

Discrimination or Leniency? A Statistical and Machine-Learning Analysis of Internal versus External Assessment across Institution Types in a District of Nepal

Krishna Chandra Paudel

Lecturer, Department of Mathematics Education, Mahendra Ratna Campus, Tahachal
ORCID: 0000-0003-3381-0248 | Email: kcpaudel@trinitycollege.edu.np

Shree Prasad Ghimire*

Lecturer, Department of Mathematics Education, Mahendra Ratna Campus, Tribhuvan University, Nepal

ORCID: 0009-0002-9697-2949 | *Corresponding author: shreeprasad.ghimire@mrc.tu.edu.np

Manil Vaidhya

Department of Software Engineering, Nepal College of Information Technology (NCIT), Pokhara University, Lalitpur, Nepal

ORCID: 0009-0003-1528-5387 | Email: manil.baidhya@ncit.edu.np

Abstract

The basic level examination for grade 8 has implemented an internal assessment system in Nepal's basic school curriculum. Schools combine internal marks with board-external marks in many high-stakes examination systems, but the question remains whether internal marks serve as a valid, consistent substitute for external attainment. Whether the validity depends on the type of institution awarding the marks remains largely unexamined in Nepal, where prior scholarship on the topic is predominantly qualitative. Using district-level records for 2788 students drawn from three types of institutions (Institutional/private, Community/public, and Alternative), this study characterizes the internal-external relationship by analyzing the agreement using statistics, ceiling-aware regression, and machine learning.

Concordance is weak, with Lin's concordance correlation coefficient not exceeding 0.15 in any type, and the form of the discrepancy differs systematically by the type of institution. Institutional marks are compressed near the scale of the ceiling and fail to discriminate attainment (area under the ROC curve, $AUC = 0.62$; a censoring-corrected Tobit slope of 0.20; a 90th-percentile with quantile slope of 0.00), whereas Community marks discriminate attainment considerably better ($AUC = 0.78$) but sit systematically above external performance. The study found that invalid internal grading has two distinct aspects, such as a discrimination failure and a level (leniency) failure, that implicate different types of institutions and that the discrimination finding is harmful to the quality of students. The results imply that internal marks are not interchangeable across institution types and that any high-stakes collection of internal and external marks should apply type moderation.

Keywords: Concordance correlation, concurrent validity, external examination, grade inflation, internal assessment, machine learning

Introduction

The evaluation system of Nepal is divided into two systems: Internal and External evaluations system. The internal examination covers 50% of the total scores, whereas the external evaluation system covers 50% of the marks in the Basic level. The respective subject teachers of their own school provide the internal evaluation marks on the basis of the overall performance of students, including attendance, classroom participation, internal examination, assignments, project work/practical work, including weekly/monthly/ terminal examinations as decided by the school system. Nowadays, discrepancies have been found between the scores of students provided by the subject teachers as internal evaluations and the scores obtained from the board examinations. Therefore, questions have been raised about the validity of the internal examination to establish the internal consistency in the scores obtained by students.

A student reported achievement by blending two distinct sources of information: marks awarded internally by the school based on continuous, classroom-embedded assessment, and marks awarded externally by a national examination board based on a teacher-made test as a terminal test. The interpretive question that arises from this arrangement is whether the internal component measures the same underlying construct as provided by the external examinations or not. It is a matter of determining the concurrent validity (American Educational Research Association et al., 2014). The question has been raised in the time of urgency of the COVID-19 disruption, when several examination boards temporarily replaced external examinations with school-based internal assessment, a shift that raised widespread concern about grade increase and about the consistency of grading system from internal and external standards across schools (National Examinations Board, 2021).

These two types of evaluation system practices in Nepal are informative and currently under-studied in the context of Nepal. 1) The results of the Secondary Education Examination explicitly combine external and internal components into a single reported outcome, so that the two measures co-exist at the level of the individual student and can be compared directly. 2) Nepali schools are classified under a distinctive category. Nepali scholarship on internal assessment is, moreover, predominantly qualitative in nature, leaving the quantitative concordance question - how strongly, and in what form, internal and external marks actually relate to one another is largely open.

Evaluation System at the Basic Level in Nepal

The main purpose of the internal assessment system, particularly at the basic level, is intended to monitor students' learning progress, identify learning difficulties, support their overall development, and prepare them for the next level of study. The system combines formative and summative assessment. Internal evaluation is grounded essentially in the formative approach. It is conducted continuously through classroom activities, homework, project work, practical work, class tests, and other learning activities. It allows students to recognise their own level of learning and provides teachers with information about students' strengths and weaknesses, enabling adjustments to teaching and the provision of remedial support. The School Education Sector Plan

(SESP) recognizes formative assessment as a core element of the pedagogical process. The Basic Level Examination (BLE), administered at grade 8, has served as the main summative assessment for evaluating students' achievement at the end of basic education. The two components are therefore intended to serve complementary purposes: internal evaluation underlines continuous learning and feedback, while external evaluation provides a standardized measure of achievement.

This study gives a fixed, municipal-level dataset of internal and external marks linked to two questions: a) How strong, and of what form, of their relationship between internal and external assessments? b) Does the consistency of internal marking differ systematically by institution type? The contribution of the study is threefold: 1) it offers a concordance-and-agreement treatment of the internal-external relationship, rather than the correlation and mean difference approach common in this literature; 2) it decomposes "invalid grading" into two empirically separable facets: discrimination and level bias that implicate different institution types; and 3) it demonstrates the discrimination aspect is strong to the student-quality confound that weakens the kindness feature, thereby sharpening what can, and cannot, be concluded from marks-only administrative data.

Statement of the Problems

Internal assessment covers an increasingly consequential place in Basic-Level examination in Nepal. The assumption that internal marks are a trustworthy, interchangeable proxy for external attainment has rarely been tested with quantitative rigour. Therefore, the authors were motivated to conduct this study. In this context, three types of issues have been raised about the internal assessment system at the basic level, particularly in grade 8. Firstly, it is not established for the Nepali context whether school-awarded marks track the same underlying construct as board-administered marks, or the degree to which they diverge. Where such divergence exists, it challenges the fairness of any high-stakes decision promoting the selection or certification of combinations of the two components, even if they were equivalent.

The second problem was related to the institution type, as three main categories of schools, such as Community, Institutional (Private), and Alternative schools, operate under markedly different incentive structures, resourcing, and accountability pressures. Prior grade-inflation research was conducted almost entirely in two-way private/public systems abroad. The prior literature offers no direct guidance on whether, and how, concordance differs across this divide, specifically in the environment of the schools of Nepal. Third, the methodological problem: the small body of existing Nepali work on internal assessment is predominantly qualitative, relying on teacher and student perceptions rather than large-sample quantitative analysis of actual marks. Where quantitative comparisons of internal and external marks have been attempted elsewhere, they typically rely on simple correlation and mean-difference statistics, which conflate two conceptually distinct problems whether marks order students correctly (discrimination) and whether marks are systematically too high or too low (level bias) into a single, uninformative number. A ceiling-compressed set of internal marks, for instance, can show low variance and yet fail entirely to discriminate between strong and weak students, a pattern that a variance-based or the analysis of correlations would mislabel as consistent in grading system. These problems mean that policymakers and examination authorities lack an evidence base for deciding whether, and

how, internal marks should be weighted, scaled, or moderated when combined with external marks. A gap in this study was designed to close the district-level administrative data agreement statistics and machine learning.

Objectives of the Study

The general objective of the study is to examine the statistical relationship between internally and externally awarded assessment marks across institution types in a district of Nepal, using a combination of classical statistical and machine-learning methods. Specifically, this study has the following objectives:

1. To determine the strength and functional form (linear, monotone, or non-linear) of the relationship between internal and external marks.
2. To examine whether the concordance and consistency of internal marking differs systematically across Institutional, Community, and Alternative schools.
3. To draw policy implications for the moderation and aggregation of internal and external scores for high-level decisions.

Research Questions

To fulfill the above objectives, this study is guided by three principal research questions:

1. How strong, and of what functional form (linear, monotone, or non-linear), is the relationship between internal and external marks?
2. Does the concordance and consistency of internal marking differ systematically by institution type?
3. Can inconsistencies in the evaluations of students be decomposed into separable aspects of discrimination?

Review of Literature

In this section, we reviewed the different theoretical and empirical alignments of assessment systems in different institutions.

Theoretical Literature

The following section deals with the theoretical alignment of the data. It measures the concurrence between the internal and external marks of the students obtained in their Grade 8 BLE examination.

Concurrent validity

The theoretical literature associated with this study for comparing internal and external marks is the concept of concurrent validity. The degree to which one measure correlates with another measure of the same construct administered at approximately the same time (American Educational Research Association et al., 2014). Under an assumption of strong concurrent validity, internal and external marks should be interchangeable, with high correlation, close numerical agreement, and stable rank ordering of students. The Standards for Educational and Psychological Testing frame validity not as a fixed property of a test but as a property of the inferences drawn

from scores in a particular use. This framing is directly relevant to internal marks that are frequently aggregated with external marks for high-stakes purposes such as promotion, certification, and selection.

Two Aspects of Consistency: Discrimination and Level Bias

A central theoretical alignment in this study is to separate two operational meanings of consistency between internal scores that can diverge from one another. Discrimination (relative validity) is the degree to which internal marks predict the external mark of the students. It is a group based, rank-based property that captured by correlation, concordance, quantile slopes, and use of statistics. Level bias (leniency) is the degree to which internal marks sit systematically above (or below) external performance. It is a between-group, location-based property captured by the mean internal-external gap, like model-based inflation scores. A school can score well on one aspect and poorly on the other. Uniformly high internal marks clustered near the scale ceiling exhibit low variance with poor discrimination, because a small constant mark cannot distinguish strong students from a weak one. Treating low variance as evidence of consistency therefore conflates two conceptually distinct psychometric problems, a conflation this study is explicitly designed to avoid.

Ceiling effects and Changed Measurement

Where a measurement scale has a fixed upper bound, as in a 0-50 mark scheme, scores that would otherwise exceed the ceiling are reduced to it, a phenomenon formally analogous to right-censoring in econometrics (Tobin, 1958). The ordinary least-squares regression applied to censored data understates the true (latent) slope of the underlying relationship. The Tobit model corrects for this by explicitly modelling the censoring mechanism, allowing the analyst to distinguish a genuinely flat relationship from one that only appears flat because of ceiling compression. The use of quantile regression (Koenker & Bassett, 1978) provides a complementary, non-parametric window onto the same phenomenon by allowing the internal and external variation in slope to vary the conditional distribution of marks, thereby enlightening the tracking of attainment breaks that slow down specifically at the top of the mark distribution.

Discrimination in Signal Detection

Framing an internal mark as a diagnostic signal for external pass/fail status draws on signal detection theory: the AUC statistic summarizes the probability that a randomly chosen student with pass marks obtains a higher internal mark than a randomly chosen fail student. This decision-analytic framing is theoretically distinct from a simple correlation, because it asks specifically whether the internal mark can be used to separate students into meaningful attainment categories, which is closer to how marks are actually used in practice for decision-making.

Connection with theory of Grade inflation

The theory of Grade inflation explains that the absence of external moderation in marks is awarded by interested students, which is upward over time or to sit above an external, standardized benchmark. The consequences for the awarding institution are asymmetric. This theoretical

expectation motivates the type of institution which provide the incentive structures across institutions, communities, and alternative schools as a key moderating variable in the internal-external relationship.

Review of Empirical Literature

In the process of reviewing the previous studies related to internal assessments are presented below.

Concurrent validity of Internal Marks

A comparison of teacher-assigned grades with external examination results has historically found a moderate level of correlation. It was found to be $r = .40$ to $.60$ in different studies, a pattern some scholars have read as evidence that grades are excessively subjective. In this context, Arencibia Alemán et al. (2024) found substantially higher agreement ($r = .64$ to $.86$) under a system explicitly aligned between internal and external assessment, and they use regression analyses to identify the residual sources of discrepancy that remain favorable conditions. This work establishes a methodological template rather than reporting only its correlation and a comparison benchmark drawn from a tightly regulated system, against which looser systems can be judged.

Grade inflation and institution type

Internal grade inflation is a well-documented practice concentrated in private secondary schools that has demonstrated consequences for fairness of access to higher education (Costa-Santos et al., 2018; Neves et al., 2017; Nata et al., 2014). This literature has produced replicable, regression-based measures of inflation for confounding differences in student intake quality (Neves et al., 2017). It was shown that the internal advantage of private schools can shrink or even reverse, and downstream outcomes such as later academic performance (Costa-Santos et al., 2018). Two limitations were found across the body of work: the institutional contrast in binary form between institutional versus public, and the analytic approach relies on mean differences and linear regression rather than statistics that separate discrimination from level bias.

Use of Correlations between the Assessments

Outside the education-assessment literature, the clinical-measurement literature has long recognized that correlation is an inadequate summary of agreement between two methods in the same quantity. Bland and Altman (1986) give the plotting the difference between two measurements against their mean and report the mean bias in 95% limits of agreement. An approach that separates systematic bias from random scatter. Lin (1989) suggests a concordance correlation coefficient (CCC) that penalizes both departure from a 1:1 relationship and departure from the identity line, making it strictly more conservative than a simple correlation. Shrout and Fleiss (1979) develop interclass correlation coefficients for the closely related problem of inter-rater reliability. These agreement-based tools are widely used in clinical measurement but have been comparatively little exploited in educational-assessment research.

Grading Behavior in Machine Learning

The method of unsupervised and unsupervised-adjacent machine learning is a tool for discovering structure in assessment. Finite mixture models (McLachlan & Peel, 2000) can recover latent sub-populations, such as an aligned achiever group versus a systematically lenient group, by using the joint distribution of two correlated scores, yielding a continuous, model-based severity or inflation score rather than a binary classification. Isolation forests (Liu et al., 2008) provide a model-agnostic method for identifying inconsistent cases in multivariate data without requiring a labelled training set, and have been applied across fraud-detection and quality-control settings to problems structurally similar to identifying irregular grading. The unbiased assessment system is associated with how teachers' attitudes change throughout their careers for sustaining positive engagement with learners across all career stages (Ghimire, 2025). The importance of Random Forest classification, together with permutation-based variable importance in cross-validated gradient boosting, provides a standard toolkit (Pedregosa et al., 2011) for asking whether a categorical grouping variable can be recovered from continuous features, and for quantifying out-of-sample predictive validity.

Research gap

Research on internal assessment in Nepal is dominated by qualitative studies of teacher and student perceptions in school-based, continuous assessment practices. A large-sample, quantitative study of internal-external concordance that explicitly incorporates a distinctive three-way classification of institution types, such as: Institutional, community, or alternate type, was lacking in the researcher's knowledge of the researcher. The present study addresses this gap using authentic basic-level examination data, and it additionally introduces agreement-based statistics and machine-learning methods, as the simple coefficient of determination is insufficient and potentially misleading.

Methods and Procedures

This section included the research design, population and sample, construction and validation of tools, data collection and analysis procedure, results and discussion, and finally, conclusion and implication of the study.

Research Design

The study follows a quantitative, cross-sectional, correlational-comparative research design based on the secondary analysis of institutional examination records. The design is correlational and examines the strength and form of association between two continuous variables (internal and external marks). A comparative association and its component aspects of discrimination differ systematically across three institution types. The design further incorporates an exploratory, data-driven component in which unsupervised and supervised machine-learning techniques are used to discover latent grading systems and the type of institution from marks only. This combination of confirmatory statistical testing and exploratory machine learning allows the study to address the research questions from complementary angles: classical inferential statistics,

which establish an association or a group, while the machine-learning components characterize the shape, stability, and practical reparability of the involved patterns.

Population

The population of this study includes all the students who appeared in the Basic Level Examination (grade 8) based on the existing curriculum of Nepal. The internally awarded mark and an externally awarded mark were recorded on the common 0-50 scale for the subject, level, and year covered by the dataset. This population is stratified, on the basis of basic level examinations, into three types of institutions: Institutional schools, community schools, and alternate (non-formal or alternative-modality schools). The target population is fully enumerable from district examination records, rather than requiring an external sampling frame to be constructed.

Sample and Sampling Procedures

The study used a census (complete-enumeration) approach rather than a probability sample; the district and examination cycle under study were obtained from the Municipality Education Office. It was excluded from the analytic dataset records by removing with a missing internal or external mark were excluded from the analytic dataset. Since neither component can be meaningfully imputed without additional covariates that are not available in the dataset. After this screening and refining the data with $n = 2788$ students, distributed across institution type as follows: Institutional school (n) = 2176, Community (n) = 563; and Alternative (n) = 49. The final sample is taken as a representative sample by the census method of the records provided by the education department of the municipality of 2082 BS.

Construction and Validation of Tools

The secondary source of data was taken from the education office of the education department of the municipality, with the scores obtained by the students in their basic level examination. This shows that the scores of students were collected from authentic sources. The examination was taken on a cluster basis to ensure its validity. The test items used in the examinations were a) constructed based on the existing education system of the Municipality and b) subjected to analytic validation procedures applied to the resulting dataset before practical analysis.

Background of the Internal and External marks

In basic level education, external marks are derived from a common, board-set, and centrally moderated examination administered under the municipal board's grading directive (National Examinations Board, 2021), which is held at the end of grade eight and is therefore standardized across all institution types by construction. Internal marks are obtained from continuous, classroom-embedded formative assessment evaluated by the teachers in classwork, homework, projects, and practical work. The periodic tests were conducted by their own teachers in respective schools. School Education Sector Plan's formative-assessment guidelines are therefore not standardized across schools by construction; this asymmetry is precisely the

methodological reason the study exists. The equal (50:50) external-internal weighting in Basic Level examination. The equal (50:50) external-internal weighting used for these records is consistent with National Examinations Board practical subjects, for which the board departs from its more common division of 75 for theory and 25 for internal evaluations. (National Examinations Board, 2021).

Data validation and quality screening

Before substantive analysis, the dataset was screened for completeness by removing the absent students for the validity (marks were checked against the permissible 0-50 scale) of the data record institution type. It was not possible to validate the dataset against a multilevel, school cluster, nor to attach student-level covariates such as gender or socio-economic status. These are acknowledged as intrinsic limitations of the available records rather than remediable defects in the validation procedure.

The study itself constructs derived analytic instruments such as the k-means grading-regime clusters, the Gaussian-mixture inflation score, and the classification models were validated using standard machine-learning diagnostics rather than classical psychometric reliability coefficients. Since they are statistically constructed rather than test instruments administered to respondents. The number of clusters retained in the k-means analysis was assessed using the silhouette coefficient, and the stability of cluster assignment was assessed using the bootstrap Adjusted Rand Index (ARI); the predictive models (random forest, gradient boosting) were validated using 5-fold cross-validation and permutation-based variable importance, together which provide a sample check against overfitting; and the classification-based discrimination measure was validated against an explicit, stated pass-threshold on the external mark.

Data Collection and Analysis Procedure

Formal permission (Letter of Ethical Approval) and a formal request letter were submitted to access and use basic-level examination records; it was sought and obtained from the relevant municipal education authority. Student-level internal and external marks, together with the institution-type code, were extracted from the examination records obtained from the education office for the municipality for the level examination cycle under study. To maintain confidentiality, each student's data were coded as S0001, S0002 ... instead of their real names, and schools were recorded by type: Institutional school (Sc1), Community (Sc2), and Alternative (Sc3). The extracted records were transferred into a secure analytic dataset containing no personally identifying fields, and were then screened for completeness and range validity as described below. All subsequent statistical and machine-learning procedures were conducted in Python, principally using the scikit-learn library (Pedregosa et al., 2011) for machine-learning procedures and the statsmodels library (Seabold & Perktold, 2010) for classical statistical and econometric models, and the full analysis pipeline was retained as a reproducible notebook.

Statistical analysis Procedures

An Association between internal and external marks was measured using Pearson, Spearman, and Kendall correlation coefficients and, to detect non-linear dependence, a linear correlation coefficient was used. Agreement, as distinct from correlation, was quantified using Lin's concordance correlation coefficient (CCC) (Lin, 1989), the absolute-agreement intraclass correlation coefficient (ICC(A,1)) (Shrout & Fleiss, 1979), and Bland-Altman bias with 95% confidence-level agreement were used (Bland & Altman, 1986). Because internal marks are right-censored at the scale ceiling of 50, a Tobit model (Tobin, 1958) was fitted to improve the latent slope of the internal-external relationship, and its ordinary least squares was measured by regression (Koenker & Bassett, 1978). It was used to test whether the internal-external slope collapses at high quantiles of the internal-mark distribution, a pattern diagnostic of ceiling decoupling. As a decision-analytic view of discrimination, the external mark was dichotomized at an illustrative pass threshold, and the area under the ROC curve (AUC) of the internal mark as a classifier of external pass/fail status was computed. Moreover, Group differences between internal-external gaps across institution types were tested using one-way ANOVA.

Machine-learning analysis

Unsupervised grading regimes were extracted using k-means clustering applied to the three-dimensional feature space (External, Internal mark and Internal-external gap), with the number of clusters determined using the silhouette coefficient and the stability of cluster membership assessed using the bootstrap Adjusted Rand Index (ARI). A two-component Gaussian mixture model (McLachlan & Peel, 2000) was fitted to the joint (external, internal) distribution; the posterior probability of membership in the higher-gap component was used as a continuous, model-based inflation score, which was then aggregated by institution type. A random-forest classifier was trained to predict institution type from the mark-based features, with permutation importance used to rank feature contributions an isolation forest (Liu et al., 2008) was used to show the anomalous internal-external gaps within each institution type.

Results and Discussion

The relationship between internal marks and external marks in every institution type, most extremely in Institutional schools, where internal marks quantity up near the scale ceiling: 79% of Institutional students received an internal mark. Similarly, the following Table 1 summarizes the descriptive and concordance patterns across the three types of schools.

From the Table 1, it was found that the means of different types of schools in internal and external evaluations. The table also shows the Pearson correlation coefficient between the scores obtained by students in internal and external evaluations.

Table 1.
Descriptive Statistics and Concordance by Institution Type

Type	n	External Mean	Internal Mean	Pearson (r)	Lin's CCC	% Internal $\geq$ 45
Institutional	2176	27.5	46.2	0.31	0.07	79
Community	563	18.0	38.8	0.51	0.14	30
Alternative	49	3.3	31.3	0.51	0.05	10

The following figures indicate that the agreement between the two mark sources is poor throughout the sample: Lin's CCC is 0.07, 0.14, and 0.05 for Institutional, Community, and Alternative schools, respectively, with mean internal-external gaps of 18.7, 20.8, and 28.0 marks, and wide limits of agreement in every type. Rank-based and information-theoretic measures corroborate a real but weak dependence that is weakest for Institutional schools (Spearman = 0.29, Kendall = 0.21, mutual information = 0.08, distance correlation = 0.31) and comparatively stronger for Community schools (Spearman = 0.53, distance correlation = 0.49). The weak coefficients are therefore best read as a substantive empirical finding rather than as an artefact of the chosen statistical model.

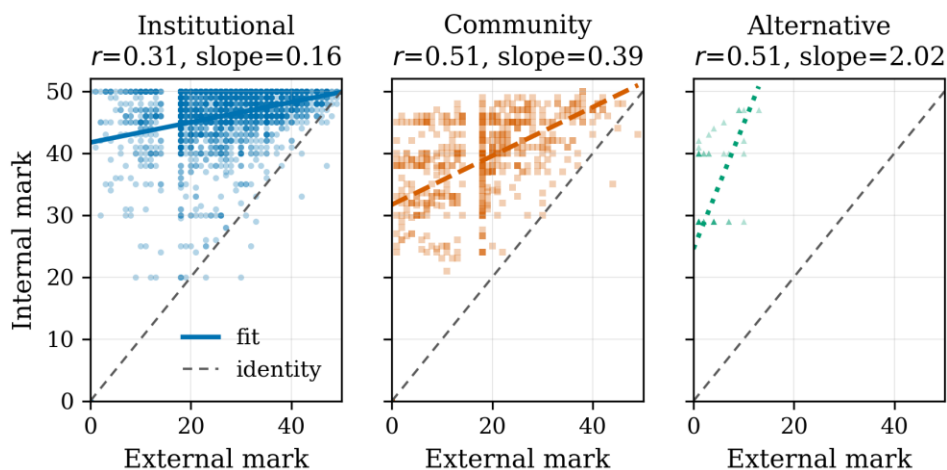


Figure 1. *Internal versus external marks by institution type, with the ordinary least-squares fit for Community and Alternative schools.*

Ceiling-aware Regression

The low Institutional association is only a part of the mechanical artefact of the ceiling. After correcting for censoring at 50 marks, the Tobit internal-on-external slope for Institutional schools rises from 0.16 to 0.20, remaining close to flat; the corresponding Community and Alternative slopes are essentially unchanged by the correction, as given in Table 2. Quantile regression makes the ceiling effect explicit: the Institutional slope falls from 0.11 at the median to 0.00 at the 90th percentile, a complete decoupling of internal marks from attainment at the top of

the distribution, whereas the slope of community schools declines more gently, from 0.32 to 0.11. Internal marks therefore progressively stop tracking attainment as they approach the ceiling, and this pattern is simple in institutional schools.

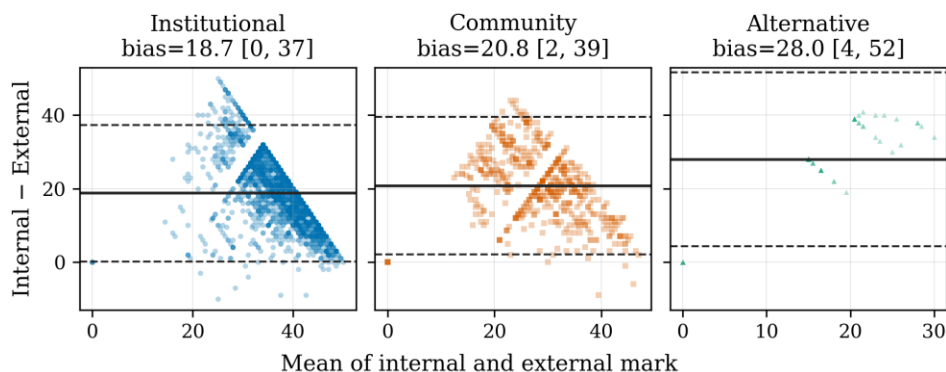


Figure 2. Bland-Altman agreement plots (internal minus external, against their mean) by institution type.

Table 2.

Censoring-corrected (Tobit) and quantile slopes of internal on external marks

Type	OLS slope	Tobit slope	Slope @ q(0.5)	Slope @ q(0.9)
Institutional	0.16	0.20	0.11	0.00
Community	0.39	0.40	0.32	0.11
Alternative	2.02	2.02	1.50	0.75

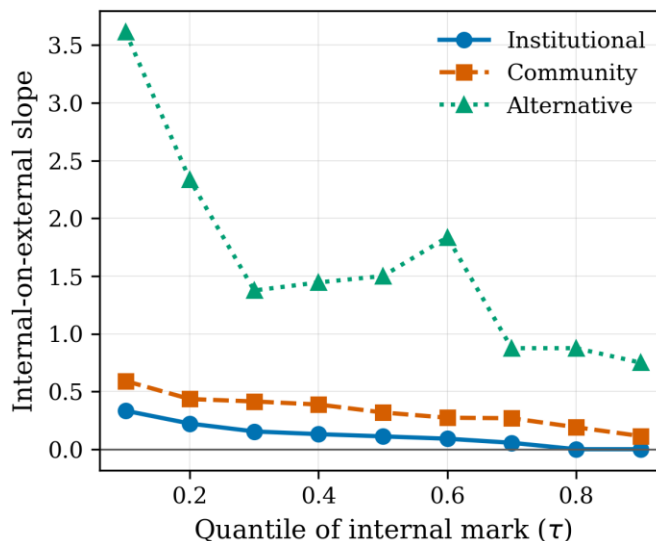


Figure 3. Quantile-Regression Slope of Internal or External Marks

Recast as a classifier of external pass/fail status, the internal mark discriminates poorly for institutional schools (AUC = 0.62) and noticeably better for community schools (AUC = 0.78); the overall AUC across the sample is 0.73. The Alternative AUC is undefined at the illustrative

threshold used, since virtually all Alternative students fall below this threshold. The ultimately replaced with the Board's official external pass mark. The calibration view is presented in Figure 4, which conveys the same finding intuitively: for a given internal mark, the mean external mark actually lies below the identity line, and the Institutional curve is both the lowest and the flattest of the three high internal marks that do not correspond to proportionally higher external attainment.

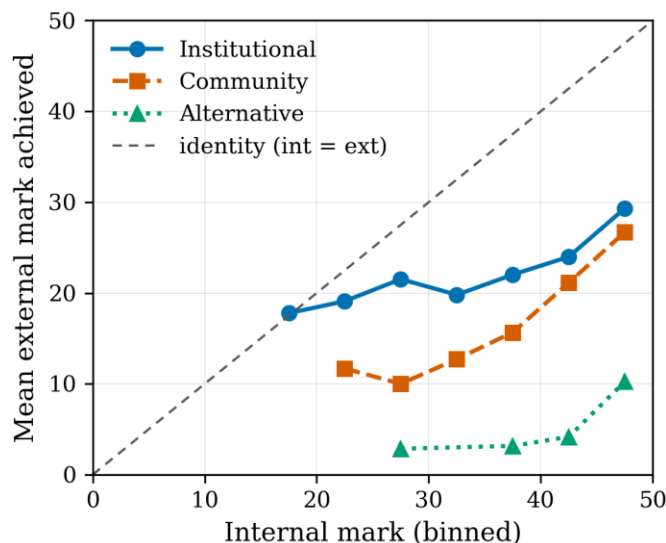


Figure 4. Calibration of internal marks against external attainment.

The mean internal-external gap differs significantly across institution types (ANOVA $F = 31.1$, $p < .001$, but the effect size is small at 0.02), and the variance of the gap does not differ significantly across types. Pairwise leniency contrasts are modest between institutional and community schools (Cohen's $d = -0.22$, 95% CI $[-0.32, -0.13]$) and considerably larger between Institutional schools and Alternative schools ($d = -0.97$ and -0.74) for the two types of schools. The differences that matter substantively are concentrated in discrimination and calibration in the raw fluctuation (variance) of the gap.

Grading Regimes and a Model-Based Inflation Score

The cluster analysis gives three stable grading systems (bootstrap ARI = 0.97; the silhouette statistic favors $k = 2$ at 0.40, with $k = 3$ retained for interpretability): an aligned-high-attainment regime (external mean 34, internal mean 47), a lenient system (external mean 18, internal mean 46), and a harsh-low regime (external mean 13, internal mean 28). Institutional students fall almost entirely within the two ceiling-internal systems (96%), whereas Community and Alternative students spread more broadly into the lower-internal systems as given in Table 3. The mixture-model inflation score is 0.23, 0.69, and 0.91 for Institutional, Community, and Alternative schools, respectively, the reverse of the ceiling-compression ordering. The score measures large absolute internal-over-external gaps, which peak among the many low-external-

attainment students that were found in community and alternative schools, rather than among the high-attaining institutional students whose marks are compressed at the ceiling.

Table 3.

Grading-regime Membership Share within each Institution Type

Type	Aligned-high	Lenient	Harsh-low
Institutional	0.56	0.40	0.04
Community	0.23	0.47	0.29
Alternative	0.00	0.45	0.55

Classification of Grading-Style

If institution type grade is genuinely distinctive, a classifier should be able to recover institution type from marks alone. A random forest trained on (external mark, internal mark, gap) attains 77% accuracy and a macro- F_1 of 0.58, well above the three-class chance level of approximately 0.33, but the signal classification is concentrated in institutional schools ($F_1 = 0.86$), while in the case of community and alternative cases are frequently misassigned to the institutional class. On a reduced, two-feature model (external, internal), balanced accuracy is 0.63 (chance = 0.33), with internal and external marks contributing comparably to the prediction. Grading style is therefore type-informative but far from deterministic. An isolation forest trained on the same features identified evidently different irregularity rates by institution type: 3.0% of institutional records, 9.1% of community records, and 46.9% of Alternative records were identified in the grading system of Alternative-school. A finding consistent with its large and highly variable internal and external gaps.

Conclusion and Implications

Internal assessment in this municipality is only weakly and non-linearly concordant with external examination performance, and the nature of the discrepancies differs systematically by institution type. Inconsistent internal grading has two separable faces: a discrimination failure that is most acute, and most robust to the student-quality confound, in institutional (private) schools. A level (leniency) failure that is larger in community and alternative schools. The practical consequence is that internal marks should not be treated as equivalent across institution types in any high-stakes aggregation of internal and external assessment. Future work including school-level identifiers and student-level covariates could convert the currently associational leniency finding into a properly adjusted, multilevel estimate. In parallel, strengthening continuous assessment practice itself through competency-based tasks, constructive feedback, remedial instruction, and more diverse assessment methods remains important for making basic-level assessment more inclusive, equitable, and supportive of meaningful learning, independent of whatever moderation scheme is ultimately adopted for high-stakes aggregation.

Examination authorities should consider type-aware moderation or scaling of internal marks before collection with external marks, rather than a single, institution-blind weighting scheme. Because Institutional schools show a discrimination failure rather than only a leniency failure, moderation approaches that only rescale the mean (e.g., simple leniency correction) will not resolve the institutional problem; a discrimination-aware remedy, such as broadening the effective internal mark range or triangulating with school-based moderation panels, is required. For community and alternative schools, where the primary problem is level bias rather than discrimination, conventional mean-adjustment or z-score standardization by school type is more likely to be effective.

Practical implications for schools

1. Schools, and Institutional schools in particular, should examine whether their internal assessment practices allow sufficient headroom to distinguish strong from average-performing students, rather than converging the majority of students toward the top of the mark scale.
2. Establishing continuous, competency-based assessment, together with structured moderation of internal marks within and across schools, would directly address the discrimination failure identified in this study.

Implications for future research

1. Future studies should seek school-level identifiers to permit multilevel modelling that separates within-school from between-school variation in grading behavior.
2. Collection of student-level covariates (e.g., gender, subject, socio-economic status, and prior attainment) would allow the leniency finding to be adjusted for student-quality confounds and converted from an association into a more defensible causal claim.
3. Replication across basic level examination cycles and subjects would clarify whether the discrimination versus leniency pattern identified generalizes beyond a single cohort.

Limitations

The data record institution type but not a school identifier, excluding the multilevel model that would separate within-school and between-school variation in grading. No student-level covariates are available to adjust the leniency estimates for student quality. The Alternative subgroup is small ($n = 49$), and its estimates should accordingly be treated as breakable. The study is cross-sectional and confined to a single district, which limits both generalizability and any temporal inference about trends in grading behavior. The external pass threshold used to compute the AUC is illustrative and should be replaced with the examination board's official value in the application of any policy. Finally, the exact provenance of the 50-50 external-internal scale (subject, level, and year) should be stated precisely by the data provider, since this detail frames the population to which the study's results generalize.

Ethical Considerations

The analysis uses aggregate, de-identified examination records containing no personally identifying information; the data were obtained and used with the permission of the relevant municipal education authority for research purposes. No individual student can be identified from the marks or institution-type codes. All statistical and machine-learning procedures underlying this study are fully reproducible from the accompanying analysis notebook with Python.

Acknowledgments

The researchers are immeasurably indebted for the approval, assistance, and support provided by the University Grants Commission, Nepal, and all other entities that made this study possible.

Conflict of Interest

The authors declare that there is no conflict of interest related to the publication of this research.

Funding

This article is the by-product of the small-scale research project funded by the University Grants Commission for the first author. Hence, there was partial financial support for the publication of the research article.

Authors' Contribution

Since this is a collaborative work conducted by three authors, all authors have contributed simultaneously to the completion of this research article. The First author has collected the necessary data for the completion of small-scale research and prepared the draft of this article. Similarly, the second author has formatted the research according to the research article with the necessary theoretical and empirical literature, and the third author has analyzed the numerical data by using data analysis software. Finally, all three authors have reviewed the draft and refine accordingly.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Arencibia Alemán, J. M., Sandsør, A. M. J., Zachrisson, H. D., & Blömeke, S. (2024). Teacher-assigned grades and external exams: Sources of discrepancy. *Assessment in Education: Principles, Policy & Practice*, 31(2), 94–115.
<https://doi.org/10.1080/0969594X.2024.2338764>
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307–310.
[https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)

- Costa-Santos, C., Vieira-Marques, P., Costa-Pereira, A., Ferreira, M. A., & Freitas, A. (2018). Do students from public schools fare better in medical school than their colleagues from private schools? If so, what can we learn from this? *BMC Medical Education*, 18(1), 51. <https://doi.org/10.1186/s12909-018-1162-x>
- Ghimire, S. P. (2025). Teaching trajectories of mathematics teachers: Teachers' attitudes towards commitment to students' learning. *Education and Development*, 34(1), 97–110. <https://doi.org/10.3126/ed.v34i1.80292>
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometric*, 46(1), 33–50. <https://doi.org/10.2307/1913643>
- Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255–268. <https://doi.org/10.2307/2532051>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE. <https://doi.org/10.1109/ICDM.2008.17>
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Wiley.
- Nata, G., Pereira, M. J., & Neves, T. (2014). Unfairness in access to higher education: A 11 year comparison of grade inflation by private and public secondary schools in Portugal. *Higher Education*, 68(6), 851–874. <https://doi.org/10.1007/s10734-014-9748-7>
- National Examinations Board. (2021). *Letter grading directive for classes 1–12* (Government of Nepal, Ministry of Education, Science and Technology). <https://neb.gov.np>
- Neves, T., Ferraz, H., & Nata, G. (2017). Social inequality in access to higher education: Grade inflation in private schools and the ineffectiveness of compensatory education. *International Studies in Sociology of Education*, 26(2), 190–210. <https://doi.org/10.1080/09620214.2016.1191966>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 92–96). <https://doi.org/10.25080/Majora-92bf1922-011>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36. <https://doi.org/10.2307/1907382>