

Next-Unit English Writing Scores under Repeated Cohort Transfer



Gula Da¹ & Jinxing Zhao^{2,*}

¹Foreign Languages School, Pioneer College of Inner Mongolia University, P.R. China

²School of Mathematical Sciences, Inner Mongolia University, P.R. China

Abstract: A single future cohort can make an 80% prediction interval look stable. This paper predicts the next English writing score from information available at the current unit, then repeats the same train–calibrate–test transfer on seven successive entering cohorts. Scores are converted to a percent of that year’s full mark before modeling. The de-identified records are English writing assessments from ten undergraduate cohorts during 2016–2025 at the School of Mathematical Sciences, Inner Mongolia University. Across the seven transfers, last-score MAE is 5.40–5.68 points. Linear quantile P50 lowers it to 4.81–5.15, and quantile LightGBM to 4.80–5.30. Raw LightGBM P10–P90 intervals cover 69.4%–76.7% of outcomes at mean widths of 12.9–14.6. One pooled conformal correction moves coverage to 77.8%–82.0% and width to 15.1–16.6 on the six transfers whose calibration cohort has more than 200 students; the remaining transfer, calibrated on 121 students, reaches 83.8% at width 17.2. A separate correction for each target unit changes LightGBM coverage by at most 1.7 percentage points. The raw linear interval covers 77.5%–81.1%. Score history accounts for most of the point-error reduction. A student-clustered bootstrap does not find a year in which LightGBM has a lower MAE than the linear quantile. A short explanation interface repeats these numbers under a fixed contract. We did not collect teacher evaluations of the interface.

Keywords: quantile regression, conformal prediction, educational prediction, uncertainty, cohort transfer

1. Introduction

For an instructor considering support before the next writing assessment, the uncertainty around a forecast matters alongside its central estimate. We therefore report a median prediction together with a P10–P90 interval. An interval that covers approximately 80% of scores in one entering cohort may perform differently in a later cohort; evaluating a single transfer cannot establish how consistently that coverage is maintained.

Using only information recorded through unit t , we predict the percentage score at the student’s next recorded unit. The conditional 10th, 50th, and 90th percentiles are fitted by quantile regression (Koenker & Bassett, 1978), and a residual quantile from the preceding entering cohort supplies the conformal correction (Vovk et al., 2005; Romano et al., 2019). We repeat this procedure for test cohorts entering between 2019 and 2025. In each transfer, the models

are trained on earlier cohorts and calibrated on the immediately preceding cohort before being evaluated once on the test cohort. Students remain in a single role within each transfer.

Across the seven transfers, both quantile models improve on using the previous score as the prediction. The reduction in absolute error is roughly half a point to one point. Uncalibrated LightGBM intervals fall below 80% coverage in all seven transfers. Pooled calibration brings coverage to the high 70s or low 80s, with intervals approximately 15–17 points wide, while corrections estimated separately by target unit produce little further change. The raw linear intervals also vary across cohorts, reaching 77.5% coverage at their lowest. These differences would be obscured by reporting only one held-out year. The feature ablations attribute most of the point-prediction accuracy to score history, with little change after attendance and homework are removed.

Corresponding Author: Jinxing Zhao

School of Mathematical Sciences, Inner Mongolia University, P.R. China

©The Author(s) 2026. Published by BONI FUTURE DIGITAL PUBLISHING CO., LIMITED This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Student-clustered confidence intervals do not distinguish the two quantile models in every test year.

2. Related Work

Conditional quantiles can be estimated by minimizing pinball loss (Koenker & Bassett, 1978), using quantile regression forests (Meinshausen, 2006), or fitting a histogram-based boosting model such as LightGBM with a quantile objective (Ke et al., 2017). For interval construction, split conformal regression calibrates a point predictor using residuals from held-out observations (Lei et al., 2018). Conformalized quantile regression instead corrects an interval formed by fitted conditional quantiles (Romano et al., 2019). Its finite-sample coverage result relies on exchangeability of calibration and test observations (Vovk et al., 2005), a condition that may fail across entering cohorts. Under covariate shift, weighting is possible when the relevant likelihood ratio is known (Tibshirani et al., 2019). Other approaches address distributional change by reweighting earlier observations (Foygel Barber et al., 2023) or adapting the nominal level online (Gibbs & Candès, 2021). We do not estimate such weights here. It refits on earlier cohorts, calibrates on the previous cohort, and reports the coverage that the next cohort actually attains. Reviews of student-performance prediction usually pool records from one course offering (Namoun & Alshanqiti, 2021). A random split of unit records can place the same student in training and testing. Every split below keeps a student inside one cohort.

3. Method

3.1 Prediction task and features

Raw scores use a full mark of 100 for cohorts C2016–C2023 and 150 for C2024–C2025. Before any feature is built, each observed score is replaced by $100 \times \text{score} / \text{score_max}$. For a student after unit t , the feature vector contains the current percent score, the historical mean and standard deviation, the previous score, the recent score change, current and historical attendance, current and historical

homework completion, initial level, the unit index, and the pre-specified difficulty of the current and next units. Although two school codes appear in the records, they refer to the same college; we therefore omit a school indicator. Student, class, and teacher identifiers are excluded. Attendance and homework from unit $t+1$ are excluded.

3.2 Quantile models

For τ in $\{0.1, 0.5, 0.9\}$, both models minimize the sum of the pinball loss $\rho_\tau(y - f(x))$ on the training rows of that transfer, where $\rho_\tau(u) = \max\{\tau u, (\tau - 1)u\}$. The linear model uses an intercept and the same features after median imputation and standardization fitted on that transfer's training cohorts only. Coefficients are estimated by iteratively reweighted least squares. LightGBM uses 350 boosting rounds, learning rate 0.03, 15 leaves, minimum leaf size 40, feature fraction 0.85, bagging fraction 0.85, and L1/L2 penalties of 0.2/1.0. We use the fitted median, P50, for point prediction. P10 and P90 form the raw interval. Quantile crossing is removed by sorting the three outputs before calibration.

3.3 Conformal calibration

The symmetric nonconformity score on a calibration row is $s = \max\{Q_{10}(x) - y, y - Q_{90}(x)\}$. We set $\alpha = 0.2$ and take the empirical quantile at level $\text{ceil}((n+1)(1-\alpha))/n$. Under pooled calibration, all calibration rows contribute to a single correction q . The horizon-specific alternative computes q separately for each target unit. A target unit absent from the calibration cohort receives the pooled correction. After calibration, we restrict the endpoints to the percentage scale, giving $[\max(0, Q_{10}(x) - q), \min(100, Q_{90}(x) + q)]$. A correction estimated from one cohort is used only for the following cohort. We report the coverage observed in these transfers without claiming a finite-sample guarantee, since exchangeability across cohorts is not established.

3.4. Explanation interface

The interface generates wording only after the numerical predictions have been fixed. Its four rule-based templates use de-identified fields:

- The case note states the predicted median and calibrated interval and can suggest no more than two observable actions.
- The uncertainty note uses interval width to indicate uncertainty. If that width is above the 75th percentile of raw interval widths in the corresponding calibration cohort, the note avoids a strong recommendation.
- The refusal template identifies which required field is missing when the current score, historical attendance, or historical homework is unavailable. It offers no explanation for the missing information.
- The screening summary gives counts and proportions only. It assigns a high flag when the calibrated lower bound is below 60 on the percentage scale and a moderate flag when that bound is below 70.

The templates must reproduce P10, P50, and P90 without alteration and exclude names as well as student, class, and teacher codes. They must also avoid attributing outcomes to motivation, family circumstances, or psychological causes. We use 60 and 70 as screening thresholds to describe the resulting workload, rather than as institutional pass marks. The evaluation checks compliance with these

template rules; it does not assess the instructional value of advice generated by a large language model.

4. Experimental Design

We analyze de-identified English writing assessments from ten undergraduate cohorts entering the School of Mathematical Sciences at Inner Mongolia University between 2016 and 2025. Direct identifiers were removed before model development. The dataset comprises 14,483 rows for 2,806 students in mathematics, computational science, and statistics, with 54 class codes and 18 teacher codes. Neither of the two school codes is included among the predictors. Assessment dates extend from September 2016 to May 2026 because follow-up for the 2025 cohort continues beyond its entry year. Of the students, 610 contribute four unit rows, 1,133 contribute five, and 1,063 contribute six. Missing values occur in 1,009 rows for writing scores, 623 for attendance, 733 for homework completion, and 1,456 for initial level. Table I summarizes the assessed units, full marks, student counts, and mean percentage scores based on observed records for each entering cohort. Cohort size ranges from 121 students in C2018 to 372 in C2024; mean scores range from 65.8 to 74.3. Thus, a calibration correction may be transferred between cohorts whose means differ by several points. The largest adjacent decrease is from 74.3 in C2023 to 67.4 in C2024, when the full mark first changes to 150.

TABLE I. Entering Cohorts in the Analytical File

Cohort	Students	Units	Full mark	Mean (%)
C2016	228	1-5	100	65.8
C2017	264	1-6	100	71.1
C2018	121	1-4	100	66.6
C2019	315	1-5	100	71.8
C2020	287	1-6	100	69.3
C2021	244	1-5	100	74.2
C2022	348	1-4	100	68.9
C2023	296	1-6	100	74.3
C2024	372	1-5	150	67.4
C2025	331	1-6	150	71.9

For each test year Y between 2019 and 2025, we train on cohorts entering before $Y-1$, calibrate on cohort $Y-1$, and evaluate on cohort Y . Feature-imputation medians and standardization parameters are estimated anew from that transfer's training rows. Table II lists the available samples.

The number of training forecasts increases from 2,011 in the first transfer to 8,007 in the last. C2018 provides the calibration sample for the 2019 test cohort and contains 121 students; subsequent calibration cohorts each include at least 244 students. In four test years, at least one assessed unit has no counterpart in the calibration cohort.

TABLE II. Sample Used in Each Transfer

Test year	Train rows	Cal. rows	Cal. students	Test rows	Missing units
2019	2,011	332	121	1,127	5
2020	2,343	1,127	314	1,295	6
2021	3,470	1,295	287	908	—
2022	4,765	908	244	969	—
2023	5,673	969	348	1,365	5, 6
2024	6,642	1,365	296	1,341	—
2025	8,007	1,341	371	1,507	6

We measure point-prediction accuracy by MAE across forecast rows. Interval quality is coverage and mean width on those rows. Because a student may contribute multiple forecasts, the coverage denominator counts forecasts rather than independent students. The seven test sets stay separate. We therefore report coverage for each test cohort rather than combining the seven rates.

5. Results

5.1 Point error

Absolute errors for the successive test cohorts are listed in Table III. Last-score MAE stays between

TABLE III. Point Error on Each Next Cohort

Test year	N	Last MAE	Linear MAE	LightGBM MAE
2019	1,127	5.57	4.88	4.90
2020	1,295	5.63	4.96	5.05
2021	908	5.68	5.15	5.30
2022	969	5.58	4.92	4.97
2023	1,365	5.40	4.81	4.93
2024	1,341	5.43	4.83	4.80
2025	1,507	5.53	4.88	4.91

A student-clustered bootstrap, with 2,000 resamples of the fixed test predictions, puts a 95% interval on the MAE difference. In every year, both fitted medians minus the previous score lie entirely below zero. LightGBM minus the linear quantile lies entirely above zero in 2020–2023, so the linear fit is better in those four years. The same interval contains zero in 2019, 2024, and 2025. In no test year does the interval lie wholly below zero in favor of LightGBM. These bootstrap intervals reflect resampling of students with predictions held fixed, without retraining either model.

5.40 and 5.68. Linear quantile P50 is 4.81–5.15. LightGBM P50 is 4.80–5.30; it is lower than the linear median only in 2024, by 0.03 points. The linear reduction from the previous score is 0.53–0.69 points. For the linear model, RMSE ranges from 6.06 to 6.48 and R^2 from 0.50 to 0.57; the corresponding LightGBM ranges are 6.16–6.71 and 0.49–0.56. Last-score RMSE is 6.82–7.21 and R^2 is 0.36–0.46. Pinball loss at the median is 2.40–2.58 for the linear quantile and 2.40–2.65 for LightGBM.

5.2 Score history

To examine the contribution of feature groups, we repeated the transfers using mean-regression LightGBM and omitted one block at a time. With all features included, MAE ranges from 4.84 to 5.21. Removing the current score together with all score-history variables increases MAE to 6.39–7.42 and reduces R^2 to 0.05–0.24. By comparison, removing attendance and homework yields an MAE of 4.83–5.20, while excluding current- and next-unit difficulty yields 4.84–5.41. Score history therefore provides most of the predictive information in this dataset.

5.3 Interval coverage

Coverage and mean interval width are reported together in Table IV. Before calibration, LightGBM covers 69.4%–76.7% of outcomes with widths of 12.9–14.6, falling short of the nominal 80% in every test year. Pooled calibration supplies a correction for all target units. For the six transfers with at least 244 students in the calibration cohort, it produces coverage of 77.8%–82.0% and widths of 15.1–16.6. In the 2019 transfer, calibration instead relies on 332 forecasts from C2018: coverage reaches 83.8%, with a mean width of 17.2. Estimating corrections separately by unit changes LightGBM coverage by no more than 1.7 percentage points relative to pooling. The largest difference occurs in 2022, when

TABLE IV. Coverage (%) and Mean Width on the Percent Scale

Test year	Lin. raw	Lin. pooled	LGB raw	LGB pooled
2019	80.1/15.6	86.6/17.7	71.0/12.9	83.8/17.2
2020	80.2/15.6	81.1/15.9	70.0/12.9	78.8/15.8
2021	77.5/15.3	76.8/15.1	69.4/13.7	77.8/15.9
2022	79.1/16.0	80.2/16.7	73.2/13.9	81.0/16.6
2023	77.7/15.4	79.1/15.9	74.7/14.0	81.2/16.5
2024	81.1/16.0	81.8/16.4	76.7/14.6	82.0/16.4
2025	79.9/15.5	78.7/15.1	75.9/14.1	79.3/15.1

In 2019, 2020, 2023, and 2025, at least one test unit is absent from the calibration cohort, so the horizon rule uses the pooled quantile on that unit. Table IV retains these forecasts, with the pooled correction replacing an unavailable unit-specific residual quantile.

5.4 Subgroups and screening cuts

Within each test year, the three majors differ by at most 0.49 MAE points (computational science in 2021). Across majors, the pooled LightGBM interval covers 77.2%–85.8% of outcomes, and the major with the lowest coverage changes between years. Grouping students by initial-level quartiles defined within each test year also produces no consistent

TABLE V. Screening with the Pooled LightGBM Lower Bound

Cut	Flagged (%)	Precision (%)	Recall (%)
60	22.4–51.3	16.3–33.6	87.8–95.2
70	72.9–92.3	35.4–61.7	97.4–99.5

6. Discussion

Six of the nine steps between adjacent cohort means in Table I exceed 4.5 points on the percent scale. The steps are +5.3, –4.5, +5.2, –2.5, +4.9, –5.3, +5.4, –6.9, and +4.5, from C2016–C2017 through

pooled calibration gives 81.0% coverage and width 16.6, compared with 82.7% and 17.3 for unit-specific calibration.

Coverage of the uncalibrated linear interval also varies around the 80% target. It covers 80.1%, 80.2%, and 81.1% in 2019, 2020, and 2024, and 77.5%–79.9% in the other four years, at widths 15.3–16.0. For the linear model, pooled calibration sometimes increases the distance from nominal coverage: in 2021, coverage decreases from 77.5% to 76.8%. In 2025 it moves coverage from 79.9% to 78.7%. Using the same small C2018 calibration sample, the pooled linear interval reaches 86.6% coverage and a mean width of 17.7 in 2019.

ordering. The lowest quartile coverage is 71.8%, for the third quartile in 2021. Forecasts with a missing initial level are not uniformly worse.

Table V uses the pooled LightGBM lower bound on the percent scale. A forecast is flagged when that bound falls below the cut, and the event is a target score below the same cut. Recall is high because the interval is about 16 points wide. Precision stays low because only 4.7%–18.3% of scores are below 60, while the 60-point rule flags 22.4%–51.3% of forecasts. The 70-point rule flags 72.9%–92.3%. These cuts describe a screening workload. They do not estimate the probability that a student will score below the cut.

C2024–C2025. Several of those shifts are about as large as the MAE in Table III. The calibration cohort is therefore not a small perturbation of the test cohort. A correction that was exact for one entering class can be wide, or slightly off center, for the next class. The

ranges in Table IV are the empirical size of that mismatch.

The sharpest drop is from 74.3 in C2023 to 67.4 in C2024. C2024 is also the first cohort scored out of 150, after conversion to a percent of full marks. Models for that transfer are trained on C2016–C2022 and calibrated on C2023. Pooled LightGBM then covers 82.0% of C2024 outcomes at a mean width of 16.4. The following transfer, into C2025, covers 79.3% at width 15.1. One year with a new full mark does not by itself set the width for every later year, and it does not leave coverage at the raw-quantile level of 76.7%.

If a student has no record for an intervening unit, we predict the next unit actually recorded for that student. Horizon-specific calibration is indexed by

this target unit. A unit found only in the test cohort has no calibration residuals from which to estimate its own correction, so we apply the pooled quantile. This fallback allows the evaluation to retain those forecasts and provides a correction for every test unit.

Table VI compares pooled and unit-specific calibration for both models. For LightGBM, the largest differences are 1.7 percentage points in coverage and 0.7 points in mean width. Unit-specific calibration is not consistently closer to 80% for the linear model: coverage changes from 76.8% to 77.3% in 2021 and from 78.7% to 79.0% in 2025. Estimating an additional quantile from the observations for one unit within a single cohort does not yield shorter intervals here.

TABLE VI. Pooled and Unit-Specific Intervals

Test year	LGB pool	LGB unit	Lin. pool	Lin. unit
2019	83.8/17.2	84.6/17.4	86.6/17.7	86.6/18.1
2020	78.8/15.8	79.9/16.1	81.1/15.9	81.1/15.9
2021	77.8/15.9	78.3/16.2	76.8/15.1	77.3/15.5
2022	81.0/16.6	82.7/17.3	80.2/16.7	81.2/17.1
2023	81.2/16.5	81.5/16.6	79.1/15.9	79.6/16.0
2024	82.0/16.4	81.4/16.2	81.8/16.4	81.4/16.2
2025	79.3/15.1	78.2/15.2	78.7/15.1	79.0/15.4

6.1 What is refit

We fit the models independently for each of the seven transfers. For test year Y , only forecast rows from cohorts entering before $Y-1$ are used to estimate the linear model's imputation medians, standardization parameters, and pinball-loss coefficients. We also retrain LightGBM on those rows, retaining the settings in Section III-B and random seed 20260920. Calibration then uses cohort $Y-1$ alone. Test-year observations play no part in fitting the models or estimating the conformal correction.

7. Limitations

The data cover one writing course at a single college, so performance under another curriculum remains untested. Calibration for the first transfer relies on the small C2018 cohort. Differences in the units assessed across adjacent cohorts further limit

where unit-specific corrections can be estimated. Reported coverage concerns forecast rows and provides no guarantee for any particular student. Converting scores to percentages makes the 100- and 150-point scales numerically comparable but cannot account for changes in grading standards. We checked the explanation layer against its template rules; teachers did not evaluate it, and the study measured no effect on learning outcomes.

8. Conclusion

The seven successive cohort transfers show how performance varies beyond a single held-out year. In each transfer, earlier cohorts provide the training data, the immediately preceding cohort provides calibration residuals, and the next cohort is evaluated once, with students kept in separate roles. Both quantile models achieve next-unit MAE of approximately 4.8–5.3 on the percentage scale,

compared with approximately 5.4–5.7 when the previous score is used directly. The linear median reduces MAE by 0.53–0.69 points. In the mean-regression LightGBM ablations, removing the current score and score history eliminates the improvement over the previous-score baseline. Removing attendance or homework has little effect, while the effect of excluding unit difficulty varies; school codes are not included as predictors. The student-clustered bootstrap favors the linear median over LightGBM in 2020–2023. For 2019, 2024, and 2025, the confidence interval for the MAE difference includes zero. None of the seven intervals establishes lower MAE for LightGBM. These comparisons condition on the fitted models, since bootstrap resampling does not involve retraining.

Without calibration, LightGBM's P10–P90 intervals cover 69.4%–76.7% of outcomes at mean widths of 12.9–14.6, below the nominal 80% throughout. Pooled conformal calibration gives 77.8%–82.0% coverage in the six transfers calibrated on at least 244 students, with widths of about 15–17 points. The first transfer, calibrated on 121 students, reaches 83.8% coverage at width 17.2. Unit-specific corrections differ from pooled calibration by at most 1.7 percentage points in coverage and do not produce shorter intervals. The uncalibrated linear intervals range from 77.5% to 81.1% coverage, and pooling does not bring every transfer closer to 80%. Cohort means also change by several points, including the decrease from 74.3 to 67.4 alongside the switch from a 100-point to a 150-point full mark. These shifts help explain why a correction estimated for one entering cohort may perform differently in the next, although interval width alone does not isolate the effect of cohort differences.

Screening at a calibrated lower bound of 60 flags roughly a quarter to a half of forecasts and identifies most outcomes below 60, but precision is only 16%–34%. A threshold of 70 flags most forecasts. These rates indicate the workload associated with the screening rules; they are neither individual probabilities of scoring below a threshold nor institutional pass criteria. The explanation interface

reproduces the median and interval while excluding names, class codes, and teacher codes from its wording. Whether instructors benefit from these explanations remains untested. Across the seven transfers, the empirical result is an MAE near 5 points and calibrated intervals about 15–17 points wide, with coverage varying by cohort around the 80% target.

Data Availability

The student-level records are not publicly released. They contain course assessments from one college in which direct names have been replaced with codes. Class and teacher codes remain in the file but are omitted from the predictors. Public summaries comprise the cohort sizes, mean percentage scores, and tables reported here. The supplementary material provides code for the seven transfers, including the LightGBM settings listed in Section III-B, together with a separately generated simulated example. The simulated records do not represent a subsample of the students in this study.

Conflict of Interest

The authors declare that there are no conflicts of interest to this work.

Acknowledgement

This research was funded by: 2026 Annual Research Cultivation Project of Pioneer College of Inner Mongolia University.² The 13th Batch China Foreign Language Education Fund Project of Foreign Language Teaching and Research Press

References

- R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.
- V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.
- Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in *Advances in Neural Information Processing Systems* 32, 2019.
- N. Meinshausen, "Quantile regression forests," *J. Mach. Learn. Res.*, vol. 7, pp. 983–999, 2006.
- G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A

- highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems* 30, 2017.
- J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, “Distribution-free predictive inference for regression,” *J. Amer. Statist. Assoc.*, vol. 113, no. 523, pp. 1094–1111, 2018.
- R. J. Tibshirani, R. Foygel Barber, E. J. Candès, and A. Ramdas, “Conformal prediction under covariate shift,” in *Advances in Neural Information Processing Systems* 32, 2019.
- R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, “Conformal prediction beyond exchangeability,” *Ann. Statist.*, vol. 51, no. 2, pp. 816–845, 2023.
- I. Gibbs and E. J. Candès, “Adaptive conformal inference under distribution shift,” in *Advances in Neural Information Processing Systems* 34, 2021.
- A. Namoun and A. Alshanqiti, “Predicting student performance using data mining and learning analytics techniques: A systematic literature review,” *Appl. Sci.*, vol. 11, no. 1, Art. no. 237, 2021.

How to Cite: Da, G. & Zhao, J. (2026). Next-Unit English Writing Scores under Repeated Cohort Transfer. *Journal of Global Humanities and Social Sciences*, 7(5), 316-323 <https://doi.org/10.61360/BoniGHSS262020860504>