45 ordered student pairs, the difficulty-adjusted scores correctly rank 40, with 2 reversed and 3 tied, an improvement over raw GPA's 34 correct, 3 reversed, and 8 tied.

What the adjustment does not correct. Residual distortions remain, and they are informative about the limits of the example.

(i) s_6 is still ranked above s_4 and s_5. Both discordant pairs involve s_6 versus students in the lower half of the c_{Hard} cohort. Student s_6 receives an A (4.0) in c_{Easy}, which adjusts to $4.0 - \frac{3}{5} = 3.4$, while s_4 receives a C (2.0) in c_{Hard}, which adjusts to $2.0 + \frac{3}{5} = 2.6$. The difficulty correction closes the raw two-point gap to 0.8, but cannot eliminate it entirely. The main reason is *grade quantisation*: s_4 sits just below the B/C cutoff in a five-person hard course and is rounded down to C, while s_6 sits atop a five-person easy course and is rounded up to A. The cap converts a modest true-ability gap into a two-grade-level gulf that an average difficulty adjustment cannot fully undo. With finer grade resolution or larger classes, this residual would typically shrink.

(ii) Adjacent ties cannot be broken. Students with identical transcripts—$\{s_4, s_5\}$, $\{s_7, s_8\}$, and $\{s_9, s_{10}\}$—receive identical difficulty-adjusted scores. The procedure has no within-transcript variation to exploit. Breaking these ties would require either finer grading within courses or additional information, such as within-course percentile rank, that the cap discards by construction.

Lessons for the general results. The example illustrates both the value and the limits of transcript-based difficulty adjustment in a setting with coarse grading, grade quantisation, and a binding cap.

First, *connectivity suffices for identification, but the precision of recovery depends on the richness of overlap.* The c_{Hard}–c_{Easy} difficulty gap is identified only by chaining through c_{Med}. In a noisy setting, such chained estimates may be less stable than direct comparisons. This is consistent with the broader message of Theorem 3, although that theorem is an oracle perturbation result rather than a literal finite-sample variance formula for a fixed sparse graph.

Second, *grade quantisation is an independent source of error.* Even with perfectly estimated difficulties, the coarse three-level grade scale destroys within-bin ordering information. This is a form of measurement loss absent from the continuous-performance benchmark in Theorems 2 and 3. Proposition 3 shows that the latent index remains identifiable under an ordered-response model, but the present example should not be read as a full efficiency analysis for sparse ordinal data.

Third, *transcript-based adjustment corrects systematic cross-course distortion better than raw GPA, but does not eliminate all ranking errors in coarse finite samples.* In this example there are no idiosyncratic shocks, so all distortion comes from the cap and the coarseness of the grading scheme. In practice, topical match, effort variation, and grading noise would introduce additional error, especially for students with thin transcripts. The perturbation theorem in Theorem 3 gives directional recovery in a dense latent-score benchmark with known observation structure; student-specific uncertainty in sparse transcript settings lies beyond the scope of that theorem.

Finally, the example treats the observed grade matrix $\mathbf{G}$ as a meaningful measurement object: cross-course overlap lets an observer estimate difficulty and recover a more informative ranking than raw GPA alone. But this logic presumes a sufficiently stable relationship between performance and grades. That raises a governance question: if instructors strategically adjust grading standards, when does $\mathbf{G}$ remain informative enough for transcript-based adjustment to work? The next section turns to that institutional prerequisite.

5 Strategic Grade Inflation and Policy Design

The transcript-based identification results in Section 4.2 treat the university's grade data as arising from a sufficiently stable measurement relationship. Formally, the outside observer sees a matrix $\mathbf{G}$ with entries $G_{ic} = g(P_{ic})$, where latent performance satisfies $P_{ic} = a_i - d_c + \varepsilon_{ic}$. When that relationship is stable enough, overlap in enrolments allows an observer to estimate

relative course difficulties and construct transcript-based adjustments. Strategic grading behaviour can undermine that premise.

We, therefore, complement Section 4.2 with a stylised game-theoretic analysis of instructor incentives. The purpose of this section is deliberately limited: to isolate one mean-grade externality and compare policy instruments that target it without mechanically compressing within-course variation. This is not a complete equilibrium theory of all ways grading systems can lose information. In particular, distributional compression, nonlinear threshold shifts, syllabus redesign, and other margins that may alter the informational content of $\mathbf{G}$ without moving the course mean are discussed explicitly below, but are not fully modelled in the reduced-form game.

5.1 The Grading Game

There are m instructors, one per course. In term t, instructor c chooses a course policy (grading leniency, curving, assessment design, and related margins) that determines the *mean raw grade* in that course, denoted $\bar{G}_{c,t} \in [\underline{G}, \overline{G}]$. If one prefers to parameterise strategies by an "effective difficulty" $\tilde{d}_{c,t}$, assume the mapping $\bar{G}_{c,t} = \phi(\tilde{d}_{c,t})$ is strictly decreasing so that choosing $\tilde{d}_{c,t}$ is equivalent to choosing $\bar{G}_{c,t}$.[12] Let $\bar{\mathbf{G}}_t := (\bar{G}_{1,t}, \ldots, \bar{G}_{m,t})$.

This reduced-form choice variable is intended to capture the margin relevant for grade inflation. It intentionally collapses several underlying instructor choices into a one-dimensional choice of the course mean. That is enough to analyse the externality in average grades, but it does not exhaust the ways instructors could erode information: distributional compression or threshold shifts that leave the mean unchanged are outside this reduced-form game and are discussed separately below. To connect to the identification model, recall that latent performance satisfies

$$P_{ic,t} = a_i - d_c + \varepsilon_{ic,t}, \qquad G_{ic,t} = g(P_{ic,t}),$$

where $g : \mathbb{R} \to G$ is the institutionally intended performance-to-grade mapping. Strategic inflation operates by altering the effective mapping from $P_{ic,t}$ to observed $G_{ic,t}$ (for example, via curving or leniency), shifting the distribution of $G_{ic,t}$ and hence

$$\bar{G}_{c,t} := \frac{1}{n_{c,t}} \sum_{i \in S_{c,t}} G_{ic,t},$$

without a corresponding shift in latent performance.

[12] To avoid confusion with the structural course-difficulty parameter d_c in the performance model $P_{ic,t} = a_i - d_c + \varepsilon_{ic,t}$, we reserve d_c for the fixed latent difficulty to be inferred from transcript structure, and use $\tilde{d}_{c,t}$ (or equivalently $\bar{G}_{c,t}$) for the instructor's policy choice in the grading game.

A key strategic force behind grade inflation is competition for enrolment, evaluations, or internal resources: holding content fixed, a course that awards higher grades becomes more attractive relative to its peers. We capture this with a standard logit-style share function

$$s_c(\bar{\mathbf{G}}_t) \;=\; \frac{\exp(\eta \bar{G}_{c,t})}{\sum_{k=1}^{m} \exp(\eta \bar{G}_{k,t})}, \qquad \eta > 0, \tag{7}$$

where $s_c(\bar{\mathbf{G}}_t)$ can be interpreted as the fraction of students, evaluation weight, or internal demand flowing to course c.

Instructor c's payoff has three components: (i) a direct benefit from higher grades, (ii) a competitive benefit from attracting a larger share of students, and (iii) a convex cost of deviating from a professional grading norm $\bar{G}_0$:

$$U_c(\bar{G}_{c,t}, \bar{\mathbf{G}}_{-c,t}) = \alpha \bar{G}_{c,t} + \beta \log s_c(\bar{\mathbf{G}}_t) - \frac{\gamma}{2}(\bar{G}_{c,t} - \bar{G}_0)^2, \tag{8}$$

with parameters $\alpha, \beta \geq 0$ and $\gamma > 0$.

In what follows, we suppress the time index t for notational convenience when doing so will not introduce confusion.

Proposition 7 (Grade inflation equilibrium)**.** *Consider the grading game* (7)–(8) *with* $\bar{G}_c \in [\underline{G}, \overline{G}]$.

1. *The game admits a unique Nash equilibrium.*

2. *The equilibrium is symmetric:* $\bar{G}_c = \bar{G}^{\mathrm{NE}}$ *for all* c*, where*

$$\bar{G}^{\mathrm{NE}} = \Pi_{[\underline{G},\overline{G}]} \left(\bar{G}_0 + \frac{\alpha + \beta\eta(1 - \frac{1}{m})}{\gamma} \right), \tag{9}$$

   and $\Pi_{[\underline{G},\overline{G}]}$ *denotes projection onto the feasible interval.*

3. *(Competitive externality.) Let* $\bar{G}^{\mathrm{SO}}$ *be the symmetric maximiser of total welfare* $\sum_{c=1}^{m} U_c$ *subject to symmetry. Then*

$$\bar{G}^{\mathrm{SO}} = \Pi_{[\underline{G},\overline{G}]} \left(\bar{G}_0 + \frac{\alpha}{\gamma} \right), \qquad \bar{G}^{\mathrm{NE}} - \bar{G}^{\mathrm{SO}} = \frac{\beta\eta(1 - \frac{1}{m})}{\gamma} \quad \textit{(when interior)},$$

   so equilibrium mean grades are weakly inflated relative to the symmetric planner benchmark whenever $\beta > 0$.

Proof. Define

$$\Phi(\bar{\mathbf{G}}) := \sum_{c=1}^{m} \left[(\alpha + \beta\eta)\bar{G}_c - \frac{\gamma}{2}(\bar{G}_c - \bar{G}_0)^2 \right] - \beta \log \left(\sum_{k=1}^{m} e^{\eta \bar{G}_k} \right).$$

Since

$$\log s_c(\bar{\mathbf{G}}) = \eta \bar{G}_c - \log \left(\sum_{k=1}^{m} e^{\eta \bar{G}_k} \right),$$

we have

$$\frac{\partial \Phi}{\partial \bar{G}_c} = \alpha + \beta\eta\Big(1 - s_c(\bar{\mathbf{G}})\Big) - \gamma(\bar{G}_c - \bar{G}_0) = \frac{\partial U_c}{\partial \bar{G}_c}.$$

Thus the game is an exact potential game with potential Φ.

The Hessian of Φ is

$$\nabla^2 \Phi = -\gamma I_m - \beta\eta^2\Big(\mathrm{Diag}(s) - ss^\top\Big), \qquad s := \Big(s_1(\bar{\mathbf{G}}), \ldots, s_m(\bar{\mathbf{G}})\Big)^\top.$$

For any $x \in \mathbb{R}^m$,

$$x^\top\Big(\mathrm{Diag}(s) - ss^\top\Big)x = \sum_{c=1}^{m} s_c \left(x_c - \sum_{j=1}^{m} s_j x_j \right)^2 \geq 0.$$

Hence $\nabla^2\Phi$ is negative definite because $\gamma > 0$, so Φ is strictly concave on the compact convex strategy set $[\underline{G}, \overline{G}]^m$. Therefore Φ has a unique maximiser, and because the game is exact potential, that maximiser is the unique Nash equilibrium.

The potential is invariant under permutations of the courses. By uniqueness, the equilibrium must therefore be symmetric: $\bar{G}_c = \bar{G}$ for all c. For an interior optimum, the first-order condition is

$$0 = \alpha + \beta\eta\left(1 - \frac{1}{m}\right) - \gamma(\bar{G} - \bar{G}_0), \tag{10}$$

which yields

$$\bar{G} = \bar{G}_0 + \frac{\alpha + \beta\eta(1 - 1/m)}{\gamma}.$$

Projection onto $[\underline{G}, \overline{G}]$ gives (9) when the interior solution is infeasible.

For the planner benchmark under symmetry, note that if all $\bar{G}_c$ are equal then $s_c(\bar{\mathbf{G}}) = 1/m$ for every c, so $\sum_c \log s_c(\bar{\mathbf{G}}) = m \log(1/m)$ is constant and drops out. The symmetric planner therefore chooses $\bar{G}$ to maximise $m\Big(\alpha\bar{G} - \frac{\gamma}{2}(\bar{G} - \bar{G}_0)^2\Big)$, giving $\bar{G}^{\mathrm{SO}} = \bar{G}_0 + \alpha/\gamma$ (projected to feasibility). The inflation gap follows by subtraction. □

Proposition 7 isolates a simple first-moment externality. In the symmetric equilibrium, the realised grade matrix $\mathbf{G}_t$ inherits an upward shift in its column means. Controlling that drift

helps preserve more of the variation in $\mathbf{G}_t$ that transcript-based adjustment uses, although it does not by itself guarantee that all relevant information in the grade distribution is preserved.

A Second Concern about Grade Caps

As discussed in Section 4.1, grade caps have a clear informational cost: they transform a universally calibrated performance standard into a course- and cohort-specific relative ranking, violating Axiom 2 and mechanically generating the kind of cross-course reversals highlighted by Theorem 1. The grading game above suggests a second concern. Caps constrain the *reported distribution* of grades, but they need not remove the underlying competitive pressure that makes generous grading attractive in the first place.

That point should be stated carefully. The present reduced-form game models strategic behaviour through the course mean $\bar{G}_c$, so it does not formally prove what happens when instructors substitute toward other margins. But as a policy matter the concern is straightforward: if instructors cannot relax standards by reallocating reported letter grades, they may instead adjust substantive difficulty, assessment design, or threshold placement in ways that preserve the externality while moving it off the most visible margin. A cap can therefore suppress observed grade inflation while still leaving incentive pressure to soften effective standards on dimensions outside the reported distribution. The cost is that the cap has already degraded the cross-course interpretability of the transcript.

What is needed, instead, is an instrument that directly targets the mean-grade externality while preserving substantially more of the raw informational content of $\mathbf{G}_t$. One could also imagine transcript-based difficulty reporting weakening demand for easy grading, but that mechanism would require a richer model of student beliefs and course choice than we study here—although see Section 6.3 for a discussion of how eigengrades may have precisely the sort of feedback effect.

5.2 A Taylor Rule for Grades

The grading game suggests two desiderata for an anti-inflation policy. First, the policy should restrain the competitive externality that pushes mean grades above the social benchmark. Second, it should do so without mechanically dictating the within-course distribution of grades. Grade caps perform poorly on both margins. A different approach is to target the *mean* of raw grades while leaving the internal shape of the grade distribution largely unconstrained.

Our instrument borrows the logic of feedback rules from monetary policy. The analogy is limited but useful. In the macroeconomic setting, the Taylor rule (Taylor, 1993) raises the

policy rate when inflation exceeds a target, creating a stabilising force without micromanaging individual prices. The grading analogue is a one-sided penalty on course mean grades above a target $\bar{G}^\star \in [\underline{G}, \overline{G}]$:

$$\text{Penalty}_{c,t} = \lambda \cdot (\bar{G}_{c,t} - \bar{G}^\star)_+, \tag{11}$$

where $(x)_+ = \max\{x, 0\}$ and $\lambda > 0$. The modified payoff is

$$U_c^\tau(\bar{G}_{c,t}, \bar{\mathbf{G}}_{-c,t}) = U_c(\bar{G}_{c,t}, \bar{\mathbf{G}}_{-c,t}) - \lambda(\bar{G}_{c,t} - \bar{G}^\star)_+.$$

Here the penalty is written directly into instructor utility, so we are implicitly assuming that the university has some instrument with real utility consequences. Alternatively, one could imagine an institutionally imposed consequence that operates indirectly through evaluations, budgeting, or staffing.[13]

The penalty is one-sided: it activates only when the mean grade exceeds the target. Unlike a grade cap, this mechanism does *not* constrain the within-course distribution of grades. It is therefore a first-moment instrument rather than a full distributional rule.

Proposition 8 (Target implementation under a one-sided mean penalty)**.** *Consider the modified grading game with payoffs*

$$U_c^\tau(\bar{G}_c, \bar{\mathbf{G}}_{-c}) = \alpha\bar{G}_c + \beta \log s_c(\bar{\mathbf{G}}) - \frac{\gamma}{2}(\bar{G}_c - \bar{G}_0)^2 - \lambda(\bar{G}_c - \bar{G}^\star)_+, \qquad \bar{G}_c \in [\underline{G}, \overline{G}].$$

1. *The modified game is an exact potential game with strictly concave potential*

$$\Phi^\tau(\bar{\mathbf{G}}) := \sum_{c=1}^{m} \left[(\alpha + \beta\eta)\bar{G}_c - \frac{\gamma}{2}(\bar{G}_c - \bar{G}_0)^2 - \lambda(\bar{G}_c - \bar{G}^\star)_+\right] - \beta \log\left(\sum_{k=1}^{m} e^{\eta\bar{G}_k}\right).$$

   Consequently the modified game admits a unique Nash equilibrium. Because the primitives and the target are common across courses, that equilibrium is symmetric.

2. *Let $\bar{G}^{\text{NE}}$ be the unpenalised symmetric equilibrium mean from Proposition 7. Suppose $\bar{G}^\star \in [\underline{G}, \overline{G}]$ and $\bar{G}^\star \le \bar{G}^{\text{NE}}$. If*

$$\lambda \;\ge\; \alpha + \beta\eta\left(1 - \frac{1}{m}\right) - \gamma(\bar{G}^\star - \bar{G}_0), \tag{12}$$

[13] One additional issue, which we do not address here, is how to treat courses whose instructor, subject matter, or assessment structure changes enough across years that the course has no meaningful steady-state benchmark.

then the unique Nash equilibrium of the modified game is the target profile

$$\bar{G}_c = \bar{G}^\star \qquad \textit{for all } c = 1, \dots, m.$$

Proof. For any instructor c, any $\bar{\mathbf{G}}_{-c}$, and any two feasible choices $x_c, y_c \in [\underline{G}, \overline{G}]$,

$$\Phi^\tau(x_c, \bar{\mathbf{G}}_{-c}) - \Phi^\tau(y_c, \bar{\mathbf{G}}_{-c}) = U_c^\tau(x_c, \bar{\mathbf{G}}_{-c}) - U_c^\tau(y_c, \bar{\mathbf{G}}_{-c}).$$

Thus the modified game is an exact potential game.

The function Φ from Proposition 7 is strictly concave, and the additional term

$$-\lambda \sum_{c=1}^{m} (\bar{G}_c - \bar{G}^\star)_+$$

is concave. Hence Φ^τ is strictly concave on the compact convex set $[\underline{G}, \overline{G}]^m$, so it has a unique maximiser.

It remains to show that Nash equilibria coincide with maximisers of Φ^τ. Fix a Nash equilibrium $\bar{\mathbf{G}}^\dagger$. For each c, the one-dimensional function $x_c \mapsto U_c^\tau(x_c, \bar{\mathbf{G}}_{-c}^\dagger)$ is concave, so there exists $q_c \in \partial_{\bar{G}_c} \Phi^\tau(\bar{\mathbf{G}}^\dagger)$ such that

$$q_c \, (y_c - \bar{G}_c^\dagger) \le 0 \qquad \text{for all } y_c \in [\underline{G}, \overline{G}].$$

Summing over c gives

$$q^\top (y - \bar{\mathbf{G}}^\dagger) \le 0 \qquad \text{for all } y \in [\underline{G}, \overline{G}]^m,$$

which is the variational inequality characterising a maximiser of the concave function Φ^τ. Conversely, every maximiser of Φ^τ is a Nash equilibrium by exact-potential equivalence. Therefore the modified game has a unique Nash equilibrium, namely the unique maximiser of Φ^τ. Because Φ^τ is invariant under permutations of the courses, that unique equilibrium must be symmetric.

For part (ii), fix c and suppose all other instructors set $\bar{G}_k = \bar{G}^\star$ for $k \neq c$. Write

$$V(\bar{G}_c) := U_c^\tau(\bar{G}_c, \bar{G}^\star \mathbf{1}_{-c}).$$

The function V is concave in $\bar{G}_c$, so $\bar{G}_c = \bar{G}^\star$ is a best response if and only if the left derivative at $\bar{G}^\star$ is weakly nonnegative and the right derivative is weakly nonpositive.

At the symmetric target profile, $s_c = 1/m$. For $\bar{G}_c < \bar{G}^\star$ the penalty is inactive, so

$$V'_-(\bar{G}^\star) = \alpha + \beta\eta\Big(1 - \frac{1}{m}\Big) - \gamma(\bar{G}^\star - \bar{G}_0).$$

If $\bar{G}^\star < \underline{G}$ the target is infeasible, so suppose $\bar{G}^\star \in [\underline{G}, \overline{G}]$ as stated. When $\bar{G}^\star > \underline{G}$, the assumption $\bar{G}^\star \le \bar{G}^{\mathrm{NE}}$ implies $V'_-(\bar{G}^\star) \ge 0$; if $\bar{G}^\star = \underline{G}$, there is no feasible deviation below the target and this condition is vacuous.

For $\bar{G}_c > \bar{G}^\star$ the penalty is active, so

$$V'_+(\bar{G}^\star) = V'_-(\bar{G}^\star) - \lambda.$$

Condition (12) implies $V'_+(\bar{G}^\star) \le 0$. Hence $\bar{G}^\star$ is a best response to $(\bar{G}^\star, \dots, \bar{G}^\star)$, so the target profile is a Nash equilibrium. By part (i), the modified game has a unique Nash equilibrium, so that equilibrium must be exactly the target profile. □

Proposition 8 should be read as a first-moment implementation result for the stylised symmetric game. It does *not* by itself solve the cross-course confounding of ability and difficulty formalised in Theorem 1. Nor does a mean-targeting rule rule out every information-destroying response: instructors could compress the within-course distribution or shift effective thresholds in ways that leave $\bar{G}_c$ unchanged. Institutions that care about preserving measurement quality may therefore wish to monitor grade dispersion or related diagnostics alongside course means. For cross-course comparability, one still needs the transcript-based difficulty adjustment developed in Section 4.2.

The same exact-potential argument extends mechanically to predetermined course-specific targets $\{\bar{G}^\star_c\}_c$, in which case the equilibrium remains unique but is generally asymmetric. We retain the common-target case because it is the cleanest benchmark and maps directly to the inflation logic in Proposition 7.

5.3 Combining Transcript Adjustment with Mean-Targeting

The preceding analysis identifies two institutional objects that should not be conflated. First, the university may want an *externally reported comparison signal* that adjusts, as far as possible, for course difficulty. Second, it may want an *internal governance target* on the raw-grade scale that disciplines instructor incentives. These are different tasks, and in general they live on different scales.

For external reporting, let $T_{ic,t}$ denote the transcript-based adjusted signal released for student i in course c at date t. If the difficulty estimates are obtained on the same cardinal

scale as raw grades, a convenient choice is

$$T_{ic,t} = G_{ic,t} + \hat{d}_{c,t}^{\text{raw}} - \bar{d}_t^{\text{raw}}, \qquad \bar{d}_t^{\text{raw}} := \frac{1}{m}\sum_{c=1}^{m} \hat{d}_{c,t}^{\text{raw}}.$$

However, when the first stage is an ordered-response or other latent-index model, $(\hat{a}_{i,t}, \hat{d}_{c,t})$ are identified only on the latent-index scale fixed by that model's normalisation. In that case the arithmetic adjustment $G_{ic,t} + \hat{d}_{c,t} - \bar{d}_t$ should *not* be interpreted literally as a raw-grade-point correction. The institution should instead report a monotone calibrated transform of the fitted latent index, such as a posterior-ability score, percentile, or other transcript score derived from $(\hat{a}_{i,t}, \hat{d}_{c,t})$.

This distinction matters for incentives. The Taylor-style penalty must be applied to the raw grade distribution $G_{ic,t}$, because that is the object instructors manipulate directly. Applying the penalty to an adjusted transcript score $T_{ic,t}$ would allow the adjustment step itself to absorb part of the strategic inflation and thereby weaken the incentive effect of the rule. If instructor c inflates raw grades in term t, a transcript model re-estimated on the same term's data can partially rationalise the change as lower estimated difficulty. That is precisely why the penalty must be computed from raw grades and predetermined targets rather than from contemporaneously adjusted scores.

5.4 A Two-Signal Grading Architecture

The impossibility result in Section 3 suggests a two-signal architecture rather than a single reported scalar:

1. **External reporting.** Use transcript data to estimate latent abilities and course difficulties, and report the adjusted signal $T_{ic,t}$. This abandons universal calibration for the *reported* signal in order to recover cross-course comparability as well as the transcript model permits.

2. **Internal governance.** Apply the mean-targeting mechanism to the raw grade matrix $\mathbf{G}_t$, but allow the course-specific target to depend on lagged estimated class ability through an explicit *scale calibration* step.

To formalise the second component, let

$$\bar{a}_t := \frac{1}{n}\sum_{i=1}^{n} \hat{a}_{i,t}, \qquad \bar{a}_{c,t+1|t} := \frac{1}{n_{c,t+1}} \sum_{i \in S_{c,t+1}} \big(\hat{a}_{i,t} - \bar{a}_t\big),$$

where $S_{c,t+1}$ is the roster for course c in term $t+1$, fixed before instruction begins (for example, at the close of registration or at the add/drop deadline). When $(\hat{a}_{i,t})$ come from a latent-index model, centring fixes location but not scale. We therefore introduce a predetermined coefficient $\kappa_t > 0$ that converts one unit of centred latent ability into raw-grade points. The coefficient κ_t may be fixed institutionally or estimated from historical data using only information available by the end of term t, for example from a lagged calibration regression of course-mean raw grades on lagged course-average centred ability.

Using this calibration, set the term-$(t+1)$ target to

$$\bar{G}^{\star}_{c,t+1} = \Pi_{[\underline{G},\overline{G}]}\left(\bar{G}^{\star} + \kappa_t\, \bar{a}_{c,t+1|t}\right),$$

where

$$\Pi_{[\underline{G},\overline{G}]}(x) := \min\{\overline{G}, \max\{\underline{G}, x\}\}$$

denotes projection onto the feasible grade interval.

This formulation resolves the scale problem inherent in latent-index first stages: a location normalisation such as $\sum_i \hat{a}_{i,t} = 0$ is not enough to place $\hat{a}_{i,t}$ on the raw-grade scale. The coefficient κ_t is what performs that conversion.

Timing and scope. At the end of term t, the institution updates $(\hat{a}_{i,t}, \hat{d}_{c,t})$ and the calibration coefficient κ_t using only data available through term t. These objects are then treated as predetermined when setting $\bar{G}^{\star}_{c,t+1}$. Instructors choose term-$(t+1)$ grading policies against those predetermined targets, generating raw grades $G_{ic,t+1}$ and means $\bar{G}_{c,t+1}$. Only after the term is complete do the term-$(t+1)$ observations enter the next update.

This timing blocks *contemporaneous* neutralisation: an instructor who inflates grades in term $t+1$ cannot change the target against which the term-$(t+1)$ penalty is assessed. Any transcript-based re-estimation of difficulty occurs only after the current penalty has already been applied.

In the static reduced-form game, replacing the common target $\bar{G}^{\star}$ by a predetermined vector $\{\bar{G}^{\star}_{c,t+1}\}_c$ preserves the exact-potential structure, and therefore preserves uniqueness of equilibrium, though the equilibrium is then generally asymmetric. What we do *not* solve here is the full dynamic joint updating problem for $(\hat{a}_{i,t}, \hat{d}_{c,t}, \kappa_t)$ when the target vector itself is recalibrated over time. We therefore interpret the ability-indexed mean-targeting rule as a disciplined institutional architecture rather than as a complete dynamic equilibrium theorem.

In short, raw grades remain the internally governed measurement system, while $T_{ic,t}$ is the externally reported comparison signal.

6 Extensions

6.1 Grading in Matching Markets

The analysis so far treats the grade as a signal consumed by a generic outside observer. In practice, grades feed into *matching markets*: students are matched to positions (jobs, graduate-school slots, clerkships) on the basis of transcript information. Ostrovsky and Schwarz (2010) (hereafter OS) provide the canonical equilibrium theory of information disclosure in exactly this setting.

In the OS framework, schools are modelled as strategic actors who design grading policies (transcript structures) to maximise the aggregate desirability of the job placements obtained by their graduates. A central finding of OS is that schools often have a strategic incentive to compress grades—intentionally mixing students of different abilities into the same grade bins—to lift the placement outcomes of average students.

While OS provides the "macro" theory of between-school competition, our paper provides the "micro-foundation" of within-school grading. The OS framework assumes that schools are frictionless, monolithic entities that perfectly control the information content of their transcripts. In reality, schools face severe internal frictions: courses vary in difficulty, students strategically sort into classes, and individual instructors face distinct incentives to inflate grades.

By viewing our within-school mechanisms through the lens of the OS matching market, we can answer a critical institutional question: how do the structural limitations of grading and our proposed solutions (eigengrades and the Taylor rule) interact with a school's broader strategic goals? In this section, we synthesise the economic intuition of these interactions. To formalise these insights, Section A embeds our grading model into the continuous-time matching framework of OS, providing formal statements and proofs for all results discussed below.

6.1.1 Structural vs. Strategic Noise

When an employer looks at a transcript and finds it noisy or uninformative, that noise can originate from two entirely different sources. OS establish a *strategic constraint*: schools deliberately introduce noise (strategic compression) to optimise job placements. Our impossibility result (Theorem 1) establishes a *structural constraint*: a universally calibrated scalar grade inevitably confounds ability with difficulty, creating unintentional noise simply by mixing easy and hard courses.

These two concepts are highly complementary, but structural noise is deeply destructive

to a school's strategic goals. If a school wishes to strategically pool specific tiers of students to maximise their market outcomes, it must be able to control *who* is pooled with whom. As we formally show in Proposition 10 in the Appendix A, a universally calibrated grading rule fails to account for heterogeneous course difficulty and therefore cannot implement the OS-optimal equilibrium distribution of posterior abilities. Real-world grades are noisy for two independent reasons, and an institution must resolve the structural confounding of its curriculum before it can effectively exert strategic control over its disclosures.

6.1.2 Course Selection as an Information Leak

A core assumption in the standard matching literature is that the school has a monopoly on the information signal: employers only know what the school's transcript tells them. Our analysis of endogenous selection (Proposition 2) highlights a critical departure from this assumption: the choice of course can itself act as an independent signal of ability.

In a matching market, if ambitious, high-ability students systematically sort into harder courses, endogenous course selection acts as an uncontrolled information leak that bypasses the school's intended disclosure policy. As demonstrated in Proposition 12, if employers observe course difficulty alongside grades, their posterior estimates of student ability are strictly refined. This additional disclosure channel can unravel the school's strategic pooling. Our impossibility and selection results arise precisely because the grading rule is forced to be universal, turning students' course choices into a highly informative signal that pushes the market away from the school's optimal disclosure environment.

6.1.3 Filtering the Right Noise with Eigengrades

A sceptical reader might ask: if schools actually *want* to compress information to maximise placements (as OS predict), will transcript-based difficulty adjustments like eigengrades run contrary to the school's objective by revealing too much?

The answer is: not automatically. It is vital to distinguish between removing *structural* noise and undoing *strategic* information design. The eigengrade estimator is built to strip out curriculum-driven additive heterogeneity (course fixed effects such as d_c) while remaining agnostic about *monotone* post-processing of the resulting signal.

What eigengrades cannot do is resurrect information that the school deliberately destroys through pooling or quantisation. Hard coarsening compresses within-course variation and can wipe out the local covariance structure that spectral methods exploit. Proposition 13 should therefore be read as a robustness result with a negative edge: additive-noise versions of OS-style compression do *not* defeat spectral recovery, so such implementations are not

effective ways of hiding the underlying ability ranking. More general OS equilibria based on deterministic partition rules need not satisfy these conditions, and sufficiently aggressive pooling may make any transcript-based recovery impossible, regardless of estimator.

Once the school has a difficulty-adjusted signal, it remains free to apply additional disclosure policies, but the more those policies compress the signal, the less any outside observer can recover.

6.1.4 The Taylor Rule as an Implementation Mechanism

Finally, embedding our model into a matching market clarifies the necessity of the Taylor rule. The OS framework assumes that a central university planner (e.g., a Dean) can magically dictate the grading curve to achieve the school's optimal placement strategy. In reality, grading is typically decentralised: i.e., *instructors* assign grades. As modelled in Section 5, instructors face a competitive prisoner's dilemma to inflate raw grades to attract enrolments. Left unchecked, this instructor-level inflation can subvert the school-level strategic disclosure policy.

This hierarchical tension reveals the Taylor rule as crucial to institutional implementation. Proposition 14 in Section A.5 shows that setting the Taylor-rule target to the OS-equilibrium *mean* can rein in rogue instructors and halt the decentralised arms race in average grades. However, because the Taylor rule targets only the first moment, it does not in general implement the full within-course transcript *distribution* required by the OS equilibrium (variance, compression, and shape remain unconstrained). Furthermore, the eigengrade-enhanced Taylor rule moves the institution even closer to the OS logic: by allowing course-specific grade targets that reflect the estimated ability composition of each class, schools can award higher grades to higher-ability cohorts without distorting the market. The Taylor rule provides the practical governance architecture required to make matching-market grading work.

6.2 Extension: Multi-Dimensional Abilities and Skill Matches

The additive transcript model developed in Section 4.2 assumes that student ability and course difficulty can each be compressed into a single scalar fixed effect (a_i and d_c). While this captures baseline academic strength and overall grading leniency, it rules out comparative advantage. In practice, a student may have a strong portable component of general preparation together with subject-specific strengths and weaknesses: a mathematics prodigy may find a rigorous topology course perfectly manageable but an introductory poetry seminar exceptionally difficult.

To capture this richer environment, in Section B, we extend the model to a Multi-Dimensional Additive Interaction (MDAI) specification. In that extension, a student's performance is driven by a scalar baseline ability a_i, a scalar baseline course difficulty d_c, and a multi-dimensional match $u_i^\top v_c$ between the student's skill profile and the course's skill demands. This decomposition matters conceptually. The scalar term a_i remains the natural analogue of broad academic strength; the vector term u_i captures comparative advantage, not a second hidden scalar ranking in disguise.

Moving from a one-dimensional ability metric to a multi-dimensional skill space changes the object of identification. In the scalar model, the goal is recovery of a single centred ability direction. In the MDAI model, by contrast, we do *not* in general obtain a canonical total ordering of students from the interaction term alone. Vector skill profiles typically support only partial comparisons: one student may dominate another in some course families and trail in others. Section B therefore formalises the multi-dimensional extension as a problem of recovering a latent interaction *subspace* and then, where desired, constructing benchmark-specific scalar summaries on top of that subspace.

Recovering the Interaction Subspace via Double-Centring. To study comparative advantage, one must first filter out the baseline effects. In the one-dimensional model, centring isolates the centred ability vector. In the MDAI model, one instead double-centres the latent performance matrix, removing the additive row and column effects associated with a_i and d_c. What remains is a residual matrix of interaction effects: did the student perform better or worse in this course than their general level and the course's general difficulty would predict?

Applying a Singular Value Decomposition (SVD) to that residual matrix recovers the leading low-rank interaction structure. We term the resulting student-side singular vectors *Multi-Dimensional Eigengrades* (MDE). Their role is geometric rather than ordinal: they locate students in a latent skill space and recover the comparative-advantage subspace, but they are not, by themselves, a scalar ranking of "true ability."

Data Requirements: Beyond Simple Connectivity. When measuring a single dimension of baseline ability, the identification results in Section 4.2 require only that the enrolment graph be connected: one bridge course between two cohorts is enough to chain relative difficulties together.

Multi-dimensional eigengrades demand substantially richer transcripts. To identify a K-dimensional interaction structure, a single bridge course is not enough. As Section B shows, every student must be observed in at least K sufficiently informative courses, and every course must enrol at least K sufficiently diverse students, with overlap rich enough

to span the relevant skill dimensions. If transcripts are too sparse, the low-rank matrix completion problem is underdetermined. Provided the overlap is sufficiently rich and, in the latent-score benchmark of Section B, missingness is completely at random (MCAR), the MDE estimator remains statistically consistent for the interaction subspace.

Scalar Summaries Require an Explicit Benchmark. If an institution wants a one-dimensional summary in the multi-dimensional environment, it must choose that summary explicitly rather than treating it as a free by-product of the model. A natural construction is to fix a reference bundle of courses, or equivalently weights ω_c over courses, and define student i's benchmark score by expected performance on that bundle after netting out the common average course difficulty:

$$s_i(\omega) = a_i + \sum_{c=1}^{m} \omega_c \, u_i^\top v_c.$$

This is a domain-specific scalarisation, not a metaphysical universal ordering of persons. Shared core courses are especially useful here. If ω is concentrated on a common core, then $s_i(\omega)$ measures expected performance on the very set of courses that all students share, yielding a natural common benchmark for external comparison.

Naming the Dimensions and Communicating Them. The final challenge is interpretability. Spectral methods recover the relevant K-dimensional subspace, but they do not uniquely label the axes. The same interaction matrix can be written in different coordinate systems, so "Dimension 1" need not correspond to a primitive skill in any institutionally meaningful sense.

To make MDE useful for outside observers, the institution must impose additional structure if it wants interpretable coordinates. One approach is to designate *anchor courses*. For example, if a university treats a specific real-analysis course as primarily quantitative and a creative-writing course as primarily verbal, these anchors help name the recovered axes. But even with anchored coordinates, the substantive comparison problem remains benchmark-dependent: the coordinates describe a profile, whereas any scalar ranking requires either a baseline measure such as a_i or an explicit benchmark bundle such as the core-course score above.

6.3 Student Course-Selection Incentives Under Eigengrades

The analysis so far has focused on the inference problem of an outside observer and the strategic incentives of instructors. Students are strategic actors too: they choose which

courses to take, and those choices respond to the grading regime. Under raw GPA, this response is well documented and can be harmful to information quality. We now show that difficulty-adjusted transcript scores (eigengrades) can eliminate the first-order course-shopping incentive and, in noisy environments where students care about signal precision, can even tilt choices toward harder courses.

The GPA-shopping baseline. Under a universally calibrated grading rule g, student i's GPA contribution from course c is $g(\pi(a_i, d_c))$. Because π is decreasing in d and g is weakly increasing, this contribution is weakly decreasing in d_c. A GPA-maximising student therefore weakly prefers easier courses, all else equal—the "flight from strict graders" documented by Sabot and Wakeman-Linn (1991). This GPA-shopping force is one form of endogenous selection. Proposition 2 uses a different selection pattern—positive sorting of higher-ability students into harder courses—to show that endogenous course choice can overturn the monotone grade–ability relationship. The common lesson is that once course choice is endogenous, grades need not remain monotone in ability.

Incentive neutrality under exact difficulty adjustment. The eigengrade construction adjusts observed outcomes by (estimated) course difficulty. In the benchmark additive model, perfect difficulty adjustment eliminates the course-shopping incentive entirely.

Proposition 9 (Incentive neutrality)**.** *Work in the noiseless additive transcript model $P_{ic} = a_i - d_c$, and suppose the transcript mechanism recovers (a, d) exactly up to the usual location normalisation (e.g., the enrolment graph remains connected so the two-way fixed effects are identified). Let the reported eigengrade score be the centred ability component $\alpha_i := a_i - \bar{a}$, where $\bar{a} := \frac{1}{n}\sum_{j=1}^{n} a_j$ (as in Theorem 2 after fixing eigenvector scale/sign). Then, holding other students' enrolments fixed, student i's reported α_i is unchanged by any unilateral change in their course portfolio that preserves identification. In particular, under ideal eigengrades a student cannot improve their reported score by avoiding harder courses.*

Proof. In the noiseless additive model, each observed entry satisfies $P_{ic} + d_c = a_i$. Hence any difficulty-adjusted course outcome is the same across courses for a given student. More formally, fix a portfolio $\mathcal{C}_i$ and consider the system of equations $P_{jc} = a_j - d_c$ over the observed enrolment graph. When the graph is connected, any two solutions (a, d) and (a', d') differ by a common additive constant: $a' = a + \kappa\mathbf{1}$ and $d' = d + \kappa\mathbf{1}$. Imposing the normalisation $\sum_j a'_j = 0$ pins down $\kappa = -\bar{a}$, so the uniquely recovered centred ability is $a_i - \bar{a}$, independent of $\mathcal{C}_i$. The same argument applies to any other portfolio $\mathcal{C}'_i$ that leaves identification intact. □

Proposition 9 makes the incentive implication explicit: under exact eigengrades, a student gains nothing from avoiding a hard course, because the difficulty adjustment fully offsets the raw-grade penalty. The "penalising ambition" pathology illustrated in Table 1—where student s_5 receives the same raw GPA as students s_7 and s_8 despite competing in the hardest course—is corrected once difficulty is adjusted for.

Noisy eigengrades and the value of hard courses. With estimation noise, the picture is richer. For intuition, work on the latent performance (or grade-point) scale and consider the simple difficulty-adjusted average score (cf. the computation in Table 1). Let

$$P_{ic} = a_i - d_c + \varepsilon_{ic}, \qquad \hat{d}_c^{-i} = d_c + \delta_c^{-i},$$

where $\hat{d}_c^{-i}$ is a *leave-one-out* estimate of course difficulty computed without using student i's own record, and δ_c^{-i} is its estimation error.[14] Define the uncentred adjusted score

$$\hat{a}_i := \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \left(P_{ic} + \hat{d}_c^{-i} \right) = a_i + \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \left(\varepsilon_{ic} + \delta_c^{-i} \right),$$

and centre it as $\hat{\alpha}_i := \hat{a}_i - \frac{1}{n} \sum_{j=1}^{n} \hat{a}_j$. If $\mathbb{E}[\varepsilon_{ic}] = 0$ and the difficulty estimator is conditionally unbiased $\mathbb{E}[\delta_c^{-i} \mid a, d] = 0$, then $\mathbb{E}[\hat{a}_i \mid a, d] = a_i$ for every portfolio $\mathcal{C}_i$: there is no *first-order* incentive to shop for easy courses.

What can vary with $\mathcal{C}_i$ is the *precision* of the reported signal. Under the simplifying approximation that the terms $\varepsilon_{ic} + \delta_c^{-i}$ are independent across courses,

$$\mathrm{Var}(\hat{a}_i \mid a, d, \mathcal{C}_i) = \frac{1}{|\mathcal{C}_i|^2} \sum_{c \in \mathcal{C}_i} \mathrm{Var}(\varepsilon_{ic} + \delta_c^{-i}).$$

More generally, because the leave-one-out difficulty estimates are computed jointly from the transcript graph and under a normalisation, cross-course covariance terms $\sum_{c \neq c'} \mathrm{Cov}(\varepsilon_{ic} + \delta_c^{-i}, \varepsilon_{ic'} + \delta_{c'}^{-i})$ need not vanish. The displayed formula is therefore best read as a clean intuition for the diagonal component of precision, not as a literal decomposition in every implementation. Even with that caveat, students who (or whose employers) value precision will prefer portfolios that reduce the variance and covariance of the combined noise terms. Two empirically relevant channels point in the direction of harder and/or more connected courses:

1. **Difficulty-estimation precision.** Courses whose difficulty is well identified, typically

[14] Using leave-one-out (or, more generally, cross-fitting) removes a mechanical feedback channel in which enrolling in a small course changes $\hat{d}_c$ and hence one's own adjusted score.

because they have large and diverse enrolments and create rich overlap in the observation graph, have smaller $\mathrm{Var}(\delta_c^{-i})$. Taking such a course yields a more precise difficulty adjustment, improving the accuracy of $\hat{a}_i$ (and, by Theorem 3, the overall eigengrade estimates).

2. **Grade coarseness and top-coding.** With coarse letter grades, the effective measurement noise in P_{ic} is larger when most students in a course fall into the same grade bin. For a high-ability student this compression is often most severe in very easy courses, where much of the class is top-coded at the highest grade. Harder courses (or courses with greater within-class ability dispersion) reduce top-coding and preserve more within-course variation, making the resulting transcript entry more informative about relative ability.

If students (or employers) attach value to precision, then course choice under eigengrades can favour harder or better-connected courses rather than easier ones, reversing the raw-GPA "flight from strict graders" incentive.

Remark 15 (A virtuous cycle). Under raw GPA, course-shopping pushes students toward easy courses; if this concentrates students within narrow difficulty bands, it can reduce overlap across courses and weaken transcript-based difficulty identification and estimation. Under eigengrades, expected scores do not penalise challenging courses, and students who care about precision may even prefer courses that improve overlap and identification. These endogenous enrolment patterns can increase overlap and improve the precision of the eigengrade estimator (Theorem 3) for *all* students, without requiring mandated course assignments.

Remark 16 (Limits of incentive neutrality). Proposition 9 relies on the additive structure $P_{ic} = a_i - d_c + \varepsilon_{ic}$ and on difficulty adjustment being (unbiased and) accurate. If course difficulty interacts with ability in a non-additive way (e.g., a course that is disproportionately hard for lower-ability students), or if the difficulty estimator is systematically biased under severe MNAR selection (Remark 13), residual course-shopping incentives can persist. Nevertheless, even imperfect difficulty adjustment attenuates the raw-GPA shopping incentive, and the directional conclusion—that eigengrades weaken or can reverse the flight from hard courses—is robust to moderate departures from the additive benchmark.

7 Implications and Broader Grading Design Principles

The eigengrade construction and the Taylor-rule mechanism developed in Sections 4.2 and 5 represent one coherent response to the impossibility theorem, but they require institutional

infrastructure—a connected transcript database, a spectral or fixed-effects estimator, and a feedback rule on course means—that many institutions may not be ready to deploy. Nevertheless, the theoretical analysis yields a set of more general design principles that apply to any grading system, and that can guide incremental reforms even in the absence of full transcript-based difficulty adjustment.

Distributional grade caps destroy information without addressing the underlying externality. Our analysis identifies two independent failures of grade caps and other forced curve mechanisms. On the information side, a cap that fixes the fraction of students receiving each letter grade transforms a universally calibrated performance standard into a course- and cohort-specific relative ranking (Section 4.1). The effective performance threshold for a given grade becomes endogenous to the composition of the enrolled class, so the same letter grade in two courses no longer reflects a common performance standard—precisely the cross-course confounding that Theorem 1 identifies as the core problem. Worse, when course selection is endogenous, a cap institutionalises ability–grade reversals: a student of moderate ability in a weak cohort can receive a higher letter grade than a stronger student competing in a more demanding classroom. On the incentive side, caps fail to neutralise the competitive externality that drives inflation in the grading game (Section 5.1). Instructors retain the margin of reducing substantive difficulty so that the capped quota of high grades is achieved at lower true performance, shifting strategic behaviour from observable grade inflation to less observable difficulty deflation. The combination of informational damage and incentive leakage makes distributional caps a dominated instrument relative to the alternatives discussed below.

Target course-level means, not distributions, and report difficulty context alongside grades. If restraining grade inflation is the policy objective, the Taylor-rule logic of Section 5.2 suggests that the appropriate instrument is a penalty (or feedback signal) on the first moment of the within-course grade distribution rather than a constraint on its shape. A mean-targeting rule preserves the within-course variation in G_{ic} that carries information about relative student ability, while still creating incentives to restrain upward drift. Even a simple institutional practice of flagging courses whose mean grade exceeds a threshold—without mandating that only a fixed percentage receive each letter grade—is typically more informative than a distributional cap, because it preserves within-course monotonicity and does not mechanically impose cohort-specific percentile cutoffs. If enforced strongly, such a threshold can still induce instructors to shift effective standards across stronger and weaker cohorts, which is why Section 5.3 later advocates ability-adjusted mean targets rather than

a naive common benchmark. Complementarily, the "rational employer" observation from Section 4.1—that $w(g, c) = \mathbb{E}[a \mid g, c]$ can refine $w(g) = \mathbb{E}[a \mid g]$ whenever course identity carries information about difficulty—suggests a low-cost reform: reporting course-level contextual information (such as median grade, historical course average, enrolment size, and course level) alongside individual grades on transcripts. This provides outside observers with the raw material to condition on course identity when forming ability inferences, without requiring the institution to implement a formal difficulty-adjustment algorithm.

Curricular overlap enables identification. The identification results in Section 4.2 establish that a connected enrolment graph is both necessary and sufficient for recovering abilities and difficulties up to a location normalisation (Proposition 5). This implies that breadth requirements, distribution requirements, and shared core courses may serve a previously underappreciated informational function beyond their pedagogical rationale. By ensuring that students from different concentrations, tracks, or ability levels share at least some courses in common, these requirements create the cross-course bridges through which relative difficulty can be estimated—whether by a formal eigengrade system, a simpler two-way fixed-effects adjustment, or even informal employer inference from transcript patterns. In the multi-dimensional extension, those same shared core courses can also serve as a natural reference bundle for a benchmark-specific scalar summary: ranking students by expected performance on the common core is meaningful even when the full skill profile admits no universal total order. Conversely, curricula that allow students to take entirely non-overlapping course portfolios (for example, highly siloed degree programmes with no shared requirements) sever the links needed for any form of cross-course calibration. In such settings, the enrolment graph is disconnected, and Proposition 5 implies that difficulty differences across the disconnected components are fundamentally unidentified from transcript data alone.

Preserve within-course grade variation. The eigengrade estimator, and more generally transcript methods that identify students through within-course deviations, relies on within-course variation in grades to distinguish students of different abilities. Policies that compress this variation—aggressive distributional caps, narrow forced curves, or expansive pass/fail options—reduce the signal-to-noise ratio of each transcript entry. In the limit, a course in which every student receives the same grade contributes no within-course ranking information; for eigengrades it drops out after centring, although in additive fixed-effects formulations it can still matter indirectly by helping connect components of the enrolment graph. More subtly, top-coding (when a large fraction of students receive the highest available grade) eliminates

the ability to distinguish among high-performing students within a course, precisely the margin that is most valuable for competitive selection decisions. The formal counterpart is the grade-quantisation effect visible in the illustrative example of Section 4.2.4: even with perfectly estimated difficulties, the coarse three-level grade scale produces residual ranking errors that would diminish with finer resolution. Institutions that wish to maximise the informational content of their transcripts should therefore resist policies that mechanically compress grade distributions and, where possible, adopt grading scales with sufficient resolution to preserve meaningful within-course differentiation.

Endogenous course selection can be as damaging to the grade informativeness as inflation. Finally, Proposition 2 observes that endogenous sorting of students into courses can make grades not merely uninformative but actively misleading about ability—even on average and even in the absence of any grade inflation. When higher-ability students systematically select harder courses, the resulting selection on difficulty can reverse the expected relationship between grades and ability, producing a setting in which lower grades are associated with higher posterior expected ability.

And of course, we generally would not want to restrict or distort students' course choices simply for the sake of making grading more informative. Rather, once again, either directly providing information about course difficulty or somehow normalizing grades to account for difficulty endogenously is key to solving this problem.

8 Conclusion

This paper has formalised the identification problem at the heart of academic grading. A letter grade collapses two dimensions—student ability and course difficulty—into a single scalar signal, and that compression has precise consequences. Our impossibility theorem shows that on curricula rich enough to realise easy-course/hard-course reversals, no universally calibrated non-trivial threshold grading rule can guarantee ability-sufficient cross-course comparisons from a single grade. The point is not that every curriculum generates such a reversal, but that once ability and difficulty both matter, a single calibrated scalar rule cannot rule reversals out uniformly. The argument requires no parametric distributional assumptions.

We have identified three complementary responses operating along different margins of the problem. First, replacing pointwise ability sufficiency with a statistical version—higher grades imply higher expected ability—yields a condition that holds under exogenous course assignment but fails sharply under endogenous selection. Second, moving from single grades to multi-course transcripts enriches the information set enough to restore identification in

an additive latent-performance model. A connected enrolment graph identifies abilities and difficulties up to a location normalisation; in the complete-data noiseless benchmark eigengrades are simply a spectral re-expression of centred ability. For noisy or incomplete latent-score data we provide perturbation bounds for an inverse-probability-weighted spectral estimator, and for ordinal letter grades we show how a consistent ordered-response first stage can be carried into the eigengrade direction. Third, modelling grade inflation as a Nash equilibrium of a game among self-interested instructors reveals a competitive externality that blunt grade caps cannot address; a Taylor-rule feedback mechanism that penalises mean grades above a target can stabilise grade levels while preserving one key precondition for informative transcripts, namely within-course variation. The combination of eigengrades for external reporting and a Taylor rule for internal governance therefore separates the identification problem from a first-moment inflation-control problem, rather than claiming to solve every margin on which grading can lose information.

More broadly, the paper illustrates a general principle in information design: when the signal space is lower-dimensional than the state space, identification fails, and the resolution lies in either enriching the signal or exploiting structure. This principle applies beyond grading to any setting where observers must infer multi-dimensional quality from compressed signals, including product ratings, performance reviews, and competitive rankings.

Several limitations suggest directions for future work. The additive performance model $\pi(a_i, d_c) = a_i - d_c + \varepsilon_{ic}$ rules out comparative advantage: a student who excels at mathematics but struggles with poetry cannot be captured by a scalar ability parameter. We have sketched a multi-dimensional extension (Section 6.2) in which the spectral step generalises from a single eigengrade to recovery of a rank-K interaction subspace. That extension should be read carefully: it separates scalar baseline ability from comparative advantage, but it does not by itself deliver a canonical universal ranking of students. A natural next step is therefore to study institutionally meaningful scalarisations—for example, benchmark scores defined over shared core courses or other reference course bundles—together with the conditions under which such benchmarks should be chosen.

Much of the spectral analysis is most transparent at the latent-score level. When only ordinal letter grades are observed, the rigorous route is two-step: estimate the latent ordered-response index first, then apply the eigengrade map to the fitted latent matrix. Characterising primitive conditions under which particular ordered-response estimators achieve the required first-stage rate in sparse two-way panels is a useful next step.

The baseline grading game emphasises the symmetric benchmark, but Section C already shows that heterogeneous primitives $(\alpha_c, \gamma_c, \bar{G}_{0,c})$ and predetermined course-specific targets preserve the exact-potential structure and uniqueness of equilibrium, albeit with generally

asymmetric outcomes. The open problem is therefore not existence or uniqueness under heterogeneity; it is the design of *optimal* heterogeneous targets, penalties, and welfare weights when departments sustain systematically different grading norms.

The eigengrade estimator is data-hungry: students with thin transcripts or courses with few enrolments yield imprecise estimates. Practical implementation would benefit from shrinkage or Bayesian regularisation, in the spirit of the Glicko rating system (Glickman, 1999), to handle sparse observations gracefully.

Finally, our Taylor-rule analysis is static, but grade inflation unfolds dynamically as instructors respond to each other's grading choices over time. A natural extension would model this as an explicit dynamic game in which instructors adjust difficulty in response to lagged university-wide grades—in direct analogy to the New Keynesian Phillips curve (Calvo, 1983)—opening a connection to the literature on learning and convergence in repeated games with feedback.

A Appendix: Formal Connections to Matching Markets

This appendix provides the formal matching market framework and proofs supporting the discussion in Section 6.1. We embed our grading model into the matching market of Ostrovsky and Schwarz (2010) (hereafter OS).

A.1 The OS Framework

We briefly recapitulate the elements of OS that are needed for our purposes.

There is a continuum of students with true abilities $a \in [a_L, a_H]$ distributed across I schools. The distribution of ability at school i is $\lambda_i(a)$; the aggregate distribution across all schools is $\lambda(a) = \sum_i \lambda_i(a)$. Each school perfectly observes the abilities of its own students and chooses a *transcript structure*, a stochastic mapping from ability to transcripts, to maximise the total desirability of placements obtained by its graduates.

On the other side of the market there is a continuum of positions with desirabilities $q \in [q_L, q_H]$ distributed according to a density $\mu(q)$, with the total mass of positions equal to the total mass of students. Employers observe only transcripts, compute the expected ability $\hat{a}$ of each student conditional on the transcript and the publicly known transcript structure, and rank students by $\hat{a}$. Students rank positions by desirability. Because preferences are aligned on both sides, the resulting stable matching is assortative: the student of rank r (by expected ability) is matched to the position of rank r (by desirability). Write $Q(\hat{a})$ for the *desirability mapping*: the average desirability of the interval of positions assigned to students with posterior mean ability $\hat{a}$. When the distribution of $\hat{a}$ is atomless this coincides with the unique matched desirability at $\hat{a}$; under pooling atoms it should be read as the expected placement value under the equilibrium tie-breaking within that interval.

Two results from OS are central to what follows:

OS Theorem 1 (Uniqueness). In any connected equilibrium the desirability mapping $Q(\hat{a})$ is the same. The equilibrium amount of information disclosure is uniquely determined by the aggregate ability distribution $\lambda(a)$ and the position-desirability distribution $\mu(q)$; it does not depend on how students are distributed across schools.

OS Lemma 2 (Convexity). In any connected equilibrium the desirability mapping $Q(\hat{a})$ is convex on its domain $[\hat{a}_L, \hat{a}_H]$.

The convexity of Q is an equilibrium property: if Q were concave on some interval, a school producing students with expected abilities in that interval could raise its total placement desirability by pooling (mixing students together to create a single expected-ability category),

contradicting equilibrium. Conversely, strict convexity at a point implies that the equilibrium is "fair" there: students are fully separated and matched to positions on the basis of true ability.

A.2 The Impossibility Theorem in the Matching Market

We now embed our grading model inside the OS market. Suppose that the "performance" observed within a course is generated by the function $\pi(a_i, d_c)$ of Section 2, that the grading rule g satisfies within-course monotonicity (Axiom 2), universal calibration (Axiom 3), and non-triviality (Axiom 4), and that the resulting grades are the transcripts on which employers condition.

OS transcript structures are school-specific: each school i independently chooses a stochastic mapping $f_i(t \mid a)$ from ability to transcripts. By contrast, universal calibration (Axiom 3) requires that the mapping from performance to grades is the same across all courses, and hence across all schools.

Proposition 10 (Universal calibration forecloses robust OS implementation)**.** *Fix $\lambda(a)$ and $\mu(q)$ such that the connected OS equilibrium is not fully informative. Fix any grading rule g satisfying Axioms 1–4, and let $G = g(\pi(a, d))$ denote the reported grade when ability is a and course difficulty is d. Let $\hat{a}(G) = E[a \mid G]$ denote the employer's posterior mean ability given the grade, and let F_g denote the distribution of $\hat{a}(G)$ in the population.*

If course difficulty is heterogeneous and not separately observed by employers, then F_g depends on the joint distribution of (a, d), and hence on the curriculum. In particular, there exist curricula with heterogeneous difficulty (and associated course-taking patterns that preserve the same aggregate $\lambda(a)$) for which F_g differs from the OS equilibrium distribution of posterior means. Consequently, no universally calibrated grading rule can implement the OS equilibrium in a manner that is robust to heterogeneous course difficulty.

Proof. By OS Theorem 1, the connected-equilibrium distribution of posterior means is pinned down by (λ, μ) and does not depend on curricular details such as the distribution of course difficulties.

Under universal calibration, employers observe only $G = g(\pi(a, d))$ and therefore form posteriors by conditioning on an event that depends jointly on ability and difficulty. Unless course difficulty is degenerate or separately observed and conditioned on, the conditional distribution $a \mid G$ (and hence $\hat{a}(G) = E[a \mid G]$) depends on the distribution of d and on how students are allocated across difficulties.

To see that this dependence is substantive, compare two curricula with the same aggregate ability distribution $\lambda(a)$. In the first curriculum, difficulty is degenerate at some d_0, so

$G = g(\pi(a, d_0))$ is a coarsening of ability alone. In the second curriculum, difficulty takes at least two distinct values with positive probability, with course assignment independent of a. Because g is non-trivial (Axiom 4) and π satisfies the weak overlap property (Axiom 1), the set of ability levels mapped into a given grade bin differs across the two difficulty levels, so the mixture of abilities within at least one grade bin changes. This changes at least one posterior mean $\hat{a}(g_k)$ and therefore changes the distribution F_g.

Since the OS equilibrium distribution is fixed by (λ, μ), at most one of these curricula can yield F_g equal to the OS equilibrium distribution. Hence a universally calibrated grading rule cannot implement the OS equilibrium distribution in a way that is robust to heterogeneous course difficulty. □

A.3 Convexity, Risk Neutrality, and Statistical Sufficiency

In the OS framework, as new information arrives, a student's expected ability $\hat{a}_\tau$ follows a martingale: $E[\hat{a}_\tau \mid \hat{a}_{\tau'}] = \hat{a}_{\tau'}$ for $\tau > \tau'$. Because the equilibrium desirability mapping $Q(\hat{a})$ is convex, Jensen's inequality implies

$$E\Big[Q(\hat{a}_\tau) \mid \hat{a}_{\tau'}\Big] \geq Q\left(E[\hat{a}_\tau \mid \hat{a}_{\tau'}]\right) = Q(\hat{a}_{\tau'}).$$

This is not a monotone-inference analogue of Proposition 1. Rather, it is a within-school value-of-information statement: moving from $\hat{a}_{\tau'}$ to the mean-preserving refinement $\hat{a}_\tau$ raises expected placement desirability whenever Q is convex. The monotone-inference comparison within the school follows separately from assortative matching and the weak monotonicity of Q, as formalised in Proposition 11.

Proposition 11 (Within-school monotone inference under the OS equilibrium)**.** *Suppose school i chooses its transcript structure to form a connected OS equilibrium, and consider two students at school i with posterior mean abilities $\hat{a}_H > \hat{a}_L$ as implied by their transcripts. Then the desirability of the position obtained by the $\hat{a}_H$ student is weakly higher than the desirability obtained by the $\hat{a}_L$ student. Moreover, if Q is strictly convex on $[\hat{a}_L, \hat{a}_H]$, then the position desirability is strictly higher.*

Proof. In an assortative matching, higher posterior mean ability corresponds to weakly higher rank and hence weakly higher matched desirability. Equivalently, the desirability mapping Q is weakly increasing, so $\hat{a}_H > \hat{a}_L \implies Q(\hat{a}_H) \geq Q(\hat{a}_L)$.

If Q is strictly convex on $[\hat{a}_L, \hat{a}_H]$, then Q cannot be constant on that interval. Combined with weak monotonicity, this yields $Q(\hat{a}_H) > Q(\hat{a}_L)$. □

Endogenous course choice and observable course identity create an additional disclosure channel beyond the school's transcript design. The following result records this refinement formally in terms of posterior means.

Proposition 12 (Course observability refines the grade signal)**.** *Suppose employers observe both a reported grade G and course difficulty d_c (or any publicly observed course attribute that may further refine beliefs about a). Let*

$$\hat{a} := E[a \mid G] \qquad \textit{and} \qquad \hat{a}^+ := E[a \mid G, d_c]$$

denote the posterior mean ability under grades alone and under grades plus course difficulty, respectively. Then $\hat{a}^+$ is a mean-preserving refinement of $\hat{a}$ in the sense that

$$E[\hat{a}^+ \mid G] = \hat{a}.$$

If there exists a grade level g_k such that $E[a \mid G = g_k, d_c]$ is not almost surely constant as a function of d_c, this refinement is strict in posterior-mean terms:

$$\mathrm{Var}(\hat{a}^+ \mid G = g_k) > 0.$$

Consequently, for any convex function φ,

$$E\Big[\varphi(\hat{a}^+) \mid G\Big] \geq \varphi(\hat{a}),$$

with strict inequality whenever φ is strictly convex on the conditional support of $\hat{a}^+ \mid G$ and the refinement is strict.

Proof. The identity $E[\hat{a}^+ \mid G] = \hat{a}$ is the law of iterated expectations:

$$E[\hat{a}^+ \mid G] = E\Big[E[a \mid G, d_c] \mid G\Big] = E[a \mid G] = \hat{a}.$$

If there exists a grade level g_k such that $E[a \mid G = g_k, d_c]$ is not almost surely constant in d_c, then $\hat{a}^+ = E[a \mid G, d_c]$ is non-degenerate conditional on $G = g_k$, which implies $\mathrm{Var}(\hat{a}^+ \mid G = g_k) > 0$.

The convex inequality is Jensen's inequality applied conditional on G:

$$E\Big[\varphi(\hat{a}^+) \mid G\Big] \geq \varphi\Big(E[\hat{a}^+ \mid G]\Big) = \varphi(\hat{a}),$$

with strict inequality under strict convexity and non-degeneracy. □

A.4 Eigengrades Under Strategic Information Compression

The next proposition shows that a reduced-form additive-noise model of transcript compression preserves eigengrade consistency under the same asymptotic logic as our baseline setting.

Proposition 13 (Eigengrade consistency under additive-noise transcript compression)**.** *Suppose each school implements transcript compression by adding noise* after *the performance stage: latent performance is $P_{ic} = a_i - d_c$, and the reported numerical grade is*

$$\widetilde{G}_{ic} = P_{ic} + \eta_{ic},$$

This additive-noise specification is a stylised reduced-form version of OS-style compression; canonical OS equilibria are often deterministic pooling/partition rules and need not generate independent, mean-zero errors. Here (η_{ic}) are independent across (i, c) with $E[\eta_{ic} \mid a_i] = 0$ and uniformly sub-Gaussian with parameter at most $\bar{\sigma}$. Assume the observation pattern and boundedness conditions of Theorem 3, so that the eigengrade estimator is well-defined and the observation graph is connected with high probability.

Then the eigengrade direction computed from $\widetilde{G}$ is consistent under the same rate conditions as Theorem 3, up to replacing σ by $\bar{\sigma}$ in the bound.

Proof. Under the stated assumptions, $\widetilde{G}_{ic} = a_i - d_c + \eta_{ic}$ is exactly the additive latent-index model of Theorem 3 with noise term $\varepsilon_{ic} = \eta_{ic}$ and sub-Gaussian scale $\bar{\sigma}$. The proof of Theorem 3 applies verbatim, with the same centring algebra $MP_0 = (Ma)1_m^\top$ and the same eigenvector perturbation argument, yielding the stated conclusion. □

A.5 Hierarchical Incentives and the Taylor Rule

The next proposition shows that a common Taylor-rule target can implement the OS benchmark at the level of average grades while leaving higher-moment transcript structure unconstrained.

Proposition 14 (The Taylor rule as partial implementation of the OS equilibrium)**.** *Suppose the university uses a* common *Taylor-rule target $\bar{G}^\star$, sets that target equal to the mean grade implied by the OS equilibrium transcript structure, and chooses the penalty parameter λ so that the condition in Proposition 8 holds. Then the symmetric Nash equilibrium of the penalised instructor-level grading game has mean grade $\bar{G}^\star$.*

This is only a first-moment implementation result. The Taylor rule disciplines the course mean, but it does not in general implement the full OS transcript distribution, whose shape may depend on higher moments and on the exact pooling structure.

Proof. The conclusion on the mean follows directly from Proposition 8: under its stability condition, the symmetric equilibrium of the penalised grading game satisfies $\bar{G}_c = \bar{G}^\star$ for every course c. Setting $\bar{G}^\star$ equal to the OS-equilibrium mean therefore aligns the first moment.

For the second statement, note that the OS equilibrium is generally described by a distribution of transcript categories or posterior beliefs, not just by its mean. A rule that fixes only the within-course mean

$$\bar{G}_c = \frac{1}{n_c} \sum_{i \in c} G_{ic}$$

leaves the variance, tail behaviour, and pooling pattern of grades within the course largely unrestricted. Different within-course grade distributions can therefore share the same mean $\bar{G}^\star$ while inducing different posterior structures in the matching market. Hence the Taylor rule aligns the OS-equilibrium mean but does not, in general, implement the full OS transcript structure. □

B Formal Results for the Multi-Dimensional Additive Interaction (MDAI) Model

This appendix provides the formal definitions, identification conditions, and proofs for the Multi-Dimensional Additive Interaction (MDAI) model discussed conceptually in Section 6.2. It also clarifies a key limitation: absent additional structure, the interaction term identifies a skill subspace, not a canonical total ordering of students.

B.1 Model Definition and Matrix Formulation

Definition 7 (MDAI Model)**.** Let $K \geq 1$. Each student i has a baseline ability $a_i \in \mathbb{R}$ and a skill profile $u_i \in \mathbb{R}^K$. Each course c has a baseline difficulty $d_c \in \mathbb{R}$ and a skill-demand profile $v_c \in \mathbb{R}^K$. Latent performance is

$$P_{ic} = a_i - d_c + u_i^\top v_c + \varepsilon_{ic}, \tag{13}$$

where (ε_{ic}) are independent, mean-zero idiosyncratic shocks.

As in the purely additive model, (a, d) are identified only up to a common additive constant: replacing (a, d) by $(a + \kappa 1_n,\ d + \kappa 1_m)$ leaves (P_{ic}) unchanged. All results below are invariant to this shift.

To separate the additive main effects (a_i, d_c) from the interaction term (u_i, v_c), we impose

the standard mean-zero normalisations

$$\sum_{i=1}^{n} u_i = 0 \qquad \text{and} \qquad \sum_{c=1}^{m} v_c = 0,$$

so that the interaction matrix has zero row and column means.

Let $U \in \mathbb{R}^{n\times K}$ be the matrix with rows $u_i^\top$ and $V \in \mathbb{R}^{m\times K}$ the matrix with rows $v_c^\top$. Writing $a \in \mathbb{R}^n$ and $d \in \mathbb{R}^m$ for the stacked vectors of baselines, the noiseless MDAI performance matrix can be written

$$P^0 = a1_m^\top - 1_n d^\top + UV^\top.$$

The term $UV^\top$ is low rank (rank at most K) and captures multi-dimensional match effects.

A key feature of the MDAI model is a built-in change-of-basis indeterminacy. For any invertible matrix $R \in \mathrm{GL}(K, \mathbb{R})$, the pair (U, V) and the pair $(UR, VR^{-\top})$ generate the same interaction matrix $UV^\top$. Without additional structure (such as the anchor courses or pre-specified benchmark bundles discussed in Section 6.2), the data can recover the interaction subspaces and the product $UV^\top$, but not a uniquely labelled coordinate system for skills or a canonical scalar comparison.

Remark 17 (No canonical universal ranking in the interaction term). The multi-dimensional interaction component does not, by itself, induce a basis-free total ordering of students. A natural basis-free comparison relation is the coursewise dominance preorder

$$i \succeq j \quad \Longleftrightarrow \quad u_i^\top v_c \geq u_j^\top v_c \text{ for every course } c,$$

which is generally incomplete: two students may each outperform the other in different parts of the curriculum. Accordingly, the identified interaction object is the matrix $UV^\top$ (or equivalently its row and column spaces), not a canonical scalar ranking. Any one-dimensional ranking in the MDAI model therefore requires additional structure, such as ranking only the baseline component a_i or fixing a reference bundle of courses as in Definition 9.

B.2 Complete Data: Subspace Recovery

Let $M_n := I_n - \frac{1}{n}1_n 1_n^\top$ and $M_m := I_m - \frac{1}{m}1_m 1_m^\top$ be the student- and course-centring projections. To avoid notational collision with $\tilde{P}$ in Section 4.2.3, define the *double-centred* matrix $P^{\mathrm{dc}} := M_n P M_m$.

Definition 8 (Multi-Dimensional Eigengrades)**.** The *Multi-Dimensional Eigengrades* (MDE) are the matrix $\widehat{U}_K \in \mathbb{R}^{n\times K}$ whose columns are the top K left singular vectors of P^{dc}.

Theorem 4 (Subspace Recovery with Complete Data)**.** *In the noiseless MDAI model with complete data, suppose U and V have full column rank K. Then P^{dc} has rank K, and the column space of $\widehat{U}_K$ equals the column space of U.*

Moreover, the interaction factorisation is not unique: if (U, V) is feasible, then $(UR, VR^{-\top})$ is feasible for any $R \in \mathrm{GL}(K, \mathbb{R})$. If one additionally fixes an orthonormal factorisation (for example, by taking the singular vectors of $UV^{\top}$), the remaining indeterminacy reduces to an orthogonal change of basis within any singular subspace with repeated singular values.

Proof. Apply double-centring to the noiseless matrix:

$$M_n P^0 M_m = M_n(a1_m^{\top})M_m - M_n(1_n d^{\top})M_m + M_n(UV^{\top})M_m.$$

Because $M_n 1_n = 0$ and $1_m^{\top} M_m = 0^{\top}$, the additive main-effect terms vanish. By the mean-zero normalisations, $M_n U = U$ and $V^{\top} M_m = V^{\top}$. Hence

$$M_n P^0 M_m = UV^{\top}.$$

Thus $P^{\mathrm{dc}} = UV^{\top}$ in the noiseless complete-data case, so $\mathrm{rank}(P^{\mathrm{dc}}) = K$ when U and V have full column rank.

The left singular space of $UV^{\top}$ is the column space of U, so the top K left singular vectors span $\mathrm{col}(U)$. The change-of-basis indeterminacy follows from the identity

$$(UR)(VR^{-\top})^{\top} = URR^{-1}V^{\top} = UV^{\top}$$

for any invertible R. □

Definition 9 (Reference-bundle score)**.** Fix nonnegative weights $\omega_1, \ldots, \omega_m$ summing to 1, interpreted as a reference distribution over courses. The associated *reference-bundle score* for student i is

$$s_i(\omega) := a_i + \sum_{c=1}^{m} \omega_c \, u_i^{\top} v_c.$$

Equivalently, in the noiseless model,

$$s_i(\omega) = \sum_{c=1}^{m} \omega_c \, (P_{ic} + d_c),$$

so $s_i(\omega)$ is expected latent performance on the reference bundle after removing the bundle's common average difficulty. If the support of ω is a shared core curriculum, we call $s_i(\omega)$ a *core score*.

Proposition 15 (Basis invariance of benchmark-specific scalarisations)**.** *Fix weights ω. The reference-bundle score $s_i(\omega)$ is invariant to the change of basis $(U, V) \mapsto (UR, VR^{-\top})$ for any invertible $R \in \mathrm{GL}(K, \mathbb{R})$. Hence, once the baseline vector a and the interaction matrix $UV^\top$ are identified up to the usual additive normalisation in a, the ordering of students by $s_i(\omega)$ is well defined up to that same common additive shift. In particular, a shared core curriculum yields a legitimate domain-specific scalar comparison even though the full interaction profile does not generate a canonical universal ranking.*

Proof. Under the change of basis $(U, V) \mapsto (UR, VR^{-\top})$,

$$(UR)_i^\top (VR^{-\top})^\top \omega = u_i^\top R R^{-1} V^\top \omega = u_i^\top V^\top \omega = \sum_{c=1}^{m} \omega_c \, u_i^\top v_c.$$

So the interaction component of $s_i(\omega)$ is basis-invariant. The additive normalisation $a \mapsto a + \kappa 1_n$ shifts every $s_i(\omega)$ by the same constant κ, leaving all pairwise rankings unchanged. □

B.3 Incomplete Transcripts: Beyond Connectedness

With missing transcript entries, identification of $UV^\top$ becomes a low-rank matrix completion problem. Suppose the main effects (a, d) have been fixed, so the observed residuals satisfy $r_{ic} := P_{ic} - a_i + d_c = u_i^\top v_c$ for observed pairs (i, c). Let $\mathcal{G} \subseteq S \times C$ denote the bipartite observation graph.

Proposition 16 (A necessary degree condition)**.** *Fix any representation of the course profiles v_c. If a student i is observed in fewer than K courses, then u_i is not uniquely determined from the observed inner products. Symmetrically, if a course c is observed for fewer than K students, then v_c is not uniquely determined from the observed inner products.*

Consequently, a necessary condition for unique identification of interaction profiles (up to a single global change of basis) is that every student and every course has degree at least K in $\mathcal{G}$.

Proof. Fix a student i observed in k courses with indices in C_i. The observations determine the linear system

$$u_i^\top v_c = r_{ic} \qquad \text{for } c \in C_i,$$

where $r_{ic} := P_{ic} - a_i + d_c$. This is k linear equations in the K unknown coordinates of u_i. If $k < K$, the system is underdetermined for any choice of v_{C_i}, so there are infinitely many solutions. The argument for a course c is symmetric. □

Theorem 5 (A sufficient condition for sequential recovery)**.** *Assume the main effects* $(\mathbf{a}, \mathbf{d})$ *have been fixed (e.g., estimated in a first stage), and consider the residuals* $R_{ic} := P_{ic} - a_i + d_c$*, observed on* $\mathcal{G}$*; and assume the noiseless low-rank relation* $R_{ic} = u_i^\top v_c$ *holds on* $\mathcal{G}$*.*

Suppose there exist sets of students S_0 *and courses* C_0 *with* $|S_0| = |C_0| = K$ *such that all pairs in* $S_0 \times C_0$ *are observed and the corresponding residual submatrix has rank* K*.*

Suppose further that there is an ordering of the remaining vertices such that each new student is observed in at least K *previously recovered courses whose* v_c *rows span* $\mathbb{R}^K$*, and each new course is observed for at least* K *previously recovered students whose* u_i *rows span* $\mathbb{R}^K$*.*

Then the interaction profiles (U, V) *are uniquely determined up to a single global change of basis* $R \in \mathrm{GL}(K, \mathbb{R})$*.*

Proof. On the seed submatrix $S_0 \times C_0$, the rank-K condition implies the residual matrix admits a rank-K factorisation $U_{S_0} V_{C_0}^\top$ with both factors full column rank, which pins down the corresponding column spaces. Any two such factorisations differ by a change of basis $R \in \mathrm{GL}(K, \mathbb{R})$.

Proceed inductively. Suppose a set of course vectors has been recovered in a common coordinate system. For a new student i observed in courses C_i within the recovered set, the equations $u_i^\top v_c = R_{ic}$ for $c \in C_i$ determine u_i uniquely whenever the v_c for $c \in C_i$ span $\mathbb{R}^K$. The symmetric argument applies when adding a new course. The stated spanning assumptions ensure uniqueness at each step, and all recovered profiles remain tied to the same initial coordinate system, so the only remaining freedom is the initial change of basis. □

B.4 Noisy and Incomplete Multi-Dimensional Eigengrades

This section records a direct extension of the spectral perturbation argument to the MDAI setting, focusing on recovery of the interaction subspace under MCAR missingness.

Theorem 6 (Consistency of MDE under MCAR for Latent Scores)**.** *Assume the MDAI model*

$$P_{ic} = a_i - d_c + u_i^\top v_c + \varepsilon_{ic},$$

where (ε_{ic}) *are independent, mean-zero* σ*-sub-Gaussian variables. Let* $\Omega_{ic} \sim \mathrm{Bernoulli}(p)$ *be independent observation indicators (MCAR), and observe* $Y_{ic} = \Omega_{ic} P_{ic}$*. Form the inverse-probability-weighted matrix* $\widehat{P}$ *with entries* $\widehat{P}_{ic} := Y_{ic}/p$*.*

Define the double-centred matrix $\widehat{P}^{\mathrm{dc}} := M_n \widehat{P} M_m$*, and let* $\widehat{U}_K \in \mathbb{R}^{n \times K}$ *be its top* K *left singular vectors. Let* $U_K \in \mathbb{R}^{n \times K}$ *be any orthonormal basis for* $\mathrm{col}(U)$*, and let* $S := UV^\top$ *with* K*th singular value* $\sigma_K(S) > 0$*.*

Assume bounded parameters $|a_i| \le A$, $|d_c| \le D$, and $|u_i^\top v_c| \le B$. Then there exist universal constants $C, c > 0$ such that, with probability at least $1 - 2\exp(-c(n+m))$,

$$\left\|\sin\Theta(\widehat{U}_K, U_K)\right\|_{\mathrm{op}} \le C\,\frac{A+D+B+\sigma}{p}\cdot\frac{\sqrt{n}+\sqrt{m}}{\sigma_K(S)}.$$

Proof. Write $\widehat{P} = P^0 + W$, where P^0 is the noiseless MDAI matrix and

$$W_{ic} = \left(\frac{\Omega_{ic}}{p} - 1\right) P^0_{ic} + \frac{\Omega_{ic}}{p}\,\varepsilon_{ic}.$$

Then $E[W_{ic}] = 0$ and the entries of W are independent. Using $|P^0_{ic}| \le A + D + B$ and sub-Gaussianity of ε_{ic}, each W_{ic} is mean-zero sub-Gaussian with scale on the order of $(A + D + B + \sigma)/p$.

Double-centring gives

$$\widehat{P}^{\mathrm{dc}} = M_n P^0 M_m + M_n W M_m.$$

By the algebra in the complete-data result and the mean-zero normalisations in Definition 7 (so that $M_n U = U$ and $V^\top M_m = V^\top$), we have $M_n P^0 M_m = UV^\top = S$, so $\widehat{P}^{\mathrm{dc}} = S + \widetilde{W}$ where $\widetilde{W} := M_n W M_m$.

Standard operator-norm bounds for rectangular sub-Gaussian matrices imply that, with probability at least $1 - 2\exp(-c(n+m))$,

$$\|W\|_{\mathrm{op}} \le C'\,\frac{A+D+B+\sigma}{p}\,(\sqrt{n}+\sqrt{m})$$

for a universal constant C'. Since M_n and M_m are orthogonal projections, $\|\widetilde{W}\|_{\mathrm{op}} \le \|W\|_{\mathrm{op}}$.

Applying a singular-subspace perturbation bound (Wedin's $\sin\Theta$ theorem) to the rank-K signal matrix S perturbed by $\widetilde{W}$ yields

$$\left\|\sin\Theta(\widehat{U}_K, U_K)\right\|_{\mathrm{op}} \le \frac{\|\widetilde{W}\|_{\mathrm{op}}}{\sigma_K(S)}.$$

Combining inequalities and absorbing constants gives the stated bound. □

C Heterogeneous Grading Norms and Asymmetric Equilibrium

The symmetric benchmark in Proposition 7 is analytically convenient, but it is not essential for the potential-game structure. This appendix allows courses to differ in their direct benefit

from higher grades, in their professional grading norms, and in the marginal cost of deviating from those norms, while maintaining a common competitive environment.

Specifically, let the logit share function remain

$$s_c(\bar{\mathbf{G}}) = \frac{\exp(\eta \bar{G}_c)}{\sum_{k=1}^{m} \exp(\eta \bar{G}_k)}, \qquad \eta > 0,$$

with common β, η, but allow course-specific primitives

$$\alpha_c \geq 0, \qquad \gamma_c > 0, \qquad \bar{G}_{0,c} \in \mathbb{R}.$$

Instructor c's payoff is

$$U_c^H(\bar{G}_c, \bar{\mathbf{G}}_{-c}) = \alpha_c \bar{G}_c + \beta \log s_c(\bar{\mathbf{G}}) - \frac{\gamma_c}{2}(\bar{G}_c - \bar{G}_{0,c})^2, \qquad \bar{G}_c \in [\underline{G}, \overline{G}]. \tag{14}$$

Proposition 17 (Heterogeneous grading game)**.** *Consider the game* (14)*.*

1. *The game is an exact potential game with potential*

$$\Phi^H(\bar{\mathbf{G}}) := \sum_{c=1}^{m} \left[(\alpha_c + \beta\eta)\bar{G}_c - \frac{\gamma_c}{2}(\bar{G}_c - \bar{G}_{0,c})^2 \right] - \beta \log \left(\sum_{k=1}^{m} e^{\eta \bar{G}_k} \right). \tag{15}$$

2. *The game admits a unique Nash equilibrium* $\bar{\mathbf{G}}^{\,\mathrm{NE},H} \in [\underline{G}, \overline{G}]^m$*.*

3. *The equilibrium is characterised as the unique maximiser of* Φ^H*. If it is interior, then it is the unique solution to*

$$\gamma_c\Big(\bar{G}_c^{\,\mathrm{NE},H} - \bar{G}_{0,c}\Big) = \alpha_c + \beta\eta\Big(1 - s_c(\bar{\mathbf{G}}^{\,\mathrm{NE},H})\Big), \qquad c = 1, \ldots, m, \tag{16}$$

or equivalently

$$\bar{G}_c^{\,\mathrm{NE},H} = \bar{G}_{0,c} + \frac{\alpha_c + \beta\eta\Big(1 - s_c(\bar{\mathbf{G}}^{\,\mathrm{NE},H})\Big)}{\gamma_c}, \qquad c = 1, \ldots, m. \tag{17}$$

With bounds, the same equilibrium is characterised by the projection system

$$\bar{G}_c^{\,\mathrm{NE},H} = \Pi_{[\underline{G},\overline{G}]}\left(\bar{G}_{0,c} + \frac{\alpha_c + \beta\eta\Big(1 - s_c(\bar{\mathbf{G}}^{\,\mathrm{NE},H})\Big)}{\gamma_c} \right), \qquad c = 1, \ldots, m. \tag{18}$$

4. *The equilibrium need not be symmetric. In particular, any* interior *symmetric equilibrium*

would have to satisfy

$$\bar{G} = \bar{G}_{0,c} + \frac{\alpha_c + \beta\eta(1 - \frac{1}{m})}{\gamma_c} \qquad \textit{for every } c = 1, \ldots, m. \tag{19}$$

Hence symmetry requires the right-hand side of (19) *to be identical across all courses, which is a knife-edge restriction on the heterogeneous primitives. When*

$$\alpha_c = \alpha, \qquad \gamma_c = \gamma, \qquad \bar{G}_{0,c} = \bar{G}_0 \qquad \textit{for all } c,$$

Proposition 17 collapses to Proposition 7.

Proof. Define Φ^H by (15). Since

$$\log s_c(\bar{\mathbf{G}}) = \eta \bar{G}_c - \log\left(\sum_{k=1}^{m} e^{\eta \bar{G}_k}\right),$$

we have

$$\frac{\partial \Phi^H}{\partial \bar{G}_c} = \alpha_c + \beta\eta\Big(1 - s_c(\bar{\mathbf{G}})\Big) - \gamma_c(\bar{G}_c - \bar{G}_{0,c}) = \frac{\partial U_c^H}{\partial \bar{G}_c}.$$

Thus the game is an exact potential game.

Let $\Gamma := \mathrm{Diag}(\gamma_1, \ldots, \gamma_m)$ and $s := (s_1(\bar{\mathbf{G}}), \ldots, s_m(\bar{\mathbf{G}}))^\top$. The Hessian of Φ^H is

$$\nabla^2 \Phi^H = -\Gamma - \beta\eta^2\Big(\mathrm{Diag}(s) - ss^\top\Big).$$

For any $x \in \mathbb{R}^m$,

$$x^\top \nabla^2 \Phi^H x = -\sum_{c=1}^{m} \gamma_c x_c^2 - \beta\eta^2 \sum_{c=1}^{m} s_c \left(x_c - \sum_{j=1}^{m} s_j x_j\right)^2.$$

Because each $\gamma_c > 0$, this is strictly negative for every nonzero x. Hence Φ^H is strictly concave on the compact convex set $[\underline{G}, \overline{G}]^m$, so it has a unique maximiser.

We now show that Nash equilibria coincide with maximisers of Φ^H. Suppose $\bar{\mathbf{G}}^\dagger$ is a Nash equilibrium. For each c, the function $x_c \mapsto U_c^H(x_c, \bar{\mathbf{G}}_{-c}^\dagger)$ is concave, so

$$\frac{\partial U_c^H}{\partial \bar{G}_c}(\bar{\mathbf{G}}^\dagger)\,(y_c - \bar{G}_c^\dagger) \le 0 \qquad \text{for all } y_c \in [\underline{G}, \overline{G}].$$

Using the exact-potential identity $\partial_{\bar{G}_c} U_c^H = \partial_{\bar{G}_c} \Phi^H$ and summing over c yields

$$\nabla \Phi^H(\bar{\mathbf{G}}^\dagger)^\top (y - \bar{\mathbf{G}}^\dagger) \le 0 \qquad \text{for all } y \in [\underline{G}, \overline{G}]^m.$$

This is the first-order variational inequality characterising a maximiser of the concave function Φ^H. Thus every Nash equilibrium maximises Φ^H. Conversely, if $\bar{\mathbf{G}}^\star$ maximises Φ^H, then no unilateral deviation can increase Φ^H, and by exact-potential equivalence no unilateral deviation can increase any player's payoff; hence $\bar{\mathbf{G}}^\star$ is a Nash equilibrium. Therefore the game has a unique Nash equilibrium, namely the unique maximiser of Φ^H.

If the equilibrium is interior, the first-order condition $\partial_{\bar{G}_c}\Phi^H = 0$ gives (16), which is equivalent to (17). When some coordinates lie on the boundary, the corresponding projection form (18) follows from the usual Kuhn–Tucker conditions.

Finally, if an interior equilibrium were symmetric, say $\bar{G}_c = \bar{G}$ for all c, then $s_c = 1/m$ for every c, and (16) would reduce to (19). Unless the heterogeneous primitives satisfy the resulting equalities across all courses, no common $\bar{G}$ can solve the system. The homogeneous special case immediately reduces to Proposition 7. □

Remark 18 (Implication for policy targets). Proposition 17 shows that the symmetric benchmark in the main text should be read as a tractable baseline, not as a claim that all departments face identical grading incentives or norms. In applications, heterogeneity in $(\alpha_c, \gamma_c, \bar{G}_{0,c})$ makes department-specific raw-grade targets natural. The same exact-potential logic also carries over to the one-sided penalty game: replacing

$$-\lambda(\bar{G}_c - \bar{G}^\star)_+$$

by

$$-\lambda_c(\bar{G}_c - \bar{G}_c^\star)_+$$

preserves exact potential, and the modified potential remains strictly concave because the added penalty term is concave in each coordinate. Accordingly, predetermined course-specific targets preserve uniqueness of equilibrium even though the equilibrium is generally asymmetric.

D Alternative Approaches to Grade Comparability

The identification problem inherent in grading—specifically, the challenge of disentangling student ability from assessment difficulty—is a well-recognised issue in tertiary admissions. Jurisdictions globally have developed various psychometric and statistical frameworks to achieve cross-institutional comparability, ranging from iterative scaling to item response theory. Below, we review these practical approaches, translate them into our formal framework where applicable, and compare their structural properties to the eigengrade method.

D.1 Item Response Theory (IRT)

Item Response Theory (IRT) shifts the analytical focus from aggregate test scores to item-level performance, estimating a student's latent proficiency independently of the specific items encountered TIMSS & PIRLS International Study Center (2020). This framework is heavily utilised in large-scale international assessments to place varying test forms onto a unified scale.

In a Three-Parameter Logistic (3PL) model, the probability of success is modelled non-linearly to account for varying item strengths and random chance. To avoid collision with Section C's grading-game primitives (α_c, γ_c), write the IRT discrimination and guessing parameters as (λ_c, ψ_c). Let $x_{ic} \in \{0, 1\}$ denote a binary outcome (e.g., student i successfully passing an assessment item c). Substituting our parameters for baseline ability (a_i) and course/item difficulty (d_c), the probability function is:

$$P(x_{ic} = 1 \mid a_i; \lambda_c, d_c, \psi_c) = \psi_c + \frac{1 - \psi_c}{1 + e^{-\lambda_c(a_i - d_c)}} \tag{20}$$

Here, λ_c represents the discrimination parameter (the steepness of the response curve), and ψ_c represents the pseudo-guessing parameter (the probability of a correct response through random chance).

Comparison with Eigengrades. IRT provides a rigorous probabilistic framework that mathematically accommodates non-linearities, floor effects (ψ_c), and varying capacities to separate student ability tiers (λ_c). Our additive eigengrade model $(P_{ic} = a_i - d_c + \varepsilon_{ic})$ simplifies these dynamics by assuming parallel item characteristic curves. However, IRT is highly data-intensive, requiring massive item-level datasets and systematically overlapping anchor items to calibrate. Eigengrades operate on macro-level transcript data (course-level aggregates), requiring only a connected bipartite observation graph (Proposition 5). Furthermore, while IRT relies on computationally intensive maximum likelihood estimation to resolve probabilities iteratively, eigengrades offer a deterministic and computationally simpler spectral extraction suitable for sparse, high-level institutional data.

D.2 Iterative Scaling and Cohort Benchmarking

Systems such as the Australian Tertiary Admission Rank (ATAR) scale subject scores by benchmarking against the performance of a subject's cohort in all their other enrolled subjects Universities Admissions Centre (UAC). This process is mathematically analogous to the Bradley–Terry model for paired comparisons Bradley and Terry (1952), relying on iterative convergence algorithms to establish relativity between non-equable subjects.

Comparison with Eigengrades. Both iterative scaling algorithms and eigengrades fundamentally rely on the cross-sectional overlap of student portfolios to establish relative difficulty. At the level of the *overlap-based adjustment itself*, the comparison is between two ways of extracting relative performance information from a connected student–subject graph. Separate from that are implementation and governance choices about what raw inputs enter the system and how often they are re-estimated. In practice, iterative scaling systems often begin from school-assessed coursework marks or other locally generated inputs, so any school-level leniency or socio-economic disparities present in those inputs can be inherited by the subsequent scaling exercise. Here "subject-blind" means that the scaling step does not impose a structural model of subject-specific grading leniency ex ante; it benchmarks subjects through observed cross-subject cohort performance.

By contrast, the formal eigengrade construction isolates relative student and course effects within the maintained additive transcript model and is invariant to common additive course-level shifts once the relevant centring is applied. That invariance is an algorithmic property. Dynamic control of raw-grade inflation is a separate governance problem: the Taylor-rule mechanism in Section 5.2 is best read as a complementary institutional device for stabilising the raw inputs over time, not as part of the eigengrade algorithm itself.

D.3 Longitudinal and Historical Adjustment

Certain institutions lack a unified testing framework and instead apply an empirical adjustment factor to incoming grades University of Waterloo (2026). For example, the University of Waterloo identifies and corrects for grade inflation based on the historical university performance of alumni from the applicant's specific secondary school over a multi-year period.

Comparison with Eigengrades. This longitudinal approach circumvents the need for a contemporaneous connected graph of course enrolments, relying instead on historical anchors. However, it introduces a significant temporal flaw: a current high-ability student is statistically penalised (or rewarded) for the past underperformance (or success) of their school's alumni. Furthermore, the methodology breaks down when evaluating students from newly established schools or regions with sparse historical data. Eigengrades are strictly contemporaneous, calculating difficulty adjustments directly from the current cohort's overlapping performance matrix.

D.4 Quota-Based Standardisation and Comparable Outcomes

Regulators such as the UK's Ofqual utilise a policy of "comparable outcomes," heavily relying on prior attainment data (e.g., Key Stage 2 results) to predict and statistically lock the expected grade distributions for a current cohort Ofqual (2024). If provisional outcomes deviate outside strictly defined tolerances, they must be manually justified.

Comparison with Eigengrades. This approach functions effectively as a demographic quota at the cohort level. As demonstrated in our formal analysis of grade caps (Section 5.1), enforcing rigid distributions destroys individual performance signals and institutionalises rank reversals. The risks of this approach were exposed during the 2020 standardisation crisis, where demographic algorithms actively disadvantaged high-performing students enrolled in historically low-performing schools. The eigengrade method avoids this structural flaw by extracting relative ability signals without dictating the final shape or capacity of the within-course grade distribution.

References

Abowd, John M., Francis Kramarz, and David N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, 1999, *67* (2), 251–333.

Bar, Talia, Vrinda Kadiyali, and Asaf Zussman, "Grade Information and Grade Inflation: The Cornell Experiment," *Journal of Economic Perspectives*, 2009, *23* (3), 93–108.

Blackwell, David, "Equivalent Comparisons of Experiments," *The Annals of Mathematical Statistics*, 1953, *24* (2), 265–272.

Bradley, Ralph Allan and Milton E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, 1952, *39* (3/4), 324–345.

Calvo, Guillermo A., "Staggered Prices in a Utility-Maximizing Framework," *Journal of Monetary Economics*, 1983, *12* (3), 383–398.

Chan, William, Li Hao, and Wing Suen, "Grades, Course Evaluations, and Academic Choice," *Journal of Economic Behavior & Organization*, 2007, *63* (2), 272–286.

Chouldechova, Alexandra, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, 2017, *5* (2), 153–163.

Elo, Arpad E., "The Rating of Chessplayers, Past and Present," *Arco Publishing*, 1978.

Glickman, Mark E., "Parameter Estimation in Large Dynamic Paired Comparison Experiments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1999, *48* (3), 377–394.

_ , "Breaking Chess's Rating Stalemate," *Journal of Data Science and Statistical Modeling*, 2026. Reported in the Harvard Gazette, 6 February 2026.

Heckman, James J., "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *47* (1), 153–161.

Horvitz, D. G. and D. J. Thompson, "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 1952, *47* (260), 663–685.

Johnson, Valen E., "Grade Inflation: A Crisis in College Education," *Springer*, 2003.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in "Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)" 2017.

Little, Roderick J. A. and Donald B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., New York: Wiley, 2002.

Luce, R. Duncan, *Individual Choice Behavior: A Theoretical Analysis*, New York: Wiley, 1959.

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 1993, *60* (3), 531–542.

Milgrom, Paul R., "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 1981, *12* (2), 380–391.

Myerson, Roger B., "Optimal Auction Design," *Mathematics of Operations Research*, 1981, *6* (1), 58–73.

Negahban, Sahand, Sewoong Oh, and Devavrat Shah, "Rank Centrality: Ranking from Pairwise Comparisons," *Operations Research*, 2017, *65* (1), 266–287.

Ofqual, "Inter-subject comparability in GCSEs and A levels in summer 2024," Technical Report 2024.

Ostrovsky, Michael and Michael Schwarz, "Information disclosure and unraveling in matching markets," *American Economic Journal: Microeconomics*, 2010, *2* (2), 34–63.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Stanford InfoLab Technical Report*, 1999.

Pinski, Gabriel and Francis Narin, "Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics," *Information Processing & Management*, 1976, *12* (5), 297–312.

Popov, Sergey V. and Dan Bernhardt, "University Competition, Grading Standards, and Grade Inflation," *Economic Inquiry*, 2013, *51* (3), 1764–1778.

Rojstaczer, Stuart and Christopher Healy, "Where A is Ordinary: The Evolution of American College and University Grading, 1940–2009," *Teachers College Record*, 2012, *114* (7), 1–23.

Rubin, Donald B., "Inference and Missing Data," *Biometrika*, 1976, *63* (3), 581–592.

Sabot, Richard and John Wakeman-Linn, "Grade Inflation and Course Choice," *Journal of Economic Perspectives*, 1991, *5* (1), 159–170.

Spence, Michael, "Job Market Signaling," *Quarterly Journal of Economics*, 1973, *87* (3), 355–374.

Taylor, John B., "Discretion Versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, 1993, *39*, 195–214.

TIMSS & PIRLS International Study Center, "TIMSS 2019 Scaling Methodology: Item Response Theory, Population Models, and Linking Across Modes," Technical Report 2020.

Universities Admissions Centre (UAC), "How your ATAR is calculated," 2026.

University of Waterloo, "Admissions frequently asked questions | Engineering," 2026.