Are Course Grades Comparable Across Classrooms? Instructor-by-Term Heterogeneity in Undergraduate Mathematics Assessment

Heriberto Espino-Montelongo
Universidad de las Américas Puebla

September 2026

Abstract

Course grades are routinely combined across instructors and academic periods in institutional reporting and educational research, although the same numerical scale may be produced within substantially different assessment contexts. We examine this comparability problem using administrative records from undergraduate mathematics at Universidad de las Américas Puebla. The analytical unit is the classroom, defined as an instructor teaching the same course in the same academic period. A retained exploratory analysis contains 26,140 student–classroom observations, 77 instructors and 769 classrooms from 2019–2025, with a median classroom size of 30. That analysis shows visible differences in the location, spread and shape of grade distributions across instructors and period-to-period variation for the same instructor; an exploratory distribution-distance analysis also identified nontrivial clustering of grading profiles. Rather than interpreting these differences as pure instructor severity, the present paper treats them as context-associated variation that may combine assessment practice, assessment design, student composition and realized performance. The confirmatory specification therefore conditions comparisons on course and period, separates numeric awarded grades from nonnumeric adverse outcomes, evaluates distributional distances and temporal stability, and quantifies the reclassification produced when raw grades are replaced by classroom-relative standardization. The central measurement claim is deliberately limited: a common numerical grade scale does not by itself guarantee empirical exchangeability across classrooms. Classroom-relative standardization can reduce one source of contextual incomparability, but it changes the estimand from recorded achievement on an absolute institutional scale to relative position within a local distribution and should not be interpreted as a universal measure of mathematical proficiency.

Keywords: grading; assessment; grade comparability; higher education; mathematics; classroom effects; standardization

1 Introduction

Grades occupy an unusual position in higher education because they are simultaneously records of individual achievement, inputs to progression and award decisions, and convenient quantitative outcomes for institutional research. Their apparent simplicity encourages aggregation: a numerical grade can be averaged across courses, compared across students and inserted into statistical models without requiring a new measurement instrument. Yet that convenience depends on a strong assumption that is often left implicit, namely that the scale is sufficiently commensurable across the local contexts in which grades are produced. A grade of 8.5 on a 0–10 scale has the same numerical value wherever it appears, but numerical identity does not establish that it represents an equivalent position in an assessment distribution or that it was generated under equivalent criteria.

This distinction between a common scale and a common measurement meaning is well established in assessment scholarship. Sadler (2009) frames grade integrity in terms of the relationship between awarded grades and the quality, breadth and depth of the evidence on which they are based, while research on marking practices shows that academic judgement is not reducible to the mechanical application of written criteria (Bloxham, Boyd, et al. 2011; Bloxham, den-Outer, et al. 2016). Broader reviews likewise show that grades can incorporate multiple cognitive and non-cognitive elements rather than a single latent construct (Brookhart et al. 2016), and empirical work across disciplines has documented differences in both grading criteria and assessment cultures (Lipnevich et al. 2020; Ylonen et al. 2018). These findings do not imply that grades are useless measures; indeed,

aggregate college grades can be highly reliable (Beatty et al. 2015). They do imply, however, that reliability and cross-context comparability are different questions.

The present study examines that distinction in undergraduate mathematics at Universidad de las Américas Puebla (UDLAP). It emerged from a broader institutional study in which course grades were used to evaluate students' relative academic performance while studying use of the Centro de Aprendizaje de Matemáticas (CMAT). During that work, pooling raw grades across instructors and academic periods proved difficult to justify because classroom distributions differed visibly in location, spread and shape. The solution used in the performance study was to standardise outcomes within classroom, where a classroom means the same instructor teaching the same course in the same academic period. Paper 3 turns that methodological adjustment into the object of study rather than treating it as a preprocessing detail.

Our research questions are therefore about measurement rather than instructor ranking. First, how much do classroom grade distributions vary when comparisons are restricted to the same course and academic period? Second, how stable are an instructor's course-specific grade distributions across periods? Third, how much does student ranking or classification change when a raw course grade is replaced by a classroom-relative standardised outcome? Fourth, how sensitive are conclusions about comparability to classroom-size restrictions and to the treatment of nonnumeric adverse academic outcomes?

The contribution is twofold. Empirically, the study quantifies grading-context heterogeneity in a multi-year mathematics setting in which the numerical grade scale is nominally common. Conceptually, it distinguishes three objects that are often conflated: the recorded course grade, the student's position relative to classmates evaluated in the same local context, and latent mathematical proficiency. The first is observed, the second can be constructed, and the third is not identified by administrative course grades alone. Maintaining those distinctions allows grade standardisation to be evaluated as a contextual measurement choice without presenting it as a correction that reveals a student's "true" ability.

2 Grades as contextual measures in higher education

A grade can be dependable for one purpose without being exchangeable for every other purpose. Beatty et al. (2015), for example, report high reliability for first-year and overall college grade-point averages in very large multi-institutional data, which supports the use of accumulated grades as stable summaries of academic performance. That result does not establish that every component course grade has the same meaning across instructors or assessment contexts, because aggregation can be reliable even when local components contain systematic contextual variation. For the present study, comparability therefore concerns whether observations expressed on the same numerical scale can reasonably be treated as if instructor-course-period context were irrelevant.

Assessment research provides several reasons to doubt that assumption. In a think-aloud study of university grading, Bloxham, Boyd, et al. (2011) found that assessors frequently made holistic judgements and did not simply apply written criteria analytically, while Bloxham, den-Outer, et al. (2016) documented variation in the criteria noticed, ranked and scored by experienced assessors. These studies concern judgement processes rather than administrative grade distributions, but they establish a plausible mechanism through which local marking practice can generate heterogeneity even when formal standards exist. Lipnevich et al. (2020) similarly found substantial variation in the criteria used to construct course grades across American college and university syllabi, including differences in the balance between product and process criteria and in the use of examinations.

At a broader institutional level, Ylonen et al. (2018) show that mark distributions vary across disciplinary assessment cultures and caution against treating student marks as straightforwardly comparable measures of learning gain. Their argument is particularly relevant here because the same problem can occur at a finer scale: even within one disciplinary family, different courses, instructors and periods can constitute different assessment environments. Paper 3 does not presume that all observed distributional differences are undesirable or that uniform distributions would be evidence of higher assessment quality. Instead, heterogeneity matters because downstream analyses often assume away the context that produced the grade.

This distinction also constrains interpretation of standardisation. If student i receives raw grade G_{ic} in classroom c, the classroom-relative score in Equation (1) is

$$Z_{ic} = \frac{G_{ic} - \bar{G}_c}{s_c}, \tag{1}$$

Table 1: Retained exploratory evidence from the historical grading analysis.

Quantity	Historical value
Student–classroom observations	26,140
Instructors	77
Instructor–course–period classrooms	769
Median students per classroom	30
Students appearing in multiple classrooms	65.2%
Nonnumeric final-grade observations	3,217 (12.3%)
Passing share among observed numeric grades	90.57%
Passing share after historical adverse-outcome imputation	79.67%
Exploratory K-Medoids solution with maximum Silhouette Score	$k = 3$

where $\bar{G}_c$ and s_c are the classroom mean and standard deviation. The transformation removes classroom-specific location and scale from the outcome by construction, making Z_{ic} a statement about relative position rather than absolute recorded attainment. A positive value means that the student is above the classroom mean; it does not mean that the student has mastered an externally fixed amount of mathematics. Consequently, raw grades and classroom-relative scores answer different questions, and the empirical issue is not which one is universally correct but how much substantive inference depends on choosing between them.

3 Institutional context and data

UDLAP records final course outcomes on a 0–10 numerical scale, with 7.5 as the passing threshold in the study data. The broader administrative extract covers undergraduate mathematics enrolments over multiple academic periods between 2019 and 2025. A student can appear repeatedly because degree programmes contain multiple mathematics courses, and the same instructor can teach the same or different courses in more than one period. This repeated structure is analytically useful because it provides both within-student and within-instructor-course information, but it also makes a simple independent-observation model inappropriate for the final variance decomposition.

The classroom is defined as an instructor teaching the same course in the same academic period, independently of section when multiple sections share that instructor-course-period combination. This definition was adopted in the earlier project because the period itself can change the assessment context: the historical analysis displays changes in mean grades for the same instructor over time, so instructor identity alone is not treated as a stable grading environment. Conversely, combining all instructors within a course-period would erase precisely the variation that Paper 3 seeks to measure.

Table 1 records the quantities that can be verified from the retained historical report. They describe the old broad analytical workspace and are not yet the final Paper 3 sample, because the confirmatory rerun will apply a more explicit attempt/revalidation rule and a minimum numeric classroom size.

Nonnumeric outcomes require particular care because ‘BV’, ‘RT’ and ‘BA’ are not neutral missing values in the institutional context; they represent adverse academic states such as withdrawal or temporary withdrawal. For a paper about student academic performance, retaining these events through an explicit adverse-outcome construction can be preferable to complete-case deletion. For a paper about the distribution of grades actually awarded by instructors, however, assigning these states simulated numeric grades changes the construct being studied. We therefore use observed numeric final grades as the primary grading-distribution outcome, describe adverse-state composition separately, and reserve the project’s preserved lower-tail imputation procedure for sensitivity analysis.

Administrative equivalencies, revalidations and other non-attempt records are also excluded from the construct of a newly awarded course grade. This is important because a course can reappear under a later programme or degree record without representing a second independent grading event. The final Paper 3 cohort will use the canonical attempt/revalidation logic maintained in the shared analysis library rather than relying on simple row-level deduplication.

4 Analytical strategy

4.1 Classroom summaries and comparability strata

The first analytical layer describes each eligible classroom using the number of numeric grades, mean, standard deviation, median, interquartile range and numeric-grade pass rate, while adverse outcomes are tabulated in parallel. The primary minimum classroom size is 20 numeric awarded grades; thresholds of 10 and 30 form sensitivity analyses. The threshold is a distribution-estimation rule rather than a substantive exclusion criterion, because smaller classes are educationally meaningful but provide less stable empirical distribution estimates.

Heterogeneity is then evaluated primarily within course by academic-period strata. This restriction matters because a difference between, for example, an introductory course and a later calculus course is not evidence of instructor grading heterogeneity, nor is a change between academic years automatically attributable to an instructor. Conditioning on the nominal course and period creates a more defensible comparison set, although it still does not eliminate non-random sorting of students across instructors.

4.2 Distributional distance

Means and pass rates capture only parts of a grade distribution, so the confirmatory analysis will compare classroom empirical distributions using two complementary distances. The two-sample Kolmogorov–Smirnov distance records the maximum vertical separation between empirical cumulative distribution functions and is sensitive to differences in distributional shape, while the first Wasserstein distance summarises the amount of movement on the grade scale required to align two distributions. For each course-period stratum containing at least two eligible classrooms, we will summarise the median, interquartile range and upper tail of pairwise distances rather than treating a single pair as representative.

The historical analysis also clustered professor-level distributions using KS distances and K-Medoids, with the highest reported Silhouette Score at three clusters. That result is useful as exploratory evidence that grade distributions contained recognizable structure, but it is not a confirmatory taxonomy. The new analysis will condition clustering on course-period comparability and will not label clusters as “strict”, “lenient” or “high-quality”, because distributional shape alone cannot identify the mechanism that generated it.

4.3 Context-associated variance

The third layer will estimate a multilevel or cross-classified decomposition of numeric grades, beginning with course-by-period fixed effects and accounting for repeated students and classroom or instructor context when the observed overlap supports identification. The purpose is to quantify how much residual variation is associated with local classroom context after nominal course-period differences are absorbed. The classroom component will not be called an instructor causal effect: students are not randomly assigned to instructors, and assessment design, cohort composition and realized performance can all contribute to the same component.

If the preferred cross-classified model is unstable, we will use a simpler course-period fixed-effect model and report classroom residual distributions with cluster-robust uncertainty rather than forcing a poorly identified random-effects structure. This fallback is methodologically preferable to obtaining an apparently precise variance partition from a model unsupported by the data structure.

4.4 Temporal stability

For instructor-course combinations observed in at least three eligible periods, we will examine stability in mean grade, within-classroom dispersion, pass rate and full-distribution distance. This converts the historical visual observation of period-to-period movement into a formal question: to what extent does an instructor-course grading profile persist when the academic period changes? A stable instructor-course component would support the idea that part of the distribution is persistent, whereas large within-instructor-course shifts would show why an instructor-only adjustment is insufficient.

4.5 Consequences of classroom-relative standardisation

Finally, we will quantify how much a student’s position changes when analysis moves from the raw grade to Z_{ic}. Planned diagnostics include the Spearman association between the two measures, movement in pooled

percentile rank, quintile or decile reclassification, and examples in which similar raw grades correspond to materially different relative positions. These diagnostics are direct measurement consequences: they do not require a causal claim about instructors and make clear what information is changed by standardisation.

5 Results

5.1 Exploratory evidence of heterogeneous classroom contexts

The historical analysis already provides a strong reason not to assume homogeneous grading distributions. Across the retained 2019–2025 workspace, 26,140 student–classroom observations were distributed among 77 instructors and 769 instructor-course-period classrooms, with a median of 30 students per classroom. Because 65.2% of students appeared in more than one classroom, the data represent overlapping academic trajectories rather than isolated cross-sections, creating the possibility of comparing local grading contexts while recognising repeated observations.

Plots retained in the historical report show substantial differences across instructors in the location, dispersion and shape of final-grade distributions. Some distributions concentrate more heavily near the upper end of the scale, others are broader, and others place more observed mass near the passing boundary. Those plots establish heterogeneity descriptively, but they do not establish that instructors alone produced it. In particular, the broad professor-level figures pooled different courses and periods, so the dedicated Paper 3 rerun must determine how much of the apparent separation survives within course-period comparisons.

5.2 Outcome construction materially changes the lower tail

The historical data contain 3,217 nonnumeric final-grade observations, corresponding to 12.3% of student–classroom records. This is too large a component to disappear silently through complete-case analysis, but it is also too substantively distinct from an awarded grade to be treated as numeric without explicit qualification. The old report illustrates the consequence: the aggregate passing share is 90.57% when calculated from observed numeric grades, whereas the historical adverse-outcome imputation produces 79.67% passing after nonnumeric adverse states are placed into the lower grade range.

The difference does not show that either percentage is the unique "true" pass rate; it shows that the construct changes when adverse statuses are translated into simulated grades. For Paper 3, the primary distributional estimand is therefore the distribution of awarded numeric grades, while adverse-status incidence becomes a parallel classroom characteristic. The augmented outcome remains valuable as a sensitivity analysis because a classroom with many withdrawals may look unusually successful if only completed numeric grades are inspected.

5.3 Exploratory distributional clusters

The historical report computed pairwise KS distances between professor grade distributions and used K-Medoids to search for groups of similar distributional profiles. Its reported Silhouette Score was maximal at $k = 3$, and the corresponding plots showed three sets of distributions that were more similar within groups than across the full collection. This result is evidence that heterogeneity was not visually random, but the original clustering combined professor records across academic contexts and should therefore be interpreted only as a hypothesis-generating result.

A confirmatory cluster analysis, if retained in the final paper, will be secondary to direct distance summaries and will operate within explicitly comparable strata. The paper does not require a discrete cluster solution to establish contextual heterogeneity; continuous variation in classroom distributions is itself substantively meaningful.

5.4 Period belongs in the grading context

The historical report also follows the mean grade of the same instructor across periods and shows visible temporal movement. Although an illustrative series cannot estimate population-level stability, it demonstrates why the analysis should not define grading context by instructor alone. If an instructor's distribution moves as course cohort, assessment design or period conditions change, pooling that instructor's records across years can create a reference distribution that no individual classroom actually experienced.

5.5 How much does standardisation change student classification?

The historical project used classroom-standardised grades because raw grades appeared insufficiently comparable across local contexts, but it did not quantify the measurement consequence of that choice directly. Paper 3 will therefore report how many observations change pooled percentile rank or quintile/decile classification when raw grade is replaced by classroom-relative Z. This is the most direct link between the heterogeneity diagnosis and downstream educational analysis: if reclassification is negligible, contextual differences may have little practical consequence for many uses; if it is substantial, raw-grade pooling can alter which students appear relatively high or low performing.

6 Discussion

The retained evidence supports a measurement problem that is narrower than a claim about instructor quality but more consequential than a cosmetic difference in grade distributions. Undergraduate mathematics grades are recorded on a common numerical scale, yet the historical data show that the distributions producing those numbers differ across instructor contexts and can change over time. Assessment scholarship provides several plausible reasons for this pattern, including variation in criteria, holistic academic judgement and local assessment cultures (Bloxham, Boyd, et al. 2011; Bloxham, den-Outer, et al. 2016; Lipnevich et al. 2020; Ylonen et al. 2018). The administrative data cannot distinguish those mechanisms from student sorting or differences in realized achievement, so the defensible empirical object is context-associated grade variation rather than instructor severity.

This distinction matters because educational research frequently uses grades as outcomes without modelling the environments in which they were produced. A regression on raw grades can therefore combine at least two sources of variation: differences in student performance and differences associated with classroom assessment context. Restricting comparisons to course-period strata, including classroom effects or using classroom-relative standardisation can reduce the second component, but each strategy changes the target of inference. In particular, Z_{ic} answers where a student sits within the local classroom distribution, not what absolute level of mathematical knowledge the student possesses.

The difference between reliability and comparability is also important. High reliability of accumulated college grades (Beatty et al. 2015) is compatible with local grading heterogeneity because a stable aggregate can be built from components whose contexts differ systematically. Conversely, finding heterogeneous classroom distributions does not imply that grades are unreliable for progression decisions or useless as institutional records. Paper 3 instead asks whether raw course grades are exchangeable enough for analytical comparisons that ignore instructor-course-period context.

For assessment practice, the findings could motivate attention to coordination, moderation and shared standards, but the paper should not infer that uniform distributions are a desirable policy target. Mechanical equalisation of means could conceal genuine cohort differences or distort criterion-referenced assessment. Research on grading variation suggests that moderation and shared interpretation can reduce some inconsistency while recognising the role of professional judgement (Bloxham, den-Outer, et al. 2016; Hunter and Docherty 2011). The more immediate institutional implication is analytical transparency: whenever course grades are pooled, researchers should state whether and how grading context is handled and what interpretation remains after adjustment.

7 Limitations and remaining analyses

Several limitations are structural. First, students are not randomly allocated to instructors, so between-classroom distributional differences cannot be interpreted as causal grading effects. Second, the administrative data record final outcomes but not the full assessment design, weights of examinations and assignments, marking rubrics, moderation processes or changes in course content; those unobserved features are part of the mechanism rather than statistical noise. Third, classroom standardisation removes location and scale by construction and can hide absolute differences in mastery if those differences are real. Fourth, small classrooms yield unstable distribution estimates, making the minimum-size sensitivity essential. Fifth, nonnumeric adverse outcomes are substantively informative but do not belong naturally on the same measurement scale as awarded numeric grades, so any augmented outcome requires explicit modelling assumptions.

The present draft also has a development limitation: its quantitative statements come from the retained broad 'proyecto_visitas' report rather than a dedicated Paper 3 rerun. The old clustering pooled professor distributions more broadly than the confirmatory design now permits, and the old primary outcome treatment was developed for a student-performance question rather than a grading-comparability question. Those exploratory results are included to preserve provenance and motivate the study, not to freeze the final estimates.

8 Conclusion

A common numerical grading scale does not guarantee that course grades are empirically exchangeable across the classrooms in which they are produced. The retained undergraduate mathematics data show sufficient variation across instructor and period contexts to make that assumption worth testing directly, while the sensitivity of the lower tail to nonnumeric outcome construction shows that even the object called a "grade distribution" requires careful definition. The appropriate response is not to replace raw grades automatically with a single adjusted score, but to align the measurement choice with the research question: raw grades preserve the institution's recorded achievement scale, whereas classroom-standardised grades preserve local relative position. Quantifying the difference between those perspectives is necessary whenever grades are used to compare students across heterogeneous assessment contexts.

References

Beatty, A. S., P. T. Walmsley, P. R. Sackett, N. R. Kuncel, and A. J. Koch (2015). "The Reliability of College Grades". In: *Educational Measurement: Issues and Practice* 34.4, pp. 31–40. DOI: 10.1111/emip.12096.

Bloxham, S., P. Boyd, and S. Orr (2011). "Mark My Words: The Role of Assessment Criteria in UK Higher Education Grading Practices". In: *Studies in Higher Education* 36.6, pp. 655–670. DOI: 10.1080/03075071003777716.

Bloxham, S., B. den-Outer, J. Hudson, and M. Price (2016). "Let's Stop the Pretence of Consistent Marking: Exploring the Multiple Limitations of Assessment Criteria". In: *Assessment & Evaluation in Higher Education* 41.3, pp. 466–481. DOI: 10.1080/02602938.2015.1024607.

Brookhart, S. M., T. R. Guskey, A. J. Bowers, J. H. McMillan, J. K. Smith, L. F. Smith, M. T. Stevens, and M. E. Welsh (2016). "A Century of Grading Research: Meaning and Value in the Most Common Educational Measure". In: *Review of Educational Research* 86.4, pp. 803–848. DOI: 10.3102/0034654316672069.

Hunter, K. and P. Docherty (2011). "Reducing Variation in the Assessment of Student Writing". In: *Assessment & Evaluation in Higher Education* 36.1, pp. 109–124. DOI: 10.1080/02602930903215842.

Lipnevich, A. A., T. R. Guskey, D. M. Murano, and J. K. Smith (2020). "What Do Grades Mean? Variation in Grading Criteria in American College and University Courses". In: *Assessment in Education: Principles, Policy & Practice* 27.5, pp. 480–500. DOI: 10.1080/0969594X.2020.1799190.

Sadler, D. R. (2009). "Grade Integrity and the Representation of Academic Achievement". In: *Studies in Higher Education* 34.7, pp. 807–826. DOI: 10.1080/03075070802706553.

Ylonen, A., H. Gillespie, and A. Green (2018). "Disciplinary Differences and Other Variations in Assessment Cultures in Higher Education: Exploring Variability and Inconsistencies in One University in England". In: *Assessment & Evaluation in Higher Education* 43.6, pp. 1009–1017. DOI: 10.1080/02602938.2018.1425369.