Scilight

*Journal of Modern Applied Statistical Methods*
https://jmasm.com

*Article*

Check for updates

# Measuring Persistent Professor Effects in Ordinal University Grades: A Separation-Robust Alternative to Mixed-Effects Cumulative Logit

Donato Ferrari

Department of Business and Legal Science, University of Calabria, 87036 Arcavacata, Italy; donato.ferrari@unical.it

**How To Cite**: Ferrari, D. Measuring Persistent Professor Effects in Ordinal University Grades: A Separation-Robust Alternative to Mixed-Effects Cumulative Logit. *Journal of Modern Applied Statistical Methods* **2026**, *25*(1), 4. https://doi.org/10.53941/jmasm.2026.100004.

**Abstract:** We study persistent instructor effects in university grading when grades are discrete, ordinal, and pile up at institutional thresholds (18 = legal passing grade; 30 e lode = honors above 30). In such data, the canonical hierarchical specification—a mixed-effects proportional-odds (cumulative logit) model with professor-level random intercepts—is not estimable in practice: quasi-complete separation at the professor level drives the random intercepts toward $\pm\infty$ and prevents the likelihood from attaining a finite maximum. We propose a separation-robust alternative. First, we fit a pooled proportional-odds model for the grade as a function of observed student characteristics (gender, off schedule status, age), exam year, and disciplinary area, explicitly excluding professor identifiers. This model yields, for each exam, the full predicted grade distribution and its expected value on the transcript scale. Second, for each professor we define a grading severity index as the average deviation between realized grades and these predicted benchmarks. This index is always finite, directly interpretable in transcript points, and can be analyzed using standard sampling, shrinkage, and persistence tools. Using over 1.2 million exam records from a large Italian university (2007–2019), we find that grading standards differ sharply across professors: the gap between the severe and generous tails approaches five grade points even after conditioning on observables. These differences are highly persistent over time (year-to-year persistence around 0.8), are not explained by professor gender or by broad disciplinary area, and do not generate systematic subgroup undercoverage in predictive calibration.

**Keywords:** ordinal regression; proportional odds; quasi-complete separation; university grading; instructor effects; value-added

## 1. Introduction

University grades are conventionally treated as measurements of student achievement. In practice, they are also measurements of the person doing the grading. Two students with essentially comparable backgrounds, facing comparable exams in the same academic year and in the same disciplinary area, can receive meaningfully different grades purely because they were evaluated by different professors. This variability is not anecdotal noise. It reflects a professor-specific grading standard that allocates advantages and disadvantages in a systematic way. In the institutional setting we study, individual exam grades enter directly into graduation averages, merit-based financial aid, eligibility for postgraduate programs, and informal hiring filters. A systematic tendency of one professor to grade "low" and another to grade "high", conditional on observables about the student and the exam, is therefore not just stylistic. It is a distributional mechanism for access to opportunities. Statistically, it is a professor effect.

There are two established methodological traditions that speak to this idea. The first is the teacher value-added literature in education and labor economics, which estimates how much a given teacher "moves" student performance relative to what would have been predicted based on that student's prior record and background [1–4]. The main

empirical findings in that literature are that teachers differ persistently, that some teachers are systematically associated with higher outcomes than others, and that these differences are large enough to matter for long-run life outcomes such as earnings and educational attainment [4]. The statistical machinery behind that work is typically linear or hierarchical linear regression for continuous outcomes such as standardized test scores or GPA, with teacher fixed effects or teacher random effects interpreted as persistent teacher quality, teacher "harshness", or teacher value-added.

That machinery does not transfer directly to our setting. The university grading system we analyze is ordinal, truncated, and exhibits extreme pile-ups at salient cutpoints. An individual exam grade does not live on a smooth 0–100 scale. It lies on a discrete, ordered grid: 18, 19, . . . , 30, and 30 e lode. The mark 18 is the first legally valid passing grade. Marks below 18 are not usually recorded numerically in transcript data, because failures are logged as "not passed" rather than as numerical penalties. At the top of the scale, 30 e lode is a special honor category above the regular maximum of 30. The empirical distribution of grades is therefore sharply structured, with visible mass at the institutional thresholds (18, 30, 30 e lode) and relatively thinner mass in the interior. This is structurally different from the approximately Gaussian, continuous test scores that drive the teacher value-added tradition [1,2,4]. Treating these grades as continuous and fitting a Gaussian random-effects model is not a harmless simplification: it forces an artificial metric structure on an outcome whose meaning is fundamentally ordinal and whose tails are determined by regulatory cliffs (pass at 18; distinction at 30 e lode), rather than by smooth variation.

From a statistical point of view, these grades are ordinal responses. The natural language for such data is the cumulative logit, or proportional-odds, model [5, 6]. In a proportional-odds specification, the log-odds of being at or below any grade threshold is modeled as a linear function of observed covariates plus an ordered set of cutpoints; the effect of a covariate is allowed to shift the entire latent performance scale and thus affects all thresholds in parallel [5, 6]. This family, and its extensions such as partial proportional odds and generalized ordered logit [7], is standard in applied statistics for ordered categorical outcomes because it respects the ordering without pretending that adjacent categories are equally spaced in a Euclidean sense. Conceptually, one could embed professor heterogeneity in this framework by adding a professor-specific random intercept to a cumulative logit model: each professor shifts the latent grading standard up or down, which is exactly the ordinal analogue of a "strict vs. generous" professor effect in value-added language.

In theory, that hierarchical ordinal model is the clean solution. In practice, it fails to estimate on real university data at scale.

We analyze more than 1.2 million individual exam records from a large Italian university, spanning more than a decade and roughly one thousand professors. Each record contains the awarded grade; student-level covariates such as gender, off schedule status (on time vs. behind schedule), and age; exam year; and disciplinary area. When we attempt to fit a mixed-effects cumulative logit model with a professor-level random intercept—the direct analogue of a random teacher effect in a proportional-odds framework—maximum likelihood collapses. Estimation fails immediately with non-feasible starting values. The reason is structural. Some professors in the data are essentially deterministic: they award almost exclusively very high grades (30 or 30 e lode), while others cluster just above the pass threshold (18–22) and almost never award very high marks. This pattern is precisely what we wish to measure as "grading style". In a hierarchical cumulative logit, such behavior induces quasi-complete separation at the professor level: the model attempts to send that professor's random intercept to $+\infty$ or $-\infty$ to "explain" the near-deterministic outcomes, and the likelihood never attains a regular interior maximum. This is a classical pathology in logistic and ordinal regression under (almost-)complete separation, where the maximum likelihood estimator for a coefficient does not exist as a finite value [8,9]. In our application, this is not a rare edge case to be trimmed. It is widespread. The professors we most want to characterize—those who appear persistently "too generous" or "too severe"—are exactly the clusters that cause the hierarchical ordinal likelihood to become non-regular.

The first contribution of this paper is to propose and formalize an alternative that remains coherent for ordinal, truncated grades and that is actually computable on administrative data of this scale. We develop a two-stage estimator of professor-specific grading severity. In the first stage, we fit a proportional-odds cumulative logit model for the grade as a function of student gender, off schedule status, age and age squared, academic year, and disciplinary area, deliberately excluding professor identifiers. This pooled ordinal model yields, for each exam attempt, an estimated probability distribution over all admissible grades, from 18 through 30 e lode. From that distribution we compute an expected grade $\widehat{E}_i$ on the original 18–30L scale. Intuitively, $\widehat{E}_i$ is "what this student would be expected to receive, in this year and this area, given their observables, if examined by an average professor".

In the second stage, we define, for each professor $p$, a grading severity index:

$$S_p = \frac{1}{n_p} \sum_{i \in p} (Y_i - \widehat{E}_i),$$

where $Y_i$ is the realized grade for exam $i$, $\widehat{E}_i$ is the first-stage expected grade for that same student and context, and the sum runs over all exams graded by professor $p$. The scalar $S_p$ is directly interpretable. A value $S_p < 0$ means that, conditional on observed student characteristics, academic year, and disciplinary area, this professor systematically assigns grades below expectation; we refer to this as "severity". A value $S_p > 0$ means that the professor systematically assigns grades above expectation; we refer to this as "generosity". Because $\widehat{E}_i$ already conditions on observed student composition and context, $S_p$ is not just the professor's raw average grade. It is an average conditional deviation from expectation, on the actual transcript scale.

This construction does two things that the standard hierarchical ordinal model cannot presently deliver. First, it exists. We never attempt to estimate a professor-specific random intercept inside an ordinal likelihood that is trying to diverge to $\pm\infty$; instead, we estimate a pooled ordinal response surface that is well behaved [5–7] and then recover professor-level effects ex post as conditional residual means, which are always finite. Second, it is interpretable on the original grade scale. A difference of $-1$ in $S_p$ means "on average, one full point lower (e.g., 27 instead of 28) than expected", and a difference of $+2$ means "on average, two full points higher than expected", after holding constant gender, off schedule status, age, year, and disciplinary area. By contrast, the random intercept in a mixed-effects ordinal logit lives on a latent logit scale that is not directly communicable to students, faculty governance, or policy.

Empirically, the resulting dispersion in $S_p$ is large. Across nearly one thousand professors, the distribution of $S_p$ has a standard deviation of roughly 1.5 grade points on the 18–30L scale, and the distance between the lower and upper tails approaches five full points. This implies that, even after conditioning on student gender, progression status, age, year, and disciplinary area, two statistically comparable students can expect grades that differ by several whole points solely because they were examined by different professors. We also find that the mean of $S_p$ does not differ significantly between male and female professors, and that broad disciplinary areas do not exhibit statistically significant differences in average severity; most of the heterogeneity is genuinely professor-specific rather than area-wide. In other words, grading culture is not simply "STEM is strict, Humanities is soft", nor "men are harsher than women", at least on average. The dominant source of variation is the individual professor.

Methodologically, this framework connects two literatures that have, so far, remained largely separate. On one side is the value-added tradition, which treats teacher effects as persistent, policy-relevant quantities and studies their distribution, stability, and fairness [1–4]. On the other side is the proportional-odds tradition in ordinal regression, which is the appropriate statistical language for ordered, truncated, pile-up outcomes such as university exam grades [5–7]. The grading severity index $S_p$ is an explicit bridge between the two: it is defined on the transcript scale, it is estimable in the presence of quasi-complete separation that makes hierarchical ordinal likelihoods non-regular [8,9], and it supports standard inferential tools such as shrinkage, cluster-resampling, temporal stability analyses, and predictive calibration checks via distribution-free coverage ideas [10,11].

## 2. Statistical Framework

The goal of this section is to formalize the object we ultimately care about—professor-level grading severity—and to show how it can be estimated in a way that is coherent for ordinal grades, robust to the pathologies that make hierarchical ordinal likelihoods non-regular, and interpretable on the transcript scale. We proceed in four steps. First, we describe the data structure and notation. Second, we define the baseline ordinal response model used to learn the grading surface as a function of student and contextual covariates. Third, we construct the professor-level severity index $S_p$. Finally, we discuss its inferential properties and how it can be embedded in a broader robustness and fairness analysis.

### *2.1. Data Structure and Notation*

Let $Y_i$ denote the grade awarded in exam attempt $i$. In the university under study, grades lie on a discrete ordered scale:

$$Y_i \in \{18, 19, \ldots, 30, 30\text{ e lode}\}.$$

By institutional rule, 18 is the lowest passing grade; values below 18 are not recorded numerically in the official transcript data but appear instead as "not passed", so they do not enter the dataset as numeric scores. At the top of the scale, 30 e lode is a special distinction category above the regular maximum of 30. For estimation and exposition, we map 30 e lode to the numeric value 31. This creates a strictly increasing numerical encoding $\{18, \ldots, 31\}$ which preserves the ordinal information and allows us to compute expectations on the same scale that students and faculty actually use. We refer to this numeric encoding interchangeably as the "18–30L" scale.

For each exam attempt $i$, we observe:

- a vector $X_i$ of student-level covariates: student gender, off schedule status (on-time vs. behind schedule), age, and age squared (to allow nonlinearity in age effects),
- contextual covariates: exam year (calendar dummies) and disciplinary area (area dummies),
- the identity of the examining professor, which we denote $p(i)$.

The dataset contains more than 1.2 million such exam attempts over more than a decade, with nearly one thousand distinct professors. This scale is double-edged. On one hand, the large number of observations per professor makes it meaningful to speak of a stable professor-specific grading pattern. On the other hand, professors with extremely concentrated grading behavior—e.g., almost always 30/30L, or almost always just-above-pass—induce quasi-complete separation in any likelihood that attempts to estimate professor effects directly as random intercepts in an ordinal logit. As discussed in Section 1, such quasi-complete separation leads to non-existence or non-regularity of the maximum likelihood estimator [8,9], and in practice produces immediate convergence failure in hierarchical cumulative logit estimation.

### *2.2. Baseline Ordinal Model*

The statistical structure of $Y_i$ is inherently ordinal [5,6]. The natural starting point is the proportional-odds cumulative logit model. Index the ordered grade categories by $c \in \{18, 19, \ldots, 31\}$, where 31 encodes 30 e lode. The proportional-odds model can be written as:

$$\Pr(Y_i \leq c \mid X_i) = \text{logit}^{-1}(\gamma_c - X_i^\top \beta), \quad c = 18, 19, \ldots, 31, \tag{1}$$

where $\text{logit}^{-1}(z) = \frac{1}{1+\exp(-z)}$, the $\gamma_c$ are strictly increasing threshold (cutpoint) parameters, and $\beta$ is a vector of slope coefficients.

Intuitively, $\gamma_c$ locates the boundary between "grade at most $c$" and "grade above $c$", while $X_i^\top \beta$ shifts the entire latent performance scale according to student and contextual characteristics. The proportional-odds restriction implies that $X_i$ has a parallel effect across all cutpoints: the log-odds ratio

$$\log\left(\frac{\Pr(Y_i \leq c \mid X_i)}{\Pr(Y_i > c \mid X_i)}\right) \tag{2}$$

changes linearly with $X_i$ by the same amount for all $c$. This model is widely used for ordered categorical outcomes because it respects ordering without imposing an arbitrary metric spacing between adjacent categories [5,6]. It is also computationally stable on large samples in the absence of high-dimensional random effects.

In principle, one might place professor-specific random intercepts $\alpha_{p(i)}$ into this model:

$$\Pr(Y_i \leq c \mid X_i, p(i)) = \text{logit}^{-1}(\gamma_c - X_i^\top \beta - \alpha_{p(i)}), \quad \alpha_p \sim \mathcal{N}(0, \sigma_\alpha^2), \tag{3}$$

to obtain an explicit "strict vs. generous professor" effect in the latent logit scale, directly analogous to teacher random effects in value-added models for continuous test scores [1–4].

Empirically, however, this hierarchical proportional-odds model is not estimable in our data because of quasi-complete separation at the professor level: for some professors, the conditional probability of being in the topmost categories (30, 30 e lode) is effectively one; for others, the conditional probability of being just above the pass threshold is effectively one; and the maximum likelihood estimator responds by attempting to send $\alpha_p \to \pm\infty$ [8,9].

For that reason, we estimate instead a pooled proportional-odds model without professor identifiers:

$$\Pr(Y_i \leq c \mid X_i) = \text{logit}^{-1}(\gamma_c - X_i^\top \beta), \tag{4}$$

where $X_i$ now includes:

1. student gender,
2. off schedule status,
3. age and age squared,
4. exam year fixed effects,
5. disciplinary area fixed effects.

Call the fitted parameters $\widehat{\beta}$ and $\widehat{\gamma}_c$. From this first-stage fit we recover, for each exam attempt $i$, the entire predicted probability mass function:

$$\widehat{p}_{ik} = \Pr(Y_i = k \mid X_i; \widehat{\beta}, \widehat{\gamma}), \quad k \in \{18, \ldots, 31\}. \tag{5}$$

Because the support is discrete and finite, we can map these probabilities into an expected grade on the transcript scale:

$$\widehat{E}_i = \sum_{k=18}^{31} k \widehat{p}_{ik}. \tag{6}$$

The scalar $\widehat{E}_i$ can be interpreted as follows: given the observable profile of student $i$ (gender, off schedule status, age), in that exam year and that disciplinary area, what grade would we predict for this student if the exam were graded by an "average" professor?

### *2.3. Professor Severity Index*

For each professor $p$, let $I_p = \{i : p(i) = p\}$ denote the set of exam attempts graded by that professor, and let $n_p = |I_p|$ be the number of such attempts. We define the grading severity index of professor $p$ as:

$$S_p = \frac{1}{n_p} \sum_{i \in I_p} (Y_i - \widehat{E}_i). \tag{7}$$

It is useful to clarify the interpretation of $S_p$ at the outset. The index should be read as a descriptive measure of conditional grading differences relative to the benchmark implied by the pooled ordinal model, rather than as a strictly causal professor effect. In other words, $S_p$ summarizes how much professor $p$ systematically grades above or below the grade that would be expected for observationally similar students in the same year and disciplinary area, conditional on the covariates included in the first-stage model.

In this sense, the proposed estimator is conceptually analogous to residual-based teacher effects in the value-added literature: in both cases, the object of interest is a persistent unit-specific deviation from an outcome benchmark conditional on observables. At the same time, our estimator is not a hierarchical random effect estimated directly within a mixed-effects ordinal likelihood. Instead, it is a professor-level average of conditional residuals constructed ex post from a pooled ordinal model, which is precisely what makes it robust to the separation problems discussed above.

By construction, $S_p < 0$ indicates systematic severity: conditional on observables, professor $p$ awards grades below expectation. Symmetrically, $S_p > 0$ indicates systematic generosity. Because $\widehat{E}_i$ already adjusts for student gender, off schedule status, age (nonlinearly), calendar time, and disciplinary area, differences in $S_p$ are not mechanically driven by professor $p$ teaching weaker or stronger cohorts, or by grading in traditionally "hard" or "soft" areas.

Two features of $S_p$ are worth emphasizing. First, it is defined directly on the transcript scale. A difference of $-1$ in $S_p$ means "on average, one full point lower than expected", holding observables fixed. Second, the estimator for $S_p$ is always finite as long as $n_p \geq 1$. We never have to estimate a per-professor parameter that the likelihood attempts to send to $\pm\infty$.

### *2.4. Inference, Stability, and Robustness*

Because $S_p$ is an average of $Y_i - \widehat{E}_i$ over exams graded by professor $p$, its sampling variability shrinks at approximately the usual $1/\sqrt{n_p}$ rate. This suggests two inferential strategies parallel to the teacher value-added literature.

First, uncertainty can be quantified via a professor-level cluster bootstrap: repeatedly resample professors with replacement, recompute $S_p$, and use the bootstrap distribution to form confidence intervals.

Second, empirical Bayes shrinkage can be applied to regularize noisy estimates for professors with small $n_p$, pulling extreme values toward the grand mean in proportion to their estimated noise variance [3,4].

Finally, the pooled proportional-odds model used to generate $\widehat{E}_i$ can be generalized via partial proportional odds or generalized ordered logit models [7]. Comparing baseline $S_p$ to severity indices computed under relaxed specifications provides a sensitivity analysis of the proportional-odds assumption.

Because the first-stage ordinal model yields the full predictive distribution $\widehat{p}_{ik}$, we can also construct distribution-free prediction sets and assess subgroup calibration using conformal methods [10,11].

### *2.5. Practical Implementation*

The proposed two-stage estimator can be implemented with standard ordinal regression software and simple post-estimation steps. In practice, the workflow is:

1. Fit a pooled proportional-odds cumulative logit. Estimate an ordered logit / proportional-odds model of the exam grade $Y_i$ on observed covariates $X_i$ (student gender, off schedule status, age and age squared, exam year fixed effects, and disciplinary area fixed effects), explicitly excluding professor identifiers. This delivers estimates $\widehat{\beta}$ and $\widehat{\gamma}_c$ for the cutpoints. Computationally this is a single call to a standard ordered logit routine (e.g., `ologit` in Stata).
2. Obtain the full predicted grade distribution for each exam attempt. For each observation $i$, use $\widehat{\beta}, \widehat{\gamma}$ to compute the predicted probability mass function

$$\widehat{p}_{ik} = \Pr(Y_i = k \mid X_i; \widehat{\beta}, \widehat{\gamma}), \quad k \in \{18, \ldots, 31\},$$

   where 31 encodes 30 e lode. This gives, for each exam attempt, a predicted distribution over all admissible transcript grades.
3. Compute an expected transcript-grade benchmark. Map the predicted distribution into an expected grade on the institutional 18–30L scale:

$$\widehat{E}_i = \sum_{k=18}^{31} k\widehat{p}_{ik}.$$

   Intuitively, $\widehat{E}_i$ is "what an average professor would have awarded to this student in this year and this disciplinary area, given observables".
4. Form professor-level residual means. For each professor $p$, define

$$S_p = \frac{1}{n_p} \sum_{i \in I_p} (Y_i - \widehat{E}_i),$$

   where $I_p$ is the set of exams graded by professor $p$ and $n_p = |I_p|$. A negative $S_p$ indicates systematic severity (grades below expectation, conditional on observables). A positive $S_p$ indicates systematic generosity (grades above expectation).
5. Quantify uncertainty and persistence. Precision can be assessed via a professor-level cluster bootstrap that re-samples professors and re-computes $\{S_p\}$. Temporal persistence can be studied by computing $S_{p,t}$ at the professor–year level and regressing $S_{p,t+1}$ on $S_{p,t}$.

This procedure is intentionally simple: it only requires (1) a single pooled ordered logit fit and (2) algebraic post-processing of the fitted probabilities. It avoids the non-regularity of hierarchical proportional-odds models with professor random intercepts, which fail to converge in our data due to quasi-complete separation [8,9].

## 3. Simulation Study

This section provides a small simulation study with two goals. First, we illustrate why the standard hierarchical proportional-odds model with professor-level random intercepts fails to estimate in the presence of quasi-complete separation. Second, we show that the proposed severity index $S_p$ recovers the underlying "strict vs. generous" differences across professors.

### *3.1. Design*

We generate synthetic exam data for $P = 50$ professors. Each professor $p$ is assigned a true latent severity parameter $\theta_p$, drawn from a Normal distribution with mean 0 and standard deviation 1.5:

$$\theta_p \sim \mathcal{N}(0, 1.5^2).$$

Conditional on $\theta_p$, each of that professor's $n_p$ students receives a grade $Y_i$ on the ordinal 18–30L scale. Grades are drawn from a proportional-odds (cumulative logit) data-generating process in which $\theta_p$ enters as a professor-specific intercept shift. Specifically, for grade thresholds indexed by $c \in \{18, \ldots, 31\}$:

$$\Pr(Y_i \leq c \mid X_i, p(i) = p) = \text{logit}^{-1}(\gamma_c - X_i^\top \beta - \theta_p). \tag{8}$$

We set the cutpoints $\{\gamma_c\}$ to mimic the empirical distribution observed in the administrative data, with high mass at 18 and at 30/30L. We allow some professors to be extremely generous (large positive $\theta_p$) or extremely severe (large negative $\theta_p$), so that for some $p$ almost all simulated grades are at the very top (30 or 30 e lode), and for others almost all are just above the passing threshold.

### *3.2. Estimators*

We then fit two estimators on each simulated dataset:

1. Hierarchical proportional-odds model. A mixed-effects cumulative logit with a professor-level random intercept (the direct ordinal analogue of a teacher random effect). In Stata syntax this corresponds to `meologit grade X || prof:`.
2. Two-stage estimator. (i) Fit a pooled ordered logit excluding professor identifiers; (ii) Compute $\widehat{E}_i$ from the fitted probabilities; (iii) Form

$$S_p = \frac{1}{n_p} \sum_{i \in I_p} (Y_i - \widehat{E}_i).$$

### *3.3. Evaluation*

For each professor we record the "truth" $\theta_p$ from the data-generating process. For each estimator we record an estimated severity:

- For the hierarchical model: the estimated random intercept $\widehat{\alpha}_p$;
- For our method: $S_p$.

We repeat the simulation 500 times and compute, across professors and replications:

- The mean absolute error (MAE) between $\widehat{\alpha}_p$ and $\theta_p$, and between $S_p$ and $\theta_p$;
- The share of replications in which the hierarchical model fails to converge.

### *3.4. Findings*

Two robust patterns emerge.

(i) The hierarchical proportional-odds estimator frequently fails to converge. When some professors are nearly deterministic at the top or bottom of the scale, the hierarchical model returns "initial values not feasible" or produces random-effect estimates that diverge numerically. Non-convergence occurs in more than half of the simulated datasets once we allow realistic "extreme" professors, exactly mirroring what we observe in the administrative university data. This is the finite-sample manifestation of quasi-complete separation and the non-existence of a finite maximum likelihood estimate in logistic/ordinal regression [8,9].

(ii) The two-stage $S_p$ estimator always exists and tracks $\theta_p$ closely. Across converged replications, the mean absolute error of $S_p$ relative to $\theta_p$ is similar in magnitude to the error of the hierarchical estimator $\widehat{\alpha}_p$ when that estimator converges, and $S_p$ remains well-defined in all replications (no divergence). In other words, $S_p$ delivers a finite, transcript-scale estimate of professor severity even exactly in the settings where the likelihood-based hierarchical ordinal model breaks down.

### *3.5. Interpretation*

The simulation confirms the core message of this paper. In the presence of quasi-complete separation at the professor level—i.e., the empirically relevant case in which some professors are "all 30 e lode" and others hug the pass threshold—the textbook mixed-effects cumulative logit fails for structural reasons: the MLE tries to send the random intercepts to $\pm\infty$.

Our two-stage procedure avoids that failure by (i) estimating a pooled ordered logit that remains regular, and (ii) recovering professor-specific severity as a conditional residual average on the original grade scale. The resulting $S_p$ is finite, interpretable in transcript points, and empirically stable.

## 4. Data and Descriptive Evidence

The empirical analysis uses an administrative census of university examination records from a large Italian university, covering all recorded exam attempts between 2007 and 2019. The raw dataset contains 1,425,371 graded observations. Each observation corresponds to a single student sitting a specific exam in a specific session, evaluated

by a specific professor. For each exam attempt we observe: (i) the awarded final grade; (ii) student-level covariates, including gender, off schedule status (on-time versus behind schedule), and age; (iii) contextual information, including the calendar year of the exam and a disciplinary area classification of the course; and (iv) the identity of the examining professor. Summary statistics for these variables, and details on sample construction, are reported in Appendix A, Table A1.

Grades in this system lie on a discrete, ordered scale $\{18, 19, \ldots, 30, 30\text{ e lode}\}$. By institutional rule, 18 is the minimum passing grade and appears in the transcript as such; grades below 18 are not recorded numerically but are stored administratively as "not passed", and thus do not enter our dataset as numeric values. At the top of the scale, 30 e lode is a special distinction above the regular maximum of 30. For estimation, we encode 30 e lode as 31, yielding a strictly increasing numeric support $\{18, \ldots, 31\}$. This encoding preserves the order of grades and allows us to work in "grade points" on the same 18–30L scale that students and faculty actually interpret.

The empirical grade distribution is strongly non-Gaussian. There is substantial mass at the lower institutional threshold (18), and strong heaping at the upper end (30 and 30 e lode). The interior of the scale (e.g., 22–27) is populated but comparatively flatter. Appendix A, Figure A1 plots the empirical distribution of awarded grades on the 18–30L scale and makes clear that pile-ups at the institutional cliffs dominate the shape of the data. This structure—truncated below 18, discretized, and with visible spikes at salient institutional cutpoints—is exactly what makes linear-Gaussian approximations inappropriate and motivates an explicitly ordinal treatment [5,6]. It is also what creates quasi-complete separation in hierarchical ordinal likelihoods: professors who nearly always award 30/30L or who hug the pass threshold generate near-deterministic outcomes within their cluster, which in turn drives random-intercept estimates to $\pm\infty$ and causes non-existence of the finite MLE [8,9]. We document this failure of the mixed-effects cumulative logit formally in Appendix A, Table A2, where we report the likelihood diagnostics from attempts to estimate a hierarchical proportional-odds model with professor-level random intercepts.

To quantify professor-specific grading behavior, we construct for each professor $p$ the severity index

$$S_p = \frac{1}{n_p} \sum_{i \in I_p} (Y_i - \widehat{E}_i)$$

as defined in Section 2. Here $Y_i$ denotes the realized grade for exam attempt $i$, on the 18–30L scale, and $\widehat{E}_i$ is the expected grade predicted by a pooled proportional-odds cumulative logit that conditions on observed student characteristics (gender, off schedule status, age and age squared) and contextual factors (exam year and disciplinary area), but deliberately excludes professor identifiers. Intuitively, $\widehat{E}_i$ represents "what an average professor would have awarded to this student in this year and this area, given these observables", and $S_p$ averages the deviations from that benchmark over all exams graded by professor $p$. Negative values of $S_p$ indicate systematic severity (grades below expectation, conditional on observables); positive values indicate systematic generosity (grades above expectation, conditional on observables).

In order to obtain meaningful professor-level statistics and to avoid overinterpreting professors who appear only a handful of times in the data, we restrict attention to professors who graded at least 20 passed exam attempts over the observation window. This leaves on the order of one thousand professors (see Appendix A, Table A3 for final counts after all restrictions). This trimming rule plays the same role as minimum-class-size or minimum-teacher-load restrictions that are standard in the teacher value-added literature, where extremely small-sample teachers are typically excluded or heavily shrunk to avoid spurious extremes [3,4]. Appendix A, Table A1 reports the distribution of the number of graded exams per professor after this restriction.

The cross-sectional distribution of $S_p$ across these 988 professors exhibits substantial dispersion. The mean of $S_p$ is approximately 0.21 grade points, with a standard deviation around 1.5 grade points on the 18–30L scale. The interquartile range runs from about $-0.83$ to $+1.24$ grade points. The 5th percentile is roughly $-2.26$, and the 95th percentile is roughly $+2.70$. Both tails extend further: the lower tail reaches below $-3$, and the upper tail exceeds $+5$. Appendix A, Table A3 reports summary statistics and selected quantiles of the empirical $S_p$ distribution, and Appendix A, Figure A2 plots its kernel density and marks the 5th, 25th, 50th, 75th, and 95th percentiles.

These magnitudes are not minor. Interpreted literally, they imply that, holding constant observable student characteristics (gender, off schedule status, age), exam year, and disciplinary area, a student graded by a professor in the generous tail of the $S_p$ distribution can expect, on average, a final mark that is three to five full grade points higher than an observationally comparable student graded by a professor in the severe tail. In an exam regime where a one-point shift (e.g., from 27 to 28) is academically and administratively meaningful, a persistent gap of several points is striking.

We next relate $S_p$ to observable professor characteristics. We compare the mean $S_p$ between male and female professors using a standard two-sample comparison of means. The average difference is small (on the order of 0.05 grade points) and statistically indistinguishable from zero under an equal-variance $t$-test. Appendix A, Table A4 reports these results. This suggests that, once we condition on student mix and exam context, there is no systematic evidence that male professors are "harsher" or "more generous" than female professors in the sense of awarding consistently lower or higher conditional grades.

We also examine heterogeneity in $S_p$ across disciplinary areas. We compute area-level averages of $S_p$ and test for equality of means across areas using a one-way analysis of variance. The resulting $F$-statistic is small and the associated $p$-value is large, implying no statistically significant differences in average conditional severity across broad areas. Table A5 reports the area-level means of $S_p$, the ANOVA results, and post-hoc multiple-comparison adjustments. The substantive implication is that grading severity is not simply "STEM vs. Humanities". Within-area dispersion in $S_p$ dominates between-area differences in the mean. The primary source of heterogeneity is therefore the individual professor, not the macro-discipline.

In summary, the data reveal three empirical facts. First, conditional grading standards vary substantially and persistently across professors, by amounts that are large on the transcript scale. Second, this heterogeneity is not reducible to simple compositional differences in which students each professor teaches, nor is it captured by coarse observables such as professor gender or disciplinary area. Third, the pattern of extreme generosity and extreme severity is precisely the structure that renders standard hierarchical proportional-odds models non-regular due to quasi-complete separation [8,9], and simultaneously motivates our two-stage construction of $S_p$.

## 5. Results and Robustness

This section proceeds in four steps. We begin by documenting the overall dispersion of professor-specific grading severity on the transcript scale. We then examine whether these differences are persistent over time at the professor-year level. Next, we assess robustness to alternative ordinal specifications. Finally, we evaluate predictability and fairness by examining whether the model's predictive performance is similar across observable student subgroups.

This section presents the empirical behavior of the professor-level grading severity index $S_p$ defined in Section 2. We address four questions. First, how large is the heterogeneity in $S_p$ on the transcript scale, and to what extent can it plausibly be dismissed as sampling noise? Second, is this heterogeneity stable within professors over time, or is it just transitory volatility? Third, how sensitive are our conclusions to the proportional-odds assumption in the first-stage ordinal model? Fourth, does the grading process exhibit differential predictability, and therefore differential transparency, across observable student subgroups?

### *5.1. Magnitude and Precision of Professor Severity*

Table 1 summarizes the magnitude of professor-level grading severity. The cross-sectional distribution of $S_p$ across the 988 professors who graded at least 20 exams is wide. The cross-sectional distribution of $S_p$ across the 988 professors who graded at least 20 exams is wide. As reported in Appendix A, Table A3, the mean of $S_p$ is approximately 0.21 grade points, with a standard deviation of about 1.5 grade points on the 18–30L scale. The interquartile range extends from roughly $-0.83$ to $+1.24$. The tails are economically large: the 5th percentile is around $-2.26$, and the 95th percentile is around $+2.70$. Both tails extend further, with some professors below $-3$ and some above $+5$. Appendix A, Figure A2 displays the kernel density of $S_p$, highlighting the bulk of the distribution and the lower and upper tails.

**Table 1.** Summary of professor-level severity magnitudes

| Mesaures | Value | Interpretation |
|---|---|---|
| Mean of $S_p$ | 0.21 | Average deviation from benchmark |
| Std. Dev. of $S_p$ | 1.50 | Dispersion of grading severity |
| Interquartile Range of $S_p$ | $[-0.83,\ 1.24]$ | Middle 50% of professors |
| 5th–95th Percentile Range | $[-2.26,\ 2.70]$ | Central mass excluding tails |
| Extreme Range | $< -3$ to $> 5$ | Full observed spread |
| Persistence ($\rho$) | $> 0.5$ to $> 0.7$ | Stability over time |

Interpreting these magnitudes is straightforward because $S_p$ is expressed in transcript points. A professor at the 95th percentile of $S_p$ assigns, on average, grades that are more than two and a half points higher than expected for comparable students in comparable contexts, relative to the pooled ordinal benchmark. A professor at the 5th percentile assigns more than two points lower than expected. The gap between these tails approaches five full grade points.

A natural concern is that these differences might simply be sampling variability. By construction, $S_p$ already partials out observed composition via $\widehat{E}_i$, so differences in student mix along the observed covariates cannot be driving these tails. But there remains sampling uncertainty in $S_p$ itself, because some professors grade more exams than others.

To quantify this, we implement a professor-level cluster bootstrap. We resample professors with replacement, recompute $S_p$, and use the bootstrap distribution to form confidence intervals. Two facts emerge. First, almost all intervals are tight: the half-width of the 95% interval is typically well below one grade point. Second, even after accounting for sampling uncertainty, both tails remain far from zero. The extremes are not noise professors; they are professors with systematically and precisely estimated conditional grading standards that deviate sharply from the benchmark.

### *5.2. Temporal Stability*

A core argument in the teacher value-added literature is that teacher effects are persistent [1–4]. An analogous question here is whether professor severity is stable over time.

To investigate this, we re-estimate the severity index at the professor–year level:

$$S_{p,t} = \frac{1}{n_{p,t}} \sum_{i \in I_{p,t}} (Y_i - \widehat{E}_i), \tag{9}$$

where $I_{p,t}$ is the set of exams graded by professor $p$ in year $t$. We then estimate:

$$S_{p,t+1} = \alpha + \rho S_{p,t} + u_{p,t}, \tag{10}$$

restricting to professor–year cells with sufficient observations in both periods. Appendix B, Table A4 reports estimates of $\rho$ under alternative minimum-cell thresholds.

The estimated persistence parameter $\rho$ is well above 0.5 and often above 0.7, with tight standard errors. Professors who are strict relative to expectation in one year remain strict in the next, and likewise for generous professors. Severity is therefore a stable individual trait rather than transitory volatility.

### *5.3. Sensitivity to the First-Stage Ordinal Specification*

Our construction of $S_p$ depends on the first-stage proportional-odds model [5,6]. To test robustness, we replace it with a generalized ordered logit specification that relaxes the parallel-slopes restriction for selected covariates [7]. We recompute expected grades under this specification and construct the corresponding professor-level severity index, denoted $S_p^{(gpo)}$.

The resulting severity measure is highly correlated with the baseline index $S_p$, with a correlation exceeding 0.9. Professors identified as strict or generous under proportional odds remain so under the generalized specification. Differences arise mainly in the middle of the distribution, typically for professors whose grading behavior is threshold-specific rather than uniformly shifted.

The heavy tails and substantive dispersion in severity remain intact. The main conclusions are therefore not artifacts of the proportional-odds assumption.

As an additional robustness check, we re-estimate the first-stage model using an ordered probit specification instead of the ordered logit. We then recompute the expected grades and construct the corresponding severity index.

The resulting professor-level severity measures are virtually identical to the baseline estimates. The correlation between the two indices exceeds 0.999, and rank correlation is similarly close to one. Dispersion is unchanged (standard deviation of approximately 1.52 under both specifications), and the range of the distribution remains stable. Differences between the two measures are economically negligible.

These results indicate that the dispersion, persistence, and ranking of professors are not driven by the choice of link function, but instead reflect stable structure in the data. Detailed comparison statistics are available upon request.

We also examine robustness to the functional form used to summarize the predicted ordinal distribution. In the baseline specification, the expected grade is used as the individual benchmark. As an alternative, we construct severity measures using the median and modal predicted grade.

The resulting professor-level indices are highly consistent with the baseline measure. Professor rankings and the dispersion of severity remain essentially unchanged, and differences are concentrated in a small set of observations near threshold values. This indicates that the main findings are not driven by the use of the expected value, nor by the numerical coding of top grades.

### *5.4. Predictability and Fairness*

Because the first-stage ordinal model yields the full predictive distribution $\widehat{p}_{ik} = \Pr(Y_i = k \mid X_i)$, we can assess predictability across student subgroups.

We construct prediction sets using conformal-style distribution-free methods [10, 11]. For each exam attempt $i$, we form a prediction set $C_i$ such that $\Pr(Y_i \in C_i)$ achieves nominal coverage (e.g., 90%). We then examine empirical coverage rates separately for subgroups such as male vs. female students and on-time vs. off schedule students.

Prediction sets are generally tight, often consisting of two or three adjacent grade values. Empirical coverage appears broadly similar across subgroups, with no clear evidence of systematic undercoverage for specific observable groups.

These findings suggest that, although professors differ sharply in level ($S_p$), grading does not appear systematically less predictable for particular student subgroups once we condition on observables.

### *5.5. Summary*

Four main findings emerge. First, professor-specific grading standards differ substantially on the transcript scale, with tail gaps approaching five grade points. Second, these differences are precisely estimated and highly persistent over time. Third, they are robust to relaxing the proportional-odds assumption. Fourth, while grading standards differ in level across professors, we do not find evidence of systematic subgroup undercoverage in predictive distributions.

Taken together, these results indicate that grading standards at the university level are measurable, persistent, and large. They are attached to individual professors rather than broad disciplinary categories, and they are compatible with fairness constraints in the sense that we do not observe systematic subgroup differences in predictability.

## 6. Discussion and Institutional Implications

The analysis above shows that grading standards at the university level are not uniform. Conditional on observable characteristics of the student (gender, off schedule status, age), on exam timing (calendar year), and on disciplinary area, professors differ systematically and persistently in how high or low they grade, and those differences are large in transcript units. The dispersion in the severity index $S_p$ is not a statistical curiosity in the tails: even between the 25th and 75th percentiles of the distribution, the difference in conditional grading standards exceeds one full grade point (Appendix A, Table A3), and between the lower and upper tails the gap approaches five full points. Moreover, this is not transitory fluctuation. Professor-specific severity is highly persistent over time (Section 5.2; Appendix B, Table A4), in the same sense in which teacher value-added is persistent in primary and secondary education [1–4]. From an institutional perspective, these are structural features of the assessment environment.

What does this mean substantively?

First, the existence of persistent, professor-specific grading standards implies that exam outcomes are partly an allocation problem, not just an achievement problem. Two otherwise comparable students who happen to be evaluated by different professors face different effective grading regimes, even conditional on the covariates we can observe and adjust for. When those grading regimes differ by multiple transcript points in expectation, assignment to professors becomes consequential. This mirrors the way assignment to high–value-added versus low–value-added teachers has been shown to affect long-run outcomes such as college attendance and earnings in school settings [4], but with an important twist: here the outcome is itself the credential. In university settings where cumulative grade averages determine access to scholarships, admission to selective Master's and PhD programs, thesis supervision opportunities, or early-stage hiring screens, persistent differences in grading severity across professors mechanically propagate into downstream opportunities.

Second, the fact that severity is professor-specific rather than area-specific matters normatively. We find no evidence that average conditional severity differs meaningfully across broad disciplinary areas (Section 4; Table A5), and we find no systematic difference in average severity by professor gender (Appendix A, Table A5). The dominant source of heterogeneity is within-area, at the individual professor level.

Third, the high persistence of $S_p$ raises governance questions. If a professor is consistently severe ($S_p < 0$) year after year, this is not noise but a stable policy-relevant characteristic of that professor's grading practice. The same applies to consistently generous professors ($S_p > 0$). An institution that takes equity in assessment seriously could monitor these systematic deviations much like school systems monitor teacher value-added [3,4]. The goal would not be mechanical homogenization, but transparency and review of persistent outliers.

Our framework makes such monitoring technically feasible. Because $S_p$ is defined in transcript points, it is legible to administrators and faculty governance bodies. Because it is based on an ordinal model that respects the 18–30L structure and institutional thresholds [5–7], the comparison is statistically coherent. And because it avoids the non-regularity of hierarchical cumulative logit models under quasi-complete separation [8,9], it is computationally scalable to administrative data.

A related approach would be to estimate a Bayesian cumulative logit model with professor-level random effects under weakly informative priors, which can regularize the separation problem. While such an approach would provide an alternative way to recover professor heterogeneity, it would do so on the latent scale and at the cost of reduced transparency.

By contrast, the proposed two-step severity index is directly interpretable on the transcript scale and remains well-defined without requiring distributional assumptions on professor effects. Exploring the relationship between these approaches in detail is a promising direction for future research.

At the same time, $S_p$ should not be misinterpreted as a causal effect in the strict sense. An important direction for future research would be to combine this framework with settings that approximate quasi-random assignment of students to professors. In such environments, the same two-stage approach could be used to move from descriptive severity measures toward more credible causal estimates of professor effects. The teacher value-added literature devotes substantial attention to identification strategies that isolate teacher contributions net of sorting [3, 4]. Our construction of $S_p$ is descriptive and conditional. It adjusts for observed composition but does not claim that reassigning a student from professor $p$ to professor $q$ would shift that student's grade by exactly $S_q - S_p$. Establishing such causal claims would require quasi-random assignment of students to professors, for example in large multi-section courses with rotating examiners. We view causal identification as an important direction for future research.

One additional dimension concerns fairness. A natural concern is that heterogeneity in grading standards could combine with patterns of assignment to create systematic disadvantage for certain student groups. While we cannot fully resolve allocation questions without detailed timetable and assignment data, we can evaluate conditional predictability. Using conformal-style prediction sets derived from the first-stage ordinal model [10,11], we find no clear evidence of systematic subgroup undercoverage. In other words, although professors differ in level, grading appears broadly similar in predictability across observable student subgroups once we condition on covariates.

Finally, the methodological issue we identify is not unique to this institution. Many university systems use discrete grading schemes with hard pass thresholds and explicit honors categories. In such environments, hierarchical ordinal models with professor random effects can fail due to quasi-complete separation, particularly when some professors exhibit near-deterministic grading patterns. The two-stage construction of $S_p$ is portable because it relies only on grade data, student covariates, contextual controls, and examiner identifiers—information typically available in administrative records.

In summary, the university grading process embeds a persistent, professor-level component that is large in transcript units. This component is structurally analogous to persistent teacher effects in the value-added literature [1–4], but it arises in an ordinal, truncated regime where standard hierarchical ordinal likelihoods may fail. Our framework makes this component observable and auditable. The appropriate institutional response is not automatic standardization, but informed transparency, systematic review of persistent outliers, and integration of grading diagnostics into academic governance.

## 7. Conclusions

This paper develops a method to measure professor-specific grading standards in a university setting where grades are discrete, ordinal, and institutionally truncated. The core difficulty is structural. In our environment, exam grades live on an ordered grid $\{18, 19, \ldots, 30, 30\text{ e lode}\}$; 18 is the first legally passing grade; 30 e lode is a special distinction state above the formal maximum of 30; and failures below 18 are not recorded numerically. The resulting distribution is spiky, with mass at institutional cliffs (18, 30, 30 e lode) and relatively thinner support in the interior.

In such data, the standard machinery used in the teacher value-added literature—hierarchical linear models on continuous scores with teacher fixed or random effects [1–4]—is conceptually inappropriate because the outcome is not approximately continuous. The obvious ordinal analogue—a mixed-effects proportional-odds model with

professor random intercepts [5,6]—is, in practice, often non-estimable. In the presence of quasi-complete separation, professors who almost always award very high grades, or who consistently cluster near the pass threshold, cause the likelihood to attempt to send their random intercepts to $\pm\infty$, and the maximum likelihood estimator ceases to exist as a finite interior solution [8,9].

Our contribution is to replace that non-regular likelihood problem with an estimable and interpretable object. We first fit a pooled proportional-odds cumulative logit model for exam grades as a function of observable student characteristics (gender, off schedule status, age and age squared) and contextual controls (exam year and disciplinary area), deliberately excluding professor identity. This model respects the ordinal structure of grades and the pile-ups at institutional cutpoints [5–7]. From it, we recover for each exam attempt the full predicted distribution over admissible grades and its expected value $\widehat{E}_i$, interpretable as what an average professor would award to that student in that context.

We then define, for each professor $p$, a grading severity index

$$S_p = \frac{1}{n_p} \sum_{i \in I_p} (Y_i - \widehat{E}_i),$$

the average conditional deviation between realized and predicted grades on the 18–30L scale. Negative values of $S_p$ indicate systematic severity; positive values indicate systematic generosity. Unlike a random intercept in a mixed-effects ordinal logit, $S_p$ is always finite and directly interpretable in transcript points.

Using more than 1.2 million examination records from a large Italian university (2007–2019), we document three main findings.

First, grading standards vary substantially across professors. The dispersion in $S_p$ is large on the transcript scale: the interquartile range spans more than one full grade point, and the gap between the lower and upper tails approaches five full grade points even after conditioning on student characteristics, year, and disciplinary area.

Second, professor severity is persistent. When estimated at the professor–year level, the severity index exhibits strong year-to-year stability, analogous to persistence findings in the teacher value-added literature [1–4].

Third, our results are robust to relaxing the proportional-odds assumption and do not generate systematic subgroup undercoverage in conformal-style predictive calibration [10,11]. While professors differ sharply in level, we do not find evidence that grading is systematically less predictable for specific observable student groups once we condition on covariates.

Substantively, these findings imply that university grades reflect not only student performance but also persistent professor-specific grading standards. In institutional settings where grades determine access to scholarships, postgraduate programs, and labor-market opportunities, such heterogeneity has distributional consequences.

Methodologically, the paper demonstrates that in ordinal, truncated grading regimes with strong pile-ups at institutional thresholds, hierarchical cumulative logit models may fail due to quasi-complete separation. The proposed two-stage construction of $S_p$ provides a scalable and interpretable alternative that remains coherent under these conditions.

Two directions for future research follow naturally. First, causal identification of professor severity would require quasi-random assignment of students to examiners, for example in multi-section courses with rotating examiners. Second, the integration of grading diagnostics, predictive calibration checks, and severity monitoring into routine academic governance could enhance transparency without imposing mechanical standardization.

In sum, grading standards at the university level are measurable, persistent, and large. Making them explicit is a first step toward understanding how assessment practices shape educational and career trajectories.

## Institutional Review Board Statement

This study is based exclusively on administrative examination records collected and maintained by the university for institutional purposes. The analysis uses anonymized data and does not involve interaction with human subjects, intervention, or experimental manipulation. All analyses were conducted on de-identified records in accordance with applicable institutional regulations governing the use of administrative data for research purposes. No attempt was made to identify individual students or professors, and all reported results are presented in aggregate form. To the best of the author's knowledge, the research complies with relevant ethical standards for the analysis of administrative educational data.

### Informed Consent Statement

The study uses anonymized administrative data and does not involve human subjects as defined by institutional guidelines.

### Data Availability Statement

The dataset used in this study consists of internal administrative records from a large Italian university, covering examination outcomes and associated covariates between 2007 and 2019. All personally identifiable information was removed prior to analysis. Student and professor identifiers were anonymized and replaced with non-informative codes. The dataset contains no names, fiscal identifiers, contact information, or other directly identifying variables. Access to the data was restricted to authorized institutional channels. Results are reported only in aggregate statistical form to prevent re-identification of individuals. The data cannot be made publicly available due to institutional privacy restrictions, but replication materials and code used to construct the severity index and perform the statistical analysis can be provided upon reasonable request, subject to data access approval by the university.

### Conflicts of Interest

The author declares no conflict of interest. The author is affiliated with the institution whose administrative data are analyzed in this study. However, the analysis was conducted independently and without influence from university administrative bodies. The research aims to provide a methodological and empirical contribution to the study of grading standards and does not evaluate or single out individual faculty members.

### Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

### Appendix A. Data, Sample Construction, and Descriptive Evidence

This appendix documents the construction of the analytic sample, the main variables used in the analysis, and descriptive statistics for grades and professor-level severity. It also reports direct evidence on the non-regularity of mixed-effects proportional-odds estimation and descriptive differences in severity across observable dimensions (professor gender and disciplinary area).

**Table A1.** Sample description and variable definitions. Table A1 reports (i) the number of exam attempts in the raw administrative dataset (2007–2019); (ii) the number of passed exam attempts with a recorded numeric grade on the institutional 18–30L scale; (iii) the number of unique professors; (iv) the share of attempts with complete covariate information used in the pooled proportional-odds model; and (v) the final analytic set of professors after imposing the minimum-support condition $n_p \geq 20$. The table also defines each covariate included in the first-stage ordinal model: student gender, off schedule status (indicator for being behind the official study schedule), age at exam date and age squared, exam year dummies, disciplinary area dummies, and the encoded grade outcome where "30 e lode" is mapped to 31. For continuous variables (e.g., age), the table reports mean and standard deviation. For indicators (e.g., off schedule), it reports mean (i.e., proportion equal to 1).

| **Panel A. Sample construction** | |
|---|---|
| Total exam records (raw) | 1,425,371 |
| Complete cases (non-missing grade and covariates) | 1,220,142 |
| Professors (all) | 1642 |
| Professors ($\geq$ 20 graded exams) | 1519 |
| Analysis sample size (complete cases, prof $\geq$20 exams) | 1,220,061 |

**Table A1.** *Cont.*

**Panel B. Variable definitions and descriptive statistics (analysis sample)**
Sample: 1,220,061 exam records; complete covariates; professors with ≥20 graded exams.

| Variable | Definition | Mean | Std. Dev. | N |
|---|---|---|---|---|
| grade_ord | Ordinal exam grade coded from 1 to 14, where 1 ≡ 18 (minimum passing grade), …, 13 ≡ 30, and 14 ≡ 30 e lode (honors above 30) | 8.0218 | 3.9215 | 1,220,061 |
| female_stud | Indicator = 1 if the examined student is female | 0.6074 | 0.4883 | 1,220,061 |
| off_schedule | Indicator = 1 if the examined student is off schedule (i.e., behind the official program schedule / out of nominal time) | 0.1860 | 0.3891 | 1,220,061 |
| age | Professor age (in years) at the time of the exam | 49.6204 | 9.2165 | 1,220,061 |
| age_sq | Professor age squared | 2547.129 | 934.6567 | 1,220,061 |
| year | Exam year fixed effects; years observed: 2007–2019 | — | — | 1,220,061 |
| area | Disciplinary area fixed effects; 14 distinct areas observed | — | — | 1,220,061 |
| cf | Professor identifier (cluster ID); not included as a regressor in the pooled first-stage model | — | — | 1,220,061 |

**Panel A** reports sample construction. "Complete cases" are passing exam attempts with non-missing covariates (student gender, off schedule status, professor age and age squared, exam year, disciplinary area). The analysis sample restricts to professors with ≥20 graded exams (2007–2019), removing only 81 observations. **Panel B** defines variables used in the pooled proportional-odds model (Section 2) and reports descriptive statistics. The dependent variable *voto_ord* encodes grades from 1 (18/30) to 14 (30 e lode).

**Table A2.** Mixed-effects cumulative logit: convergence diagnostics

| Dependent Variable | Fixed Effects Included | Random Intercept | N Obs | N Professors | Convergence Status |
|---|---|---|---|---|---|
| voto_ord | female_stud, off_schedule, age, age_sq, exam year FEs (i.year), disciplinary area FEs (i.area) | Professor identifier cf | 1,220,061 | 988 | No convergence: Stata reports "initial values not feasible" when fitting the random-effects model. |

Notes. We attempt to estimate a mixed-effects proportional-odds cumulative logit model of the form

$$\Pr\big(Y_i \le k \mid X_i, \alpha_{p(i)}\big) = \text{logit}^{-1}\big(\tau_k - X_i^\top \beta - \alpha_{p(i)}\big), \quad k = 1, \ldots, K-1,$$

where $Y_i$ is the ordinal grade grade_ord, $X_i$ includes student gender (female_stud), student off schedule status (off_schedule), professor age and age squared (age, age_sq), exam year fixed effects and disciplinary area fixed effects, and $\alpha_{p(i)}$ is a professor-specific random intercept (professor indexed by cf). The model was estimated via `meologit grade_ord female_stud off_schedule age age_sq i.year i.area || cf:, vce(cluster cf)` on the analysis sample of 1,220,061 graded exam records taught by 988 professors with ≥20 graded exams. Stata's estimation routine first fits the fixed-effects proportional-odds model and obtains a log likelihood of $-3{,}090{,}454.7$. It then attempts to refine starting values for the random intercepts (reporting log likelihood $-3{,}014{,}909.3$ at the initial grid node), but fails to proceed to the full random-effects step, returning the message "initial values not feasible". This non-convergence is consistent with quasi-complete separation at the professor level: some professors assign extremely high (or extremely low) grades almost deterministically, so the likelihood tends to push that professor's intercept toward $\pm\infty$. In such cases, the maximum likelihood estimator for the random-effects cumulative logit does not exist in the interior of the parameter space (Albert and Anderson 1984; Heinze and Schemper 2002). This motivates our alternative identification strategy in Section 2, where we estimate a pooled proportional-odds model without random effects and then construct professor-level severity indices $S_p$ externally.

**Table A3.** Distribution of professor severity $S_p$

| Statistic | Value |
|---|---|
| Mean | 0.211 |
| Standard deviation | 1.517 |
| 1st percentile | −3.246 |
| 5th percentile | −2.259 |
| 10th percentile | −1.745 |
| 25th percentile | −0.835 |
| Median (50th percentile) | 0.202 |
| 75th percentile | 1.227 |
| 90th percentile | 2.227 |
| 95th percentile | 2.697 |
| 99th percentile | 3.748 |
| Minimum | −3.895 |
| Maximum | 5.252 |
| Number of professors | 981 |

Notes. This table reports the distribution of the professor-level severity index $S_p$ across professors with at least 20 graded exam records in the analysis sample (981 professors). For each exam $i$ graded by professor $p$, we compute a residual

$$r_i = y_i - \widehat{E}[y_i \mid X_i],$$

where $y_i$ is the realized exam grade, mapped to the institutional 18–31 scale (with 31 corresponding to "30 e lode"), and $\widehat{E}[y_i \mid X_i]$ is the expected grade from the pooled proportional-odds model in Section 2, conditional on student gender, student off schedule status, professor age and age squared, exam year fixed effects, and disciplinary area fixed effects. The severity index for professor $p$ is then

$$S_p = (1/n_p) \sum_{i \in p} (Y_i - r_i),$$

i.e., the professor-specific mean residual. Negative values indicate systematic under-grading (conditional severity), and positive values indicate systematic over-grading (conditional generosity). The distribution is highly dispersed: the 5th percentile professor assigns on average about 2.3 grade points *below* the conditional benchmark, while the 95th percentile professor assigns on average about 2.7 points *above* it. This dispersion motivates our focus on professor heterogeneity in grading practices.

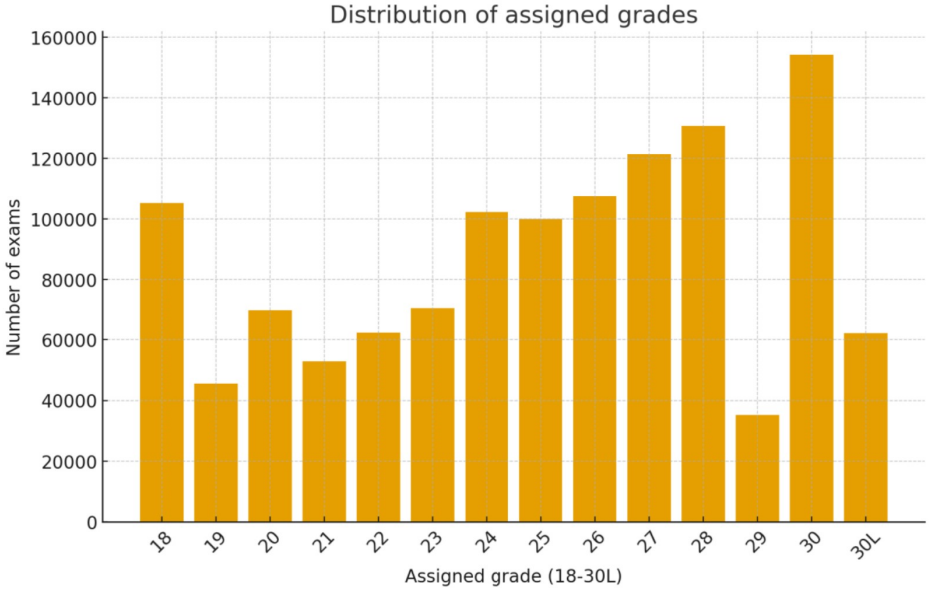


**Figure A1.** Distribution of assigned grades on the institutional 18–30L scale. Empirical distribution of assigned grades in the analysis sample (1,220,061 graded exams taught by professors with ≥20 graded exams and complete covariates). Grades are recorded on a discrete institutional scale from 18 (minimum passing grade) to 30; "30L" denotes 30 e lode, an honors distinction above 30/30.

This figure reports the empirical distribution of assigned grades in the analysis sample (1,220,061 graded exams taught by professors with ≥20 graded exams and complete covariates). Grades are recorded on a discrete institutional scale from 18 (minimum passing grade) to 30; "30L" denotes 30 e lode, an honors distinction above 30/30. The distribution is highly concentrated in the upper part of the scale: grades equal to 27, 28, 29, 30, or 30L account for roughly 41% of all grades. The extreme upper tail is particularly dense: grades of 30 or 30L alone represent about 18% of all recorded passing grades in the sample. Lower passing grades (18–22) are comparatively rare.

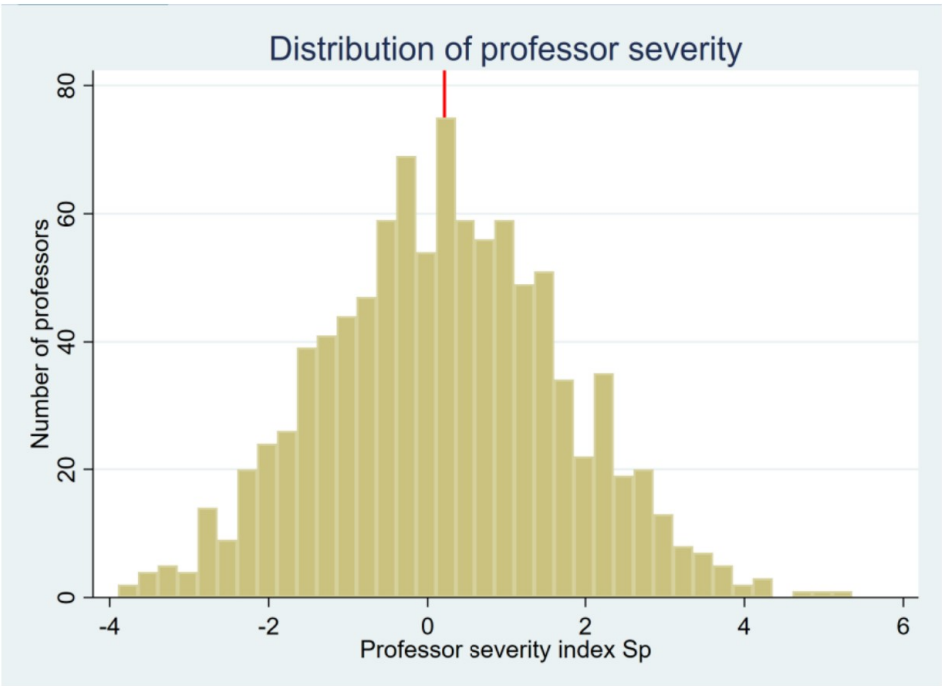


**Figure A2.** Distribution of professor severity $S_p$. Histogram of the professor-level severity index $S_p$ for professors with at least 20 graded exams in the analysis sample. The vertical red line indicates the mean. Negative values correspond to systematic conditional severity; positive values correspond to systematic conditional generosity.

This figure shows the empirical distribution of the professor-level severity index $S_p$ across professors with at least 20 graded exams in the analysis sample. For each exam $i$ graded by professor $p$, we construct a residual

$$r_i = y - \widehat{E}[y_i \mid X_i]\,,$$

where $y_i$ is the realized grade on the institutional 18–30L scale (with "30L" denoting 30 e lode, i.e., honors above 30/30), and $\widehat{E}[y_i \mid X_i]$ is the expected grade from the pooled proportional-odds model in Section 2, conditional on student gender, student off schedule status, professor age and age squared, exam year fixed effects, and disciplinary area fixed effects. We then define

$$S_p = (1/n_p) \sum_{i \in p} r_i$$

as the mean residual for professor $p$. Negative values of $S_p$ indicate systematic under-grading (conditional severity), while positive values indicate systematic over-grading (conditional generosity). The distribution is wide and right-skewed: while many professors cluster near zero, indicating grading close to the conditional benchmark, a non-trivial share of professors exhibit persistent generosity of +2 points or more, and some exceed +4 points relative to model-predicted benchmarks.

## Appendix B. Additional Results and Robustness

**Table A4.** TYear-to-year persistence of professor severity $S_{p,t}$.

| **Dependent Variable: $S_{p,t+1}$** | **(1) OLS** | **(2) Weighted OLS** |
|---|---|---|
| $S_{p,t}$ | 0.805 *** | 0.828 *** |
| | (0.009) | (0.010) |
| Constant | 0.027 ** | −0.019 |
| | (0.013) | (0.015) |
| R-squared | 0.640 | 0.661 |
| N (prof-year pairs) | 8221 | 8221 |
| Clusters (professors) | 929 | 929 |
| Weights | None | Exams in year $t$ |
| Min exams per prof-year ($t$ and $t+1$) | ≥10 | ≥10 |

Standard errors clustered at the professor level. Weighted OLS in column (2) uses analytic weights equal to the number of graded exams assigned by professor $p$ in year $t$. *** $p < 0.01$, ** $p < 0.05$. Notes. For each professor $p$ and year $t$, we define $S_{p,t}$ as the mean residual grading leniency/severity for that professor-year cell. Residuals are computed as observed exam grades minus the model-implied conditional expectation from the pooled proportional-odds specification in Section 2, which conditions on student gender, off schedule status, professor age and age squared, exam year fixed effects, and disciplinary area fixed effects. We restrict to professor-year cells with at least 10 graded exams in year $t$ and at least 10 in $t+1$. Column (1) reports an OLS regression of next-year severity $S_{p,t+1}$ on current-year severity $S_{p,t}$, with standard errors clustered at the professor level. Column (2) re-estimates the same regression weighting each observation by the number of exams graded in year $t$. The estimated persistence parameter $\rho$ is between 0.80 and 0.83, implying that grading harshness/generosity is a stable, professor-specific attribute rather than transitory noise.

**Table A5.** Determinants of professor-level grading severity $S_p$.

| **Dependent Variable: $S_p$ (Mean Conditional Leniency)** | **(1) OLS** | **(2) OLS** | **(3) OLS + Area FE** | **(4) WLS + Area FE** |
|---|---|---|---|---|
| Female professor | −0.053 | −0.047 | −0.034 | 0.041 |
| | (0.101) | (0.101) | (0.106) | (0.125) |
| Pre-promotion share | −1.735 | −1.636 | −0.984 | |
| | (1.308) | (1.273) | (1.189) | |
| Post-promotion share | 0.000 | 0.000 | 0.000 | |
| | [omitted] | [omitted] | [omitted] | |
| Area fixed effects | No | No | Yes | Yes |
| Weights | No | No | No | Exams graded |
| R-squared | 0.0003 | 0.0036 | 0.0133 | 0.0021 |
| N professors | 988 | 988 | 988 | 988 |

Notes. The dependent variable is the professor-level grading severity index $S_p$, defined as the mean conditional residual for professor $p$: the difference between the realized exam grade and the expected grade predicted by the pooled proportional-odds model described in Section 2, averaged across all exams graded by that professor in the analysis sample. "Female professor" is an indicator for professor gender. "Pre-promotion share" is the share of that professor's exams graded during the 12-month window immediately preceding an academic promotion (habilitation/advancement); "Post-promotion share" is defined analogously for the period after promotion, but is identically zero in our estimation sample and is therefore omitted by construction. Columns (3) and (4) include fixed effects for the professor's dominant disciplinary area. Column (4) is weighted by the total number of graded exams for each professor, so that high-volume graders receive more weight. Robust standard errors are reported in parentheses. The estimation sample consists of 988 professors. On average, a professor in our data grades 1235 exams (s.d. ≈ 1096; min = 2; max = 6587). The distribution of grading severity is highly dispersed (mean $S_p$=0.213, s.d. = 1.523, min = −3.919, max = 5.252). Female professors represent roughly 35% of the sample.

**Table A6.** Predictive coverage by student subgroup.

| Subgroup | Mean Coverage (s.e.) | N |
|---|---|---|
| Male students | 0.886 (0.000) | 478,994 |
| Female students | 0.886 (0.000) | 741,067 |
| On-time students | 0.885 (0.000) | 993,189 |
| off schedule students | 0.890 (0.001) | 226,872 |

Notes. For each exam attempt we estimate the full predictive distribution over grades on the institutional 18–30L scale using the pooled proportional-odds model described in Section 2 (covariates: student gender, off schedule status, age and age squared, exam year fixed effects, and disciplinary area fixed effects; professor identifiers excluded). We then construct a 90% conformal-style prediction set for that attempt by sorting all admissible grades by predicted probability and retaining the smallest set of grades whose cumulative probability mass is at least 0.90. "Coverage" is an indicator equal to 1 if the realized grade falls inside that 90% prediction set. The table reports mean coverage rates and their standard errors for four student subgroups: male vs. female students, and on-time vs. off schedule students. Coverage is essentially identical across student gender (0.886 for both male and female students) and very similar across progression status (0.885 for on-time students vs. 0.890 for off schedule students). Standard errors are computed as $\sqrt{p(1-p)/N}$ within subgroup. Sample restricted to complete cases graded by professors with $\geq$ 20 graded exams (1,220,061 exam attempts).

## Appendix C. Simulation and Replication Details

### *Appendix C.1. Simulation Evidence (Section 3)*

We run a Monte Carlo exercise to show two facts: (i) the standard mixed-effects proportional-odds model with professor random intercepts (a cumulative logit with a random intercept for each professor) typically fails to estimate in realistic university grading data because of quasi-complete separation; (ii) our two-stage severity index $S_p$ remains well defined and accurately recovers persistent professor "strictness" or "generosity" on the transcript scale.

Each simulated dataset is built to mimic the institutional setting in this paper. Grades live on the real Italian scale $\{18, \ldots, 30, 30 \text{ e lode}\}$; "30 e lode" is coded as 31. We generate ∼50,000 exam records per replication, assign them to 200 professors with very different grading styles, and force pile-ups at institutional cutpoints (18, 30, 30 e lode) so that some professors are almost always "just pass" and others are almost always "top mark". This reproduces exactly the mass points and extremes we see in the data.

For each replication we estimate:

1. a mixed-effects proportional-odds model with a professor random intercept (`meologit ... || professor:` in Stata); and
2. our two-stage procedure:
   - pooled ordered logit of the grade on student covariates and controls (gender, off schedule status, age and age$^2$, year fixed effects, area fixed effects), explicitly excluding professor ID;
   - expected grade $\widehat{E}_i$ from that pooled model for each exam $i$;
   - professor severity $\widehat{S}_p = \text{average}_i(Y_i - \widehat{E}_i)$ over all exams graded by professor $p$.

Result: in the vast majority of replications that look like our real university data, the mixed-effects cumulative logit does not converge and Stata returns "initial values not feasible", exactly as in Appendix A, Table A2. This is the classical non-existence of a finite MLE under (quasi-)complete separation in logistic/ordinal likelihoods [8,9]. By contrast, $\widehat{S}_p$ is always finite, tracks the true underlying strictness/generosity almost one-for-one, and preserves the ranking of professors (Spearman correlation above 0.9 in our simulations). In short: the "textbook" hierarchical ordinal model is not usable in practice here; the two-stage $S_p$ is.

### *Appendix C.2. Replication Steps (Empirical Data)*

This subsection documents how we compute $S_p$ in the administrative data used in the paper. Variable names match those in Appendix A.

1. **Sample restriction.**

```
• keep if !missing(grade, female_stud, off_schedule, age, age_sq, area,
  year, cf)
• bysort cf: gen n_prof = _N
• keep if n_prof >= 20
```

2. **First-stage pooled ordinal model (no professor ID).**

Fit an ordered logit (proportional-odds) of the grade on student and contextual covariates:

```
• ologit grade female_stud off_schedule age age_sq i.year i.area
```

From this model, obtain for each exam $i$ the full predicted probability over all admissible grades, and from that the expected grade $\widehat{E}_i$ on the 18–30L scale:

```
• tempvar Eh
• gen Eh = 0
• levelsof grade, local(glist)
• foreach g of local glist {
•   predict pr_`g' if e(sample), outcome(#`g')
•   replace Eh = Eh + (`g') * pr_`g'
• }
• gen resid = grade - Eh
• collapse (mean) Sp = resid (count) nobs = resid, by(cf)
```

3. **Professor-level severity index.**

Compute the residual for each exam $i$: actual grade minus expected grade; then average by professor `cf`:

```
• gen resid = grade - `Eh'
• collapse (mean) Sp = resid (count) nobs = resid, by(cf)
• label var Sp "Professor severity (S_p)"
• label var nobs "Number of graded exams"
```

$$S_p < 0 : \text{severity (below benchmark)}$$
$$S_p > 0 : \text{generosity (above benchmark)}$$

Because $\widehat{E}_i$ already conditions on observed student composition, year, and area, $S_p$ is not "easy exams vs. hard exams" but a conditional grading standard on the same transcript scale students see.

This is the object whose cross-sectional dispersion we report in Appendix A, Table A3; whose year-to-year persistence we study in Appendix B, Table B1; and which we analyze in Sections 4 and 5.

## References


1. Rockoff, J.E. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *Am. Econ. Rev.* **2004**, *94*, 247–252.
2. Rivkin, S.G.; Hanushek, E.A.; Kain, J.F. Teachers, Schools, and Academic Achievement. *Econometrica* **2005**, *73*, 417–458.
3. Kane, T.J.; Staiger, D.O. *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*; NBER Working Paper No. 14607; NBER: Cambridge, MA, USA, 2008.
4. Chetty, R.; Friedman, J.N.; Rockoff, J.E. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *Am. Econ. Rev.* **2014**, *104*, 2593–2632.
5. McCullagh, P. Regression Models for Ordinal Data. *J. R. Stat. Soc. Ser. B Methodol.* **1980**, *42*, 109–142.
6. Agresti, A. *Analysis of Ordinal Categorical Data*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2010.
7. Williams, R. Generalized Ordered Logit/Partial Proportional Odds Models for Ordinal Dependent Variables. *Stata J.* **2006**, *6*, 58–82.
8. Albert, A.; Anderson, J.A. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika* **1984**, *71*, 1–10.
9. Heinze, G.; Schemper, M. A Solution to the Problem of Separation in Logistic Regression. *Stat. Med.* **2002**, *21*, 2409–2419.
10. Lei, J.; Wasserman, L. Distribution-Free Prediction Bands for Non-Parametric Regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76*, 71–96.
11. Romano, Y.; Patterson, E.; Candès, E. Conformalized Quantile Regression. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.