




[www.ijtes.net](http://www.ijtes.net)

## Standardizing College GPA for University Senior Intake Admissions: A Moderation Tool for Equitable Evaluation

**Chenxi Dong**   
The Education University of Hong Kong, Hong Kong

**Yimin Yuan**   
The University of Adelaide, Australia

### To cite this article:

Dong, C. & Yuan, Y. (2025). Standardizing college GPA for university senior intake admissions: A moderation tool for equitable evaluation. *International Journal of Technology in Education and Science (IJTES)*, 9(4), 512-521. <https://doi.org/10.46328/ijtes.651>

The International Journal of Technology in Education and Science (IJTES) is a peer-reviewed scholarly online journal. This article may be used for research, teaching, and private study purposes. Authors alone are responsible for the contents of their articles. The journal owns the copyright of the articles. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of the research material. All authors are requested to disclose any actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations regarding the submitted work.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

# Standardizing College GPA for University Senior Intake Admissions: A Moderation Tool for Equitable Evaluation

Chenxi Dong, Yimin Yuan

---

## Article Info

### Article History

Received:

14 March 2025

Accepted:

10 August 2025

---

### Keywords

Educational assessment

Grading disparities

GPA calibration

Educational data mining

---

## Abstract

University admissions follow two distinct pathways: first-year admissions using standardized tests (e.g., SAT in the United States, GaoKao in China) and senior intake admissions, where students enter university based on their college GPA. While first-year admissions benefit from uniform comparison metrics, senior intake processes rely on college GPAs that vary significantly across institutions. This paper addresses this standardization gap by developing and validating methods to calibrate disparate college GPAs. Recognizing that raw GPA is a biased measure influenced by institutional grading policies, we propose three moderation models to align raw College GPAs ( $X$ ) with observed University Performance ( $Z$ ): (1) a Mean-Adjusted model, (2) Direct Benchmark-to-Raw Score Regression, and (3) Inter-Score Regression with Benchmark Standardization. Evaluation using synthetic datasets (training  $N=75$ ; test sets  $N=95$ ,  $N=29$ ) demonstrates that Inter-Score Regression with Benchmark Standardization produces the most substantial improvement in predictive validity ( $\Delta R^2 = +0.40$  for training,  $+0.24$  and  $+0.26$  for test sets), maintaining robustness across varying sample sizes and grade distributions. This research provides admissions officers with a standardized evaluation tool for senior intake, advancing equitable assessment practices in higher education.

---

## Introduction

University admissions processes follow two distinct pathways: first-year admissions and senior intake. For first-year admissions, universities typically rely on standardized tests (e.g., SAT in the United States and GaoKao in China) that provide uniform benchmarks for comparing applicants (Kim, 2011). In contrast, senior student admissions frequently rely on college-level GPAs as a key indicator of academic potential. However, a significant challenge arises from the inherent differences in grading systems across various colleges. Institutional grading scales and assessment methods can vary substantially, leading to inconsistent evaluations of student performance. Observed GPA can be a systematically biased measure of academic performance, influenced by the specific courses a student takes and the grading stringency within those courses and institutions (Moore et al., 2010). This inconsistency makes it difficult for university admissions committees to compare applicants fairly and equitably across different institutions (Muijtjens et al., 2008).

Consider the following scenario: Two students apply to a university. Student A has a GPA of 3.6 from a college known for grade inflation, while Student B has a GPA of 3.4 with a reputation for rigorous grading. A direct comparison of these GPAs might unfairly favor Student A. This example highlights a fundamental inequity: while first-year applicants are evaluated using standardized metrics enabling direct comparison, senior applicants are assessed using institution-specific GPAs that lack standardization (Rayevnyeva et al., 2018; Wittman, 2022).

This example highlights a fundamental inequity in evaluation methods using observed GPA: it is a systematically biased measure of academic performance. Tomkin & West's research in 2022 demonstrated that observed GPA significantly underestimates grading disparities between STEM and non-STEM courses, with differences averaging approximately 0.4 grade points on a 4.0 scale. They note that "grade offset and grade penalty studies that use observed GPAs as baselines of student ability are likely to be systematically biased" (Tomkin & West, 2022). This reinforces our argument for developing robust moderation methods that account for institutional variation in grading stringency.

This standardization challenge is not unique to senior admissions. Recent research by Molontay and Nagy (2022) analyzed data from over 24,000 students and demonstrated that adjusting existing admission metrics can significantly improve their predictive validity across disciplines. Similarly, Tesema (2014) found substantial variations in how well high school GPAs predicted university performance across different academic programs, with predictive variance ranging from 16% to 59%, depending on the discipline. These findings highlight the necessity for moderation approaches to account for institutional differences in grading practices. Our approach aligns with established educational measurement principles recognizing the need for score comparability across different contexts. As Liu et al. (2024) demonstrated in their work on fair assessment systems, statistical approaches to standardizing educational metrics can significantly improve fairness and validity. This underscores the importance of methods that can account for institutional differences in grading practices. Johnson (2003) concluded that heterogeneity in grading practices undermines academic standards and the assessment of student learning. Our research builds upon these observations by developing moderation models specifically calibrated for senior intake admissions.

To address this critical issue, the primary goal of this research is to propose and evaluate robust GPA moderation methods that can align raw College GPAs ( $X$ ) with a more calibrated measure of University Performance ( $Z$ ). Drawing inspiration from the concept of "grade offset" (Tomkin & West, 2022), our methods aim to account for these institutional grading disparities. By generating moderated GPAs ( $Y$ ), we intend to minimize the impact of these variations, enabling more equitable and accurate comparisons of applicants

This study seeks to:

- **Create Standardization Equivalence:** Develop a standardization tool for senior/transfer admissions analogous to what standardized tests provide for first-year admissions. By aligning raw College GPAs ( $X$ ) to a more reliable university performance indicator ( $Z$ ) to generate moderated GPAs ( $Y$ ), we aim to minimize grading disparities across institutions.
- **Evaluate Effectiveness:** Measure the impact of moderation by comparing the coefficient of determination ( $R^2$ ) before and after moderation. Specifically, if  $R^2(Y, Z) > R^2(X, Z)$ , it indicates that the moderated

GPA<sub>s</sub> (Y) better align with the benchmark (Z) than the original GPA<sub>s</sub> (X), suggesting improved predictive validity for university performance.

- Facilitate Implementation: Develop an accessible framework for admissions officers to apply these moderation techniques in their evaluation processes.
- Examine Generalizability: Test moderation models on datasets with varying characteristics to ensure robustness across different institutional contexts and sample sizes.

Our research addresses a critical gap by providing the first comprehensive, empirically validated framework tailored to university senior intake admissions. While previous research has focused mainly on grade standardization in secondary education, this study extends these principles to the complex landscape of higher education, offering practical and readily deployable techniques for real-world admissions contexts.

## Method

### Data Description

Our methodological approach aligns with established practices in educational assessment. This study employs three synthetic datasets to evaluate the proposed GPA moderation models: a training dataset and two test datasets. We generated these datasets using Python 3.13 with NumPy and Pandas libraries to create realistic GPA distributions that reflect typical patterns observed in higher education contexts, shown in Table 1.

Table 1. Dataset Feature Description

Column Name	Data Type	Description	Range/Values
Raw College GPA (X)	Numerical	Student's original GPA from college	0.00 to 4.33
University Performance GPA (Z)	Numerical	Student's subsequent performance GPA at university	0.00 to 4.33
College	Categorical	College Name	College 1 or 2

Figure 1 illustrates the fundamental challenge addressed by this research: significant misalignment between Raw College GPA (X) and subsequent University Performance GPA (Z).

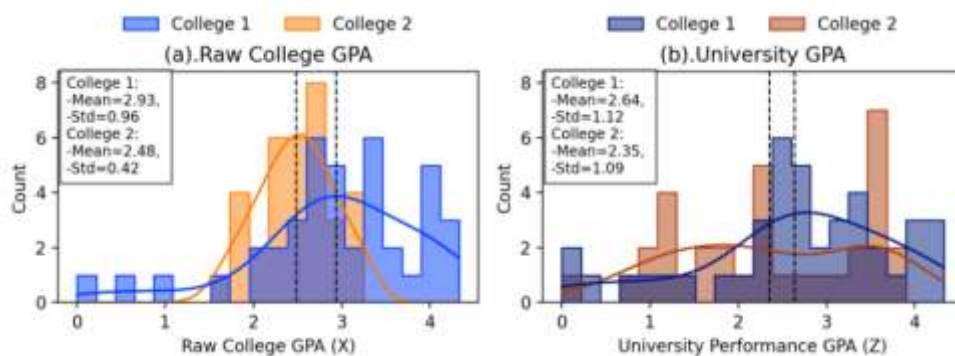


Figure 1. Comparison of Raw College GPA (X) and University Performance GPA (Z) Distributions in Training Dataset

College 1 shows right-skewed raw GPAs concentrated above 3.0, while its actual university performance follows a more symmetric distribution centered around 2.5. Similarly, College 2's raw GPAs cluster narrowly around 2.7, but its university performance spans from 1.0 to 3.5. These discrepancies demonstrate the need for effective GPA moderation.

We designed the test datasets to evaluate model robustness under different conditions. Test Dataset 1 (N=95) features substantially different means and standard deviations compared to the training set, with a wider spread of GPAs (SD = 1.35) and lower mean GPA. Test Dataset 2 (N=29) has a smaller sample size to assess performance with limited data availability. Table 2 summarizes the descriptive statistics of all three datasets.

Table 2. Descriptive Statistics of Training and Test Datasets

Dataset	N	College Count	Raw College GPA (X)	University Performance GPA (Z)
Train Set	75	N <sub>College 1</sub> = 41	Mean = 2.73	Mean = 2.51
		N <sub>College 2</sub> = 34	SD = 0.79	SD = 1.10
Test Set 1	95	N <sub>College 1</sub> = 51	Mean = 2.22	Mean = 2.40
		N <sub>College 2</sub> = 44	SD = 1.35	SD = 1.09
Test Set 2	29	N <sub>College 1</sub> = 16	Mean = 2.33	Mean = 2.08
		N <sub>College 2</sub> = 13	SD = 0.93	SD = 1.31

### Three Moderation Models

We present three GPA moderation models to improve the alignment of College GPAs (X) with University Performance GPAs (Z).

- **Model 1: Mean-Adjusted Model (Baseline)**

This baseline method shifts raw college GPAs (X) to match each institution's mean University Performance GPA (Z). Although it adjusts the central tendency, it does not address differences in grade dispersion. The moderated GPA (Y) is calculated as:

$$y = \bar{Z} + (X - \bar{X}) \tag{1}$$

Where:

- y: Moderated college GPA.
- $\bar{Z}$ : Group mean of university GPA,  $\bar{Z} = \text{mean}(Z \mid \text{College})$ .
- $\bar{X}$ : Group mean of raw college GPA,  $\bar{X} = \text{mean}(X \mid \text{College})$ .

- **Model 2: Direct Benchmark-to-Raw Score Regression**

This model uses linear regression to calibrate college GPAs (X) against university performance GPAs (Z). The moderated GPA (Y) is obtained by minimizing the sum of squared differences between moderated GPA (Y) and actual university GPAs (Z):

$$y = kX + b \tag{2}$$

Where:

- k: The slope coefficient capturing the linear relationship between X and Z.
  - b: The intercept parameter
  - All other symbols are as previously defined.
- Model 3: Inter-Score Regression with Benchmark Standardization

Similar to Model 2, this model transforms the raw college GPAs (X) into moderated GPAs (Y) to better align with the University Performance GPAs (Z). However, unlike Model 2, it also incorporates the University Performance GPA (Z) standard deviation to scale the raw college GPAs. This method builds upon the strengths of Model 2 by not only considering the linear relationship between college GPA and university performance GPA but also explicitly addressing differences in the spread of GPAs across institutions. The moderated GPA (Y) is:

$$y = X_{mean} + \beta (\bar{Z} - Z_{mean}) + (X - \bar{X}) \frac{S_z}{S_x} \quad (3)$$

Where:

- $X_{mean}$ : Global mean of raw college GPA.
- $\beta$ : The regression coefficient between raw college GPA (X) and university GPA (Z) ranges from 0 to 1.
- $Z_{mean}$ : Global mean of university GPA.
- $S_x$ : Group standard deviation of raw college GPA,  $S_x = \text{std}(X | \text{College})$ .
- $S_z$ : Group standard deviation of the university GPA,  $S_z = \text{std}(Z | \text{College})$ .
- All other symbols are as previously defined.

This method is designed to provide the most comprehensive moderation by accounting for both differences in central tendency and variability of GPAs across institutions.

### Evaluation Metrics

The effectiveness of each moderation model is assessed using the coefficient of determination ( $R^2$ ), which measures the proportion of variance in the University Performance GPA (Z) explained by the moderated GPA (Y). The change in  $R^2$  ( $\Delta R^2$ ) is calculated by:  $\Delta R^2 = R^2(Y, Z) - R^2(X, Z)$ . A positive  $\Delta R^2$  indicates that the moderation has improved the alignment between the raw GPAs and the benchmark University GPAs, with higher values signifying more effective moderation.

### Results

This section presents the comparative performance of the three GPA moderation models across training (N=75) and two test datasets using the previously defined  $\Delta R^2$  metric. Two test sets evaluated method robustness: Test Set 1 (N=95) with different standard deviations and Test Set 2 (N=29) with a limited sample size.

Table 3. Performance Comparison of GPA Moderation Models

Model	Dataset	R <sup>2</sup> (X, Z)	R <sup>2</sup> (Y, Z)	$\Delta R^2$
<b>Model 1:</b> Mean-Adjusted	Train Set	0.15	0.20	+0.05
	Test Set 1	0.08	0.07	-0.01
	Test Set 2	0.06	0.10	+0.04
<b>Model 2:</b> Direct Benchmark-to-Raw Score Regression	Train Set	0.15	0.36	+0.21
	Test Set 1	0.08	0.10	+0.02
	Test Set 2	0.06	0.27	+0.21
<b>Model 3:</b> The Inter-Score Regression with Benchmark Standardization	Train Set	0.15	0.55	<b>+0.40</b>
	Test Set 1	0.08	0.32	<b>+0.24</b>
	Test Set 2	0.06	0.32	<b>+0.26</b>

The performance improvements observed in our models align with similar gains reported in related research. The substantial improvement achieved by Model 3 ( $\Delta R^2 = +0.40$ ) suggests that comprehensive standardization approaches that address both central tendency and variability have particular utility for senior admissions contexts. This finding is consistent with research showing that incorporating institutional context into assessment models can significantly improve fairness (Liu et al., 2024).

- *Model 1 (Mean-Adjusted)*: This model showed limited improvement in the training set ( $\Delta R^2 = +0.05$ ). Critically, its performance degraded in Test Set 1 ( $\Delta R^2 = -0.01$ ), which was designed with a significantly different standard deviation ( $SD = 1.35$ ) compared to the training set ( $SD = 0.79$ ). This highlights the method's sensitivity to changes in data variance and its limited generalizability.
- *Model 2 (Direct Benchmark-to-Raw Score Regression)*: This model substantially improved over the baseline, achieving notable gains in training ( $\Delta R^2 = +0.21$ ). The model maintained consistent performance in Test Set 2 ( $\Delta R^2 = +0.21$ ), showing strong generalizability with limited data. While its performance in Test Set 1 was modest ( $\Delta R^2 = +0.02$ ), it still improved upon the raw scores, suggesting better resilience to variance changes than Model 1.
- *Model 3 (The Inter-Score Regression with Benchmark Standardization)*: This model consistently achieved the highest  $\Delta R^2$  values across all datasets. The substantial improvement in the training set ( $\Delta R^2 = +0.40$ ) was maintained in Test Set 1 ( $\Delta R^2 = +0.24$ ), demonstrating robust performance despite the difference in variance. Additionally, Model 3 performed exceptionally well in Test Set 2, achieving a  $\Delta R^2$  of +0.26, even with a small sample size, indicating excellent generalizability across diverse data conditions. The superior performance of Model 3 aligns with previous research on grade standardization. Koester et al. (2016) and Matz et al. (2017) demonstrated the importance of accounting for structural differences when comparing grades across different academic environments. Their work showed that unadjusted comparisons can lead to misleading conclusions about student performance, similar to how our raw GPA comparisons produced less reliable predictions than our more comprehensive approach in Model 3.

## Discussion

Our analysis reveals distinct performance patterns among the three GPA moderation models, as illustrated by the distributions of raw, university, and moderated GPAs for the training set (Figures 2-4).

The Mean-Adjusted Model (Figure 2) shifts the raw GPA (X) distribution for each college to align with the mean University GPA (Z) without addressing differences in variance. This limited adjustment explains its modest improvement in predictive power ( $\Delta R^2$ ).

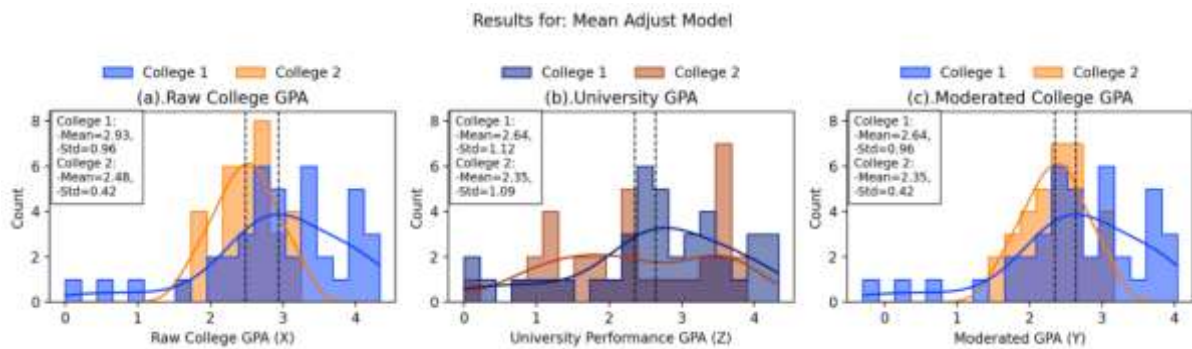


Figure 2. Distribution Comparison of Raw College GPA (X), Moderated GPA (Y), and University GPA (Z) – Model 1

The Direct Benchmark-to-Raw Score Regression (Figure 3) shows two key adjustments: College 1's mean shifts from 2.93 to 2.73, moving toward the target of 2.64, while its standard deviation changes minimally from 0.96 to 1.04. College 2's mean adjusts from 2.48 to 2.24, approaching its target of 2.35, though its distribution remains compressed with standard deviation barely increasing from 0.42 to 0.46 (far from target 1.09). This indicates that the method achieves directionally correct adjustments but falls notably short in correcting variance.

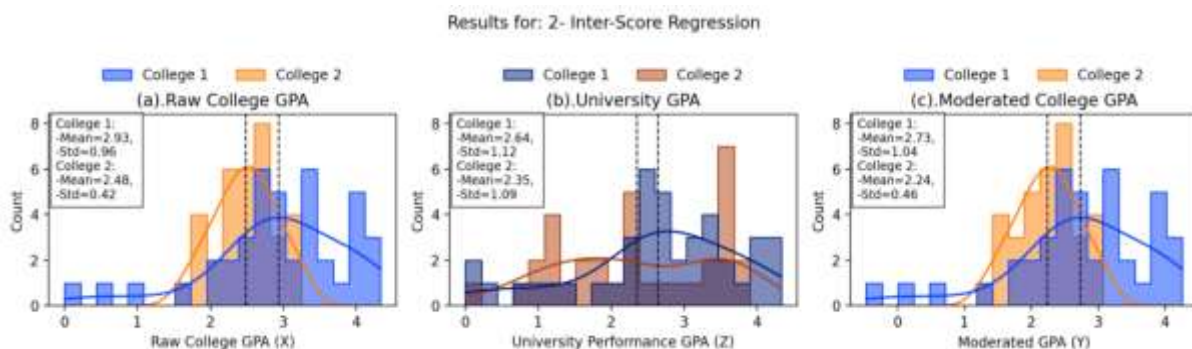


Figure 3. Distribution Comparison of Raw College GPA (X), Moderated GPA (Y), and University GPA (Z) – Model 2

The Inter-Score Regression with Benchmark Standardization (Figure 4) provides the most precise alignment. By directly scaling the standard deviation of X to match the benchmark standard deviation of Z, it ensures that both colleges' moderated GPA (Y) distributions precisely align with the benchmark. This is demonstrated by the



increase in standard deviation for College 1 from 0.96 to 1.12 and for College 2 from 0.42 to 1.09. This aligns with the understanding that grading stringency can manifest not only in lower average grades but also in the distribution and spread of grades (Tomkin & West, 2022). The formula for Model 3 allows for a global mean adjustment based on the overall relationship observed in the training data, coupled with a specific scaling of the standard deviation for each college to match the university performance benchmark.

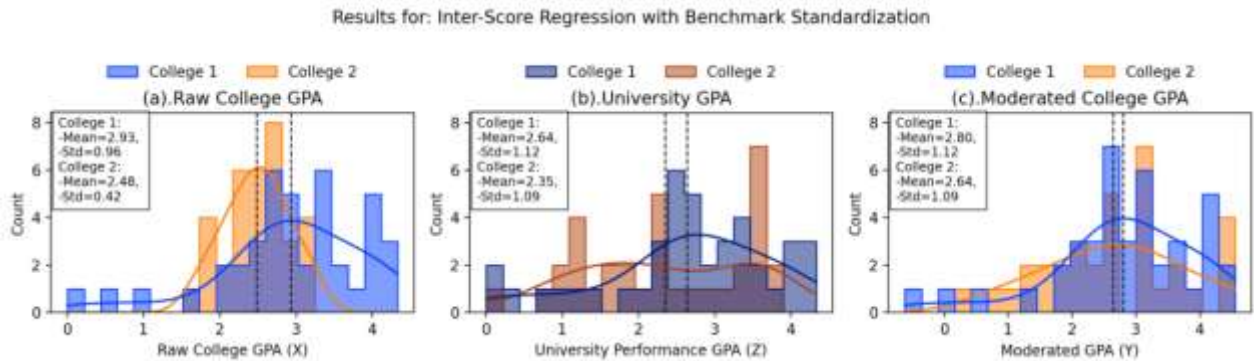


Figure 4. Distribution Comparison of Raw College GPA (X), Moderated GPA (Y), and University GPA (Z) – Model 3

These findings have important implications for equity in higher education admissions. Johnson (2003) argued that heterogeneity in grading practices can create systematic disadvantages for students from institutions with more stringent grading policies. By implementing standardization methods like our Model 3, universities can create more level evaluation systems that better reflect student capabilities rather than institutional grading conventions. For practical implementation,

Model 3 requires historical data on both college GPAs and subsequent university performance from past cohorts. Admissions offices could implement this system incrementally, beginning with programs or colleges having the most reliable historical data. By implementing this method, universities can evaluate senior intake applicants with greater confidence in the comparability of their academic credentials. While the synthetic datasets were designed to simulate realistic scenarios, future research should validate these findings using actual admissions data.

## Conclusion

This study evaluated three models for moderating college GPAs to enhance the fairness of senior student admissions decisions. Our findings demonstrate that the Inter-Score Regression with Benchmark Standardization model (Model 3) is the most effective approach, consistently achieving the highest improvements in predictive power ( $\Delta R^2$ ). Specifically, Model 3 yielded improvements of +0.40 in the training set, +0.24 in Test Set 1 (designed with differing variance), and +0.26 in Test Set 2 (with a small sample size). This model's robustness to variations in sample size and GPA distributions makes it a promising tool for ensuring a more equitable evaluation of applicants from different academic backgrounds. While Direct Benchmark-to-Raw Score Regression (Model 2) offers moderate improvement, the Mean-Adjusted model (Model 1) proves insufficient, notably when institutional GPA distributions differ significantly in spread.

Our research solves this disparity by providing a statistically rigorous approach for standardizing institution GPAs. Our findings also underscore the importance of considering mean and standard deviation in GPA standardization, moving beyond simple adjustments that may overlook the complexities of grading disparities. The standardization approach developed in this study addresses a critical equity gap in higher education admissions, potentially enabling more diverse and qualified applicants to receive fair consideration regardless of their institution of origin. As Molontay and Nagy (2022) demonstrated, improving the calibration of admission metrics can enhance fairness and predictive accuracy without introducing entirely new assessment measures. Our research extends this principle specifically to the senior intake context, where standardization has received comparatively less attention. Future research should prioritize validating these findings with real-world data and implementing these moderation techniques in university admissions practices.

## References


- Caulkins, J. P., Larkey, P. D., & Wei, J. (1996). Adjusting GPA to reflect course difficulty.
- Goldman, R. D., & Widawski, M. H. (1976). A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement*, 36(2), 381-390.
- Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. Springer Science & Business Media.
- Kim, H. K. (2011). The Compositions and the Characteristics of the Chinese National Test for University Admissions, and the Analysis on Items Concerning Chemistry. *Journal of the Korean Association for Science Education*, 31(8), 1158-1174.
- Koester, B. P., Grom, G., & McKay, T. A. (2016). Patterns of gendered performance difference in introductory STEM courses. *arXiv preprint arXiv:1608.07565*.
- Liu, J., Hui, W. W. Y., Lee, R. K. W., & Lim, K. H. (2024). Fairness And Performance In Harmony: Data Debiasing Is All You Need. *arXiv preprint arXiv:2411.17374*.
- Matz, R. L., Koester, B. P., Fiorini, S., Grom, G., Shepard, L., Stangor, C. G., ... & McKay, T. A. (2017). Patterns of gendered performance differences in large introductory courses at five research universities. *Aera Open*, 3(4), 2332858417743754.
- Muijtjens, A. M., Schuwirth, L. W., Cohen-Schotanus, J., Thoben, A. J., & Vleuten, C. P. V. D. (2008). Benchmarking by cross-institutional comparison of student achievement in a progress test. *Medical Education*, 42(1), 82-88.
- Molontay, R., & Nagy, M. (2023). How to improve the predictive validity of a composite admission score? A case study from Hungary. *Assessment & Evaluation in Higher Education*, 48(4), 419-437.
- Moore, D. A., Swift, S. A., Sharek, Z. S., & Gino, F. (2010). Correspondence bias in performance evaluation: Why grade inflation works. *Personality and Social Psychology Bulletin*, 36(6), 843-852.
- NSW Education Standards Authority. (2024). *Assessment moderation*. NSW Government. <https://www.nsw.gov.au/education-and-training/nesa/hsc/exams-and-marking/assessment-moderation>
- Rayevnyeva, O. V., Aksonova, I. V., & Ostapenko, V. M. (2018). Assessment of institutional autonomy of higher education institutions: methodical approach. *Knowledge and Performance Management*, 2(1), 75-87. doi:10.21511/kpm.02(1).2018.07

---

**Author Information**

---

**Chenxi Dong**

 <https://orcid.org/0009-0007-6656-5353>


The Education University of Hong Kong

Hong Kong

Contact e-mail: [dongchenxi123@outlook.com](mailto:dongchenxi123@outlook.com)

---

**Yimin Yuan**

 <https://orcid.org/0009-0000-3119-0490>

The University of Adelaide

Australia