International Journal of Engineering, Science and Environment
Peer-Reviewed Open Access • CC BY 4.0 • Quarterly
ijese.in

e-ISSN: 2839-5360
Vol. 1 | Issue 2 | Jul–Sep 2026
Long & Ramachandran | Calibration-Aware Evaluation of ML Models

RESEARCH ARTICLE
Received: 13 August 2026 | Accepted: 21 September 2026 | Published: 30 September 2026

Calibration-Aware Evaluation of Machine Learning Models for Survey-Based Prediction

Gabriel Long[1], **Sridhar Ramachandran**[2*]

[1] *Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, Indiana, USA*

[2] *School of Natural Sciences, Indiana University Southeast, New Albany, Indiana, USA*

***Corresponding Author:** sriramac@iu.edu

ABSTRACT

We evaluated whether school-property and electronic bullying indicators improve prediction of persistent sadness or hopelessness beyond a demographic baseline in the National Youth Risk Behavior Survey (YRBS). Baseline and expanded histogram gradient boosting models shared preprocessing and hyperparameters; evaluation combined discrimination, Brier score, calibration diagnostics, and paired uncertainty. Selection was retrospective, so the 2023 held-out partition supports internal comparison rather than an untouched final test. The selected Platt-calibrated expanded model achieved internal ROC AUC 0.716, average precision 0.644, and Brier score 0.206. Compared with the matched baseline, improvements were 0.066, 0.106, and 0.019, respectively, with conditional 95% intervals above zero. Applying the frozen pipelines to the independent 2021 wave preserved the incremental pattern, with expanded-model ROC AUC 0.716 and Brier score 0.205. Weighting sensitivities supported the comparison, while subgroup calibration varied. The findings illustrate calibration-aware feature-set evaluation and remain non-causal and research-only; they do not establish fairness or readiness for operational use.

Keywords: probability calibration, Brier score, feature-set comparison, Youth Risk Behavior Survey, bullying, cross-wave validation

1. INTRODUCTION

Applied predictive studies often ask whether a bounded set of additional measurements contributes useful information beyond a defined baseline. In applied computing with survey data, a defensible comparison should hold the analytic sample, preprocessing, and model-development boundary constant while changing the feature set under evaluation. It should also assess more than one dimension of predictive performance [1], [2]. This study illustrates that evaluation problem using public health survey data; it does not evaluate an engineering system or establish performance in an engineering setting.

Probability quality is especially important when a model reports estimated risks rather than only class labels. Discrimination describes how well a model ranks positive observations above negative observations. Calibration describes how closely predicted probabilities agree with observed outcome frequencies [3], [4]. The Brier score summarizes mean squared probability error, so lower values indicate more accurate probability estimates. A model can rank observations well and still produce probabilities that are systematically too high, too low, or too extreme.

We study this evaluation problem using the 2023 National Youth Risk Behavior Survey (YRBS). The YRBS is a documented public-use surveillance system for U.S. high school students [5], [6]. We use QN26, persistent sadness or hopelessness, as the binary outcome. The additional predictors are QN24, bullied on school property, and QN25, electronically bullied. We compare models using age, sex, grade, and race or ethnicity with otherwise matched models that also receive QN24 and QN25. The two bullying variables are grouped only by the experimental design. The algorithms receive them as separate input variables.

The substantive case is useful because bullying victimization and adolescent mental health are well-established areas of concern [7], [8], [9], [10], [11]. The methodological question is narrower. We ask whether the two bullying indicators improve prediction beyond the specified demographic baseline and whether the resulting probabilities remain well calibrated. We do not estimate causal effects, diagnose students, recommend intervention assignment, or claim that the fitted model is ready for deployment.

The study combines internal evaluation in the 2023 YRBS with a separate cross-wave evaluation. The fixed 2023-trained pipeline was applied to the compatible 2021 National YRBS without tuning, preprocessing refit, or recalibration [12], [13]. Because that wave precedes the development wave, this is reverse-time cross-wave evaluation rather than prospective temporal validation.

We therefore address two questions. First, do QN24 and QN25 add incremental predictive value for the persistent-sadness-or-hopelessness outcome beyond the demographic baseline? Second, does the calibration-aware feature-set result remain evident when the locked 2023 pipeline is applied to an independent compatible YRBS wave?

2. LITERATURE REVIEW

2.1 Bullying and adolescent mental health

The 2023 YRBS literature provides direct substantive motivation for evaluating bullying indicators alongside persistent sadness or hopelessness. Young et al. reported that bullying victimization, frequent social media use, persistent sadness or hopelessness, and suicide risk were meaningfully related in the national 2023 survey [7]. A meta-analysis of 31 studies likewise found a strong association between bullying

How to cite: G. Long and S. Ramachandran, "Calibration-Aware Evaluation of Machine Learning Models for Survey-Based Prediction," Int. J. Eng. Sci. Environ., vol. 1, no. 2, pp. 1–10, 2026.
Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. Licence: https://creativecommons.org/licenses/by/4.0/

victimization and depressive symptoms in children and adolescents [8].

School-based and electronic bullying also warrant separate measurement. Cross-national evidence links cyberbullying victimization with serious adolescent mental-health outcomes [9]. U.S. research characterizes cyberbullying as a meaningful source of psychological harm rather than a trivial extension of offline conflict [10]. Recent analysis of the 2023 U.S. YRBS also reported a strong association between school bullying and suicide attempts [11]. These studies support the substantive relevance of the two indicators. They do not establish that the variables will add held-out predictive information after a baseline feature set is fixed.

2.2 Machine learning with survey and human-centered data

Machine learning has increasingly been applied to survey and administrative data for human-centered prediction. A 2024 systematic review found a growing literature that uses machine learning with administrative and survey data to predict suicidal thoughts and behaviors, while also identifying substantial variation in design, validation, and reporting practices [14]. Large student surveys have also been used to identify predictive correlates of depression. Gohari et al. applied random forests to self-reported data from more than 56,000 Canadian students and reported useful discrimination for clinically relevant depression symptoms [15].

Other work has emphasized replication across populations. Rothenberg et al. combined theory-driven and data-driven methods to predict adolescent mental-health outcomes across multiple cultural settings [16]. Kim et al. developed substance-use prediction models using the Korea Youth Risk Behavior Web-Based Survey and externally evaluated them with U.S. YRBS and Norwegian survey data [17]. Hill et al. used questionnaire and neuroimaging data from more than 11,000 participants in the Adolescent Brain Cognitive Development study and found that questionnaire-based information carried substantial predictive signal [18]. Together, these studies show that structured human-centered datasets can support predictive modeling. They also illustrate why validation design and the meaning of model outputs must be stated precisely.

The present study differs from broad feature-discovery or cross-dataset benchmarking work. We do not search the YRBS for the largest possible predictor set. We test one researcher-specified comparison. The experimental object is the incremental value of a researcher-specified feature group under a stable data and evaluation protocol.

2.3 Calibration-aware model evaluation

Prediction studies require performance evidence that is distinct from association evidence. Transparent development, internal validation, and separation of independent evaluation data are central to credible prediction research [1]. Classification metrics also answer different questions, so a single accuracy or area-under-the-curve value is insufficient for a complete evaluation [2].

Calibration becomes central when probabilities are interpreted. Huang et al. distinguish discrimination from calibration and show why numerically reliable probability estimates require direct assessment [3]. Silva Filho et al. provide a broader review of classifier calibration, proper scoring rules, calibration diagnostics, and post hoc correction methods [4]. These principles apply beyond clinical prediction. Any engineering or data-science workflow that treats model outputs as probabilities should determine whether those probabilities are appropriately scaled on unseen data.

This literature motivates the evaluation used here. We report ROC AUC for ranking performance and average precision for precision-recall performance. We use Brier score to evaluate average probability error. We report calibration slope and intercept to assess probability scaling and centering. We also quantify paired feature-set improvements, examine survey-weight sensitivity, and evaluate the frozen selected pipeline on an independent YRBS wave.

2.4 Research gap

The literature already establishes that bullying is associated with adolescent distress and that machine learning can extract predictive signal from large human-centered datasets. The remaining gap is methodological. Researchers need compact protocols for testing whether a bounded, theoretically motivated feature group improves prediction beyond a transparent baseline without conflating feature discovery, model tuning, calibration, and final evaluation.

We address that gap with a controlled feature-set comparison that treats calibration as a central part of the evidence. The empirical findings concern YRBS respondents. Adapting the evaluation approach to other applied-computing or engineering settings would require domain-specific outcomes, data, and independent validation.

3. METHODOLOGY

3.1 Data sources and analytic samples

The development analysis used the 2023 National YRBS public-use data released by the Centers for Disease Control and Prevention (CDC) [5], [6]. The source file contained 20,103 questionnaire records. We excluded 240 rows without usable QN26 outcome coding before model comparison, yielding 19,863 observations with 8,108 positive and 11,755 negative outcomes. The analysis inputs were verified against the official 2023 fixed-width ASCII source for age, sex, grade, QN24, QN25, QN26, respondent weight, stratum, primary sampling unit, and race or ethnicity. All 20,103 records matched on every checked field.

The independent validation analysis used the 2021 National YRBS [12], [13]. The raw national file contained 17,232 records, of which 16,961 had usable coding for the persistent-sadness-or-hopelessness outcome. The 2021 external sample included 6,749 positive outcomes. We harmonized the outcome, school-property bullying, electronic bullying, age, sex, grade, and race or ethnicity by actual question wording and response meaning rather than by variable number because the YRBS item numbers changed between waves. The relevant outcome and bullying questions retained the same 12-month recall frame and Yes/No response structure. Exact item

© 2026 The Author(s) Licensed under CC BY 4.0 | Published by Engineering, Science and Environmental Research Publications | e-ISSN 2839-5360 | ijese.in

wording and response options were checked against the official national questionnaires [19], [20].

The National YRBS uses a complex multistage cluster sampling design and supplies respondent weights, strata, and primary sampling unit (PSU) identifiers [5], [6], [13]. CDC weights were used for population-description estimates and respondent-weight sensitivity analyses. Primary predictive estimators were fitted without CDC survey weights because the objective was a controlled individual-level comparison of baseline and expanded predictors among observed respondents, rather than estimation of population prevalence or a design-weighted prediction objective. This choice prioritizes comparability and does not establish population-representative predictive performance. Survey weighting is distinct from the class-balanced weights used by the logistic reference model. Respondent weighting alone does not provide complete design-based inference for stratification and clustering. Table I summarizes the study design and evaluation boundaries.

3.2 Outcome and researcher-specified feature sets

The 2023 binary outcome was QN26, persistent sadness or hopelessness. The demographic baseline contained age, sex, grade, and race or ethnicity, represented in the analysis files as q1, q2, q3, and raceeth. The expanded feature set contained the same baseline plus QN24 and QN25. The corresponding 2021 items were mapped to the same constructs after the cross-wave wording audit.

We define the two bullying indicators as a researcher-specified feature group for comparison purposes. This grouping belongs to the experimental design, not to the learning algorithm. The models receive the variables individually. We estimate their combined incremental contribution by comparing otherwise matched models with and without those two inputs.

The baseline was intentionally narrow. It provides a transparent demographic comparator and preserves a clear interpretation of the incremental comparison. A broader baseline could be scientifically useful, but it would answer a different question. The analysis does not claim that the bullying indicators remain incrementally predictive after every plausible covariate expansion.

Table I. Study design and evaluation boundaries

Component	Specification	Analytic role
Development data	2023 National YRBS	Model development and internal evaluation
Outcome	Persistent sadness or hopelessness	Binary target
Baseline	Age, sex, grade, race or ethnicity	Demographic comparator
Added feature group	School-property and electronic bullying	Researcher-specified comparison
2023 split	80/20, seed 2026	Fixed respondent-level training and internal evaluation partitions
Selection	Fixed training-side cross-validation	Brier score, calibration slope, then ROC AUC
External validation	2021 National YRBS	Frozen-pipeline reverse-time transportability
Claim boundary	Predictive, non-causal, research-only	Limits interpretation

3.3 Preprocessing and model families

Age, sex, grade, and race or ethnicity were treated as categorical variables. Missing categorical values were imputed with the most frequent category and then one-hot encoded with unseen-category handling. The binary bullying indicators were retained as numeric variables and missing values were median-imputed within the pipeline. All transformations were fitted on training data only. OneHotEncoder used handle_unknown="ignore", sparse_output=False, and drop="if_binary". The modal and median imputers were not survey-weighted, including in the weighted-estimator sensitivity.

We evaluated regularized logistic regression and histogram gradient boosting (HGB). Logistic regression served as a transparent linear reference and used the lbfgs solver, class-balanced weighting, and a maximum of 5,000 iterations. Class balancing gives the two outcome classes equal total fitting weight, changes the effective class prior, and can shift the calibration of raw probabilities. HGB provided controlled nonlinear flexibility. The retained HGB configuration used learning rate 0.01, maximum depth 5, a maximum of 600 boosting iterations, 15 maximum leaf nodes, 120 minimum samples per leaf, L2 regularization 0.01, early stopping, and random state 2026. No candidate family or HGB hyperparameter was changed in response to the 2023 held-out or 2021 external results. HGB hyperparameters had been tuned on the demographic baseline using a 12-draw randomized search scored by five-fold ROC AUC, then transferred unchanged to the expanded feature set. Logistic regression used default C=1. HGB early stopping used validation_fraction=0.1, n_iter_no_change=10, tol = 10^{-7}, and loss scoring; these internal splits also remained confined to the training data. Holding these baseline-tuned settings fixed makes the feature-set comparison more controlled: separately optimizing each model could confound the contribution of QN24 and QN25 with changes in model configuration or complexity. This approach prioritizes comparability and does not establish globally optimal hyperparameters for the expanded model.

3.4 Candidate selection, internal evaluation, and calibration

The 2023 analysis used a fixed individual-level, outcome-stratified 80/20 split with seed 2026. The training partition contained 15,890 observations and the held-out partition

© 2026 The Author(s) Licensed under CC BY 4.0 | Published by Engineering, Science and Environmental Research Publications | e-ISSN 2839-5360 | ijese.in

contained 3,973. Five stratified, shuffled folds with seed 2026 were fixed within training. Neither partitioning step enforced separation by PSU.

A diagnostic of the saved primary split found 82 distinct PSU codes in training and 81 in evaluation, with all 81 evaluation codes shared. Some codes recur across strata. Using stratum–PSU combinations, the counts were 89 in training, 88 in evaluation, and 88 shared. All 3,973 evaluation respondents therefore belonged to a stratum–PSU combination represented in training. Students within PSUs may share school-level or environmental characteristics, so cross-partition observations may not be fully independent. This clustered dependence can make internal performance optimistic relative to evaluation that separates entire clusters. It is distinct from directly using held-out features or outcomes to fit a model.

The original analysis compared calibration states and selected the headline model using 2023 held-out metrics. The same lexicographic selection rule was subsequently reapplied retrospectively to training-side cross-validation scores from nested calibration. This reanalysis does not erase the earlier consultation of held-out results or establish a completely untouched internal test. The four candidate definitions and calibration procedures were retained; the 2021 evaluation supplies separate cross-wave evidence.

The selection rule minimized cross-validation Brier score first, then minimized the absolute distance of calibration slope from 1, and then used ROC AUC as the final tie-breaker. The four fixed candidates were expanded logistic regression and expanded HGB without calibration, with Platt scaling, or with isotonic regression. Metrics were averaged across outer folds and compared at full numerical precision. This rule prioritizes probability quality; a very small numerical advantage alone does not establish practical benefit.

Calibration sensitivity considered no post hoc calibration, Platt scaling, and isotonic regression for the expanded HGB model. Probabilities were clipped to $[10^{-6}, 1 - 10^{-6}]$ before logit transformation. Platt scaling used logistic regression with $C = 10^6$, the lbfgs solver, and a maximum of 2,000 iterations on the logit of model probabilities. Isotonic regression used a monotone mapping with out-of-bounds values clipped. Final Platt fitting used out-of-fold predictions from the fixed five training folds. During training-side cross-validation, calibrated outer-fold metrics used a five-fold stratified, shuffled inner calibration layer with deterministic seeds 2026 + 500 + outer-fold index. Held-out and external outcomes were never used to fit a model or calibrator.

We evaluated ROC AUC, average precision, Brier score, calibration slope, and calibration intercept. Average precision was computed with sklearn.metrics.average_precision_score. Calibration slope and intercept were estimated jointly by logistic recalibration of the clipped probability logit using $C = 10^6$ and the lbfgs solver. A ten-bin equal-frequency reliability diagram was used for the selected 2023 model. For a prevalence-only reference, we applied the training outcome prevalence as one constant probability to every held-out observation. Brier skill was calculated as 1 minus the selected-model Brier score divided by this null Brier score.

3.5 Paired uncertainty, survey-weight sensitivity, and subgroup diagnostics

Feature-set uncertainty was quantified with a stratified paired bootstrap of the same 2023 held-out observations. Each replicate used one common set of sampled indices for the baseline and expanded models. We used 2,000 replicates with seed 20260915 and reported percentile 95% confidence intervals for ROC AUC change, average-precision change, and Brier improvement, defined as baseline Brier minus expanded-model Brier. Both uncalibrated and matched Platt-calibrated comparisons were evaluated.

Survey-weight sensitivity used the principal HGB baseline-versus-expanded comparison while holding observations, the frozen split, feature definitions, and hyperparameters constant. Models and Platt calibrators were fitted both without CDC respondent weights and with CDC respondent weights normalized to mean 1 within each fit. Each fit was evaluated with ordinary empirical metrics and CDC-weighted metrics. CDC respondent weights were not multiplied by class-balanced weights.

Subgroup diagnostics evaluated the selected expanded HGB model with Platt scaling in the 2023 held-out partition by sex, grade, race or ethnicity, and age. The adequacy rule required at least 100 positive and at least 100 negative QN26 outcomes within each evaluated category. The rule was used for this diagnostic analysis; its timing relative to result inspection is not independently established, so it is not described as prospectively specified. Counts were retained for all categories, while performance and calibration estimates were suppressed for categories failing either threshold.

Eligible categories were summarized with ROC AUC, average precision, Brier score, calibration slope, calibration intercept, and uncertainty intervals. Outcome-stratified subgroup bootstrap intervals for discrimination and Brier score used 500 replicates with base seed 20260916 plus 100 times the variable index and the category index. Calibration intervals were model-based, unpenalized binomial-GLM Wald intervals. This diagnostic assessment describes heterogeneity and does not certify algorithmic fairness.

All observation-bootstrap intervals condition on the fitted pipelines, fixed partition, and observed event counts. They omit model-development, candidate-selection, and full survey-design uncertainty. An additional paired Rao–Wu-type public-design sensitivity used 2,000 replicates (seed 20260918), sampling $n_h - 1$ public-use PSUs with replacement within each stratum and multiplying CDC weights by $n_h/(n_h - 1)$ times the PSU draw count. Replicates were formed on the full analytic sample and restricted to the evaluation domain. Its percentile limits are sensitivity intervals, not complete design-based confidence intervals, because the public data do not identify all multistage sampling probabilities or finite-population corrections. The fitted models were held fixed, and the random individual split permits shared PSUs across training and evaluation.

3.6 Cross-wave external validation

The complete 2023-trained baseline and expanded HGB pipelines, including preprocessing and Platt calibrators, were

frozen and applied to the 2021 National YRBS without hyperparameter tuning, preprocessing refit, or calibration refit. We reported ROC AUC, average precision, Brier score, calibration slope, calibration intercept, outcome prevalence, and confidence intervals. Paired incremental-value intervals used 2,000 stratified bootstrap replicates. Because 2021 precedes the development wave and was collected under different conditions, this analysis is described as cross-wave external validation or reverse-time transportability rather than prospective temporal validation [13]. The serialized fitted estimators and calibrators used for internal predictions were reused directly for external scoring, and prediction identity was checked.

3.7 Software and reproducibility

The analyses were reproduced with Python 3.13.12, scikit-learn 1.8.0, NumPy 2.4.2, pandas 3.0.0, SciPy 1.17.0, Matplotlib 3.10.8, PyArrow 23.0.0, openpyxl 3.1.5, joblib 1.5.3, and statsmodels 0.15.0. The reproducibility materials include dependency pins, fixed split and fold definitions, fitted pipelines, aggregate results, input hashes, and execution instructions. Calibrated cross-validation scores use the nested procedure described above. Raw CDC microdata are not redistributed. Fig. 1 summarizes the evaluation workflow.

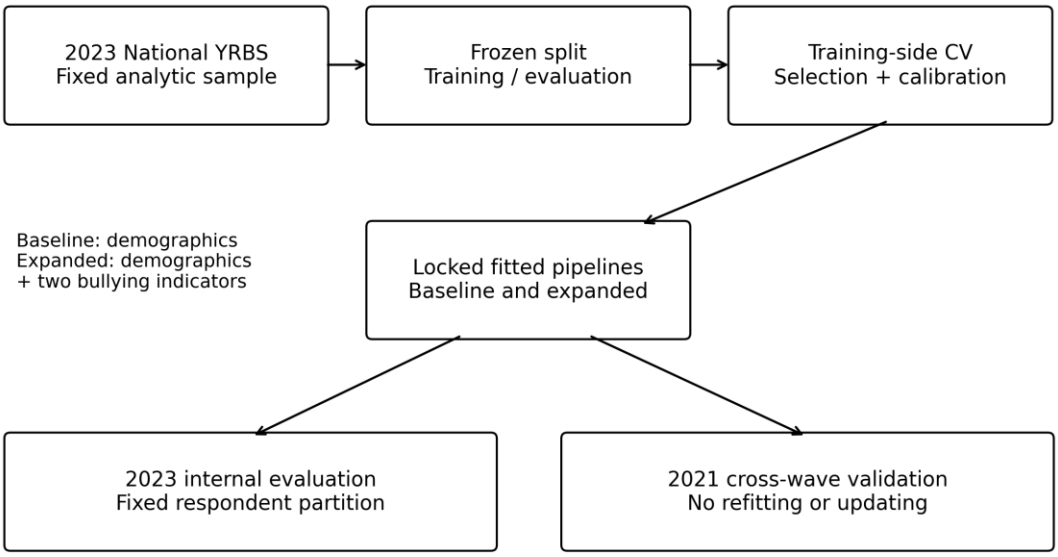


Fig. 1. Evaluation workflow. Calibration fitting uses training data only. The fixed pipelines are evaluated on the 2023 internal comparison set and the separate 2021 wave without external-wave tuning, preprocessing refit, or recalibration.

4. RESULTS AND DISCUSSION

4.1 Descriptive context

Descriptive estimates used each item's valid-response denominator in the national public-use data. In 2023, CDC-weighted prevalence was 39.7% for persistent sadness or hopelessness, 19.2% for school-property bullying, and 16.3% for electronic bullying. In the 2021 external sample, empirical outcome prevalence was 39.8% (ordinary Wilson 95% interval 39.1–40.5%) and CDC-weighted prevalence was 42.3%; no design-based interval is claimed for the latter. These contextual estimates are separate from model-performance estimates.

4.2 Training-side selection and 2023 internal evaluation

The training-side rule selected expanded HGB with Platt scaling, with mean cross-validation Brier score 0.207, calibration slope 0.984, and ROC AUC 0.716.

Table II reports the 2023 internal held-out comparison. The demographic-only logistic model achieved ROC AUC 0.650, average precision 0.547, and Brier score 0.232. The uncalibrated HGB demographic comparator produced a lower Brier score of 0.225. Adding the two bullying indicators improved both model families. The selected expanded HGB model with Platt scaling achieved ROC AUC 0.716, average precision 0.644, Brier score 0.206, calibration slope 1.005, and calibration intercept 0.014. A constant prediction using the training prevalence (6,486/15,890) gave a null Brier score of 0.242 on the same 3,973 held-out observations. The selected model's Brier skill score was 0.147, a 14.7% reduction in mean squared probability error relative to that reference.

Expanded logistic regression and expanded HGB produced almost identical 2023 ROC AUC values, but their probability estimates differed. The logistic reference used class-balanced fitting, which changes the effective class prior and can shift raw probability calibration. Its negative calibration intercepts and higher Brier scores should therefore not be interpreted as an inherent calibration disadvantage of logistic regression. The main feature-set comparison uses matched HGB pipelines.

© 2026 The Author(s) Licensed under CC BY 4.0 | Published by Engineering, Science and Environmental Research Publications | e-ISSN 2839-5360 | ijese.in

Table II. 2023 internal held-out model comparison

Model	Features	Calibration	ROC AUC	Avg. precision	Brier	Slope	Intercept
Logistic	Baseline	None	0.650	0.547	0.232	0.987	−0.364
HGB	Baseline	None	0.650	0.538	0.225	1.072	0.036
HGB	Baseline	Platt	0.650	0.538	0.225	1.033	0.024
Logistic	Expanded	None	0.716	0.643	0.213	0.974	−0.369
HGB	Expanded	None	0.716	0.644	0.206	1.046	0.028
HGB	Expanded	Platt	0.716	0.644	0.206	1.005	0.014
HGB	Expanded	Isotonic	0.714	0.619	0.207	0.980	−0.001

4.3 Incremental feature-set value and paired uncertainty

The main research question concerns the change produced by adding the two bullying indicators to the HGB demographic comparator. Under matched Platt calibration, the 2023 baseline achieved ROC AUC 0.650, average precision 0.538, and Brier score 0.225. The expanded model achieved ROC AUC 0.716, average precision 0.644, and Brier score 0.206.

The paired bootstrap shows that the internal improvements are not represented only by point estimates. ROC AUC increased by 0.066 (conditional percentile 95% CI 0.052–0.080), average precision increased by 0.106 (0.084–0.128), and Brier score improved by 0.019 (0.015–0.023). The retained uncalibrated comparison produced the same ROC AUC and average-precision changes and a nearly identical Brier-improvement interval. Table III reports the paired comparisons. Fig. 2 displays the internal and cross-wave increments.

Table III. 2023 paired feature-set improvements under matched calibration states

Calibration	Metric	Baseline	Expanded	Improvement	95% CI
None	ROC AUC	0.650	0.716	+0.066	0.052 to 0.080
None	Average precision	0.538	0.644	+0.106	0.084 to 0.128
None	Brier score	0.225	0.206	+0.019*	0.015182 to 0.023011
Platt	ROC AUC	0.650	0.716	+0.066	0.052 to 0.080
Platt	Average precision	0.538	0.644	+0.106	0.084 to 0.128
Platt	Brier score	0.225	0.206	+0.019*	0.015060 to 0.023110

*Brier improvement = baseline Brier − expanded-model Brier. Positive values indicate lower probability error. Intervals are conditional percentile bootstrap limits for fixed predictions and observed event counts; they omit fitting, selection and full survey-design uncertainty.

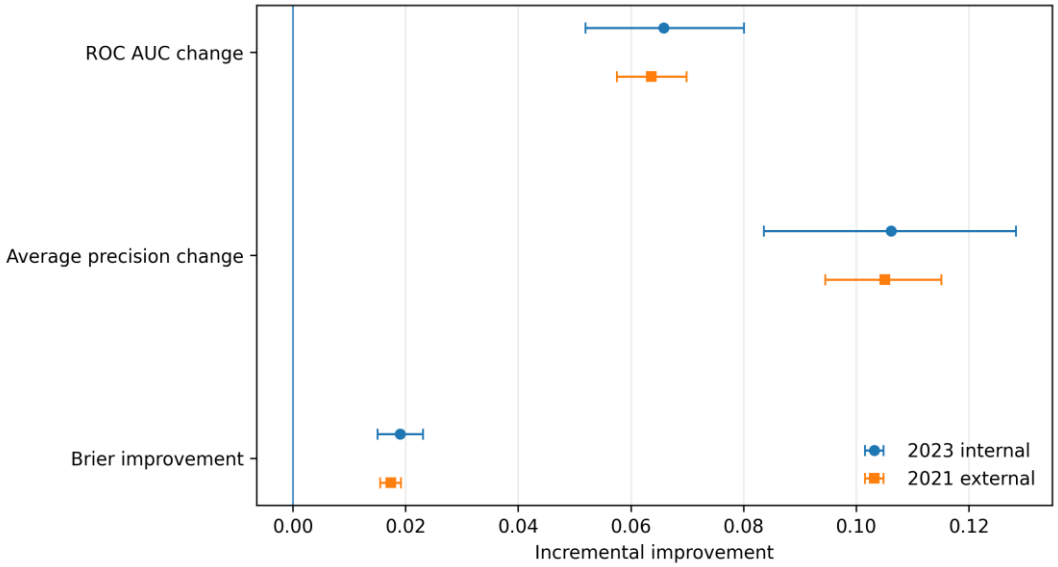


Fig. 2. Incremental feature-set improvements in the 2023 internal comparison and 2021 cross-wave external validation. Points show paired differences and bars show conditional 95% percentile bootstrap intervals. Brier improvement is baseline Brier minus expanded-model Brier.

© 2026 The Author(s) Licensed under CC BY 4.0 | Published by Engineering, Science and Environmental Research Publications | e-ISSN 2839-5360 | ijese.in

Table IV. 2023 calibration sensitivity for the expanded HGB model

Calibration	ROC AUC	Avg. precision	Brier	Slope	Intercept
None	0.716	0.644	0.206	1.046	0.028
Platt	0.716	0.644	0.206	1.005	0.014
Isotonic	0.714	0.619	0.207	0.980	−0.001

4.4 Calibration sensitivity and reliability

Platt scaling preserved ROC AUC (0.716) and average precision (0.644) relative to uncalibrated expanded HGB. Brier score changed from 0.206178 to 0.206122, a numerically very small improvement that does not by itself establish practical benefit. The calibration slope moved from 1.046 to 1.005, closer to 1, and the intercept changed from 0.028 to 0.014. These probability-scaling diagnostics, preserved discrimination, and the calibration-aware selection rule explain retaining Platt scaling. The main empirical finding remains the incremental predictive value of QN24 and QN25. Isotonic calibration produced an intercept close to zero but reduced ranking performance and had a higher Brier score than Platt scaling (Table IV).

The reliability diagram in Fig. 3 uses ten equal-frequency bins. The selected model remains close to the 45-degree line across the observed probability range, while local deviations remain visible. The purpose of the diagram is diagnostic rather than to imply perfect calibration.

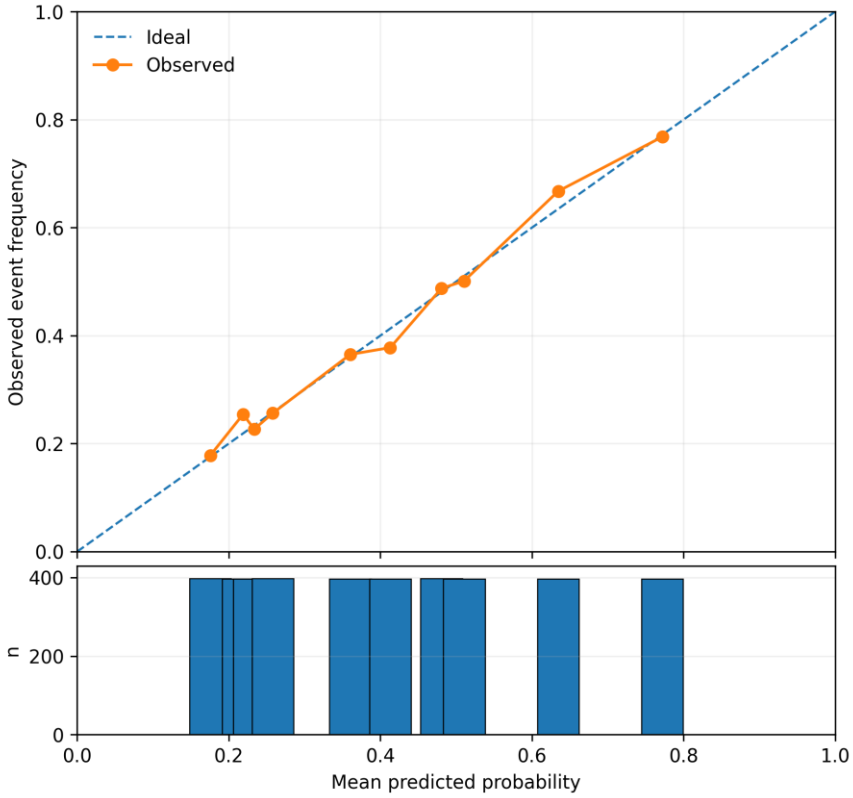


Fig. 3. Reliability diagram for the selected expanded HGB model with Platt scaling on the 2023 internal evaluation set. Ten equal-frequency bins show mean predicted probability and observed event frequency, with the lower panel showing the number of observations per bin.

4.5 Survey-weight sensitivity

The direction of the bullying-feature increment was stable under all four combinations of HGB model fitting and evaluation with or without CDC respondent weights. ROC AUC improvement ranged from 0.066 to 0.080, average-precision improvement ranged from 0.101 to 0.108, and Brier improvement ranged from 0.019 to 0.021. The full comparison is reported in the reproducibility supplement. These results show that the main feature-set result is not reversed by respondent-weight sensitivity. They do not constitute full design-based inference for the stratified clustered survey design.

The additional public-design sensitivity resampled within 16 strata containing 89 stratum–PSU combinations, with no singleton stratum. These combinations use 82 distinct PSU codes because some codes recur across strata. For the not-survey-weighted fitted models evaluated with CDC weights, the matched Platt comparison yielded a ROC AUC improvement of 0.073 (95% sensitivity interval 0.048–0.099), average-precision improvement of 0.101 (0.065–0.137), and Brier improvement of 0.020 (0.014–0.028). These results

© 2026 The Author(s) Licensed under CC BY 4.0 | Published by Engineering, Science and Environmental Research Publications | e-ISSN 2839-5360 | ijese.in

support the direction of the comparison under the stated cluster-resampling approximation, without claiming full design-based coverage. This resampling conditions on the existing split and fitted models. It does not remove the shared clusters between development and evaluation.

4.6 Subgroup calibration diagnostics

Among categories meeting the 100-event and 100-non-event rule, calibration slopes ranged from 0.939 to 1.127 for sex, 0.910 to 1.034 for grade, 0.977 to 1.030 for age, and 0.842 to 1.379 for race or ethnicity. Table V reports racial/ethnic subgroup counts and calibration estimates. This range shows that aggregate calibration close to 1 does not imply equivalent calibration across groups. Slopes below 1 can indicate predictions that are too extreme within a group, whereas slopes above 1 can indicate insufficient spread, interpreted jointly with the calibration intercept.

Three racial/ethnic categories had fewer than 100 events (4, 57, and 90), and the missing-race/ethnicity category had 23 events and 48 non-events. Their calibration estimates were not reported. Other sparse or missing categories likewise failed the rule. Intervals were wider in smaller eligible categories, and the range of point estimates is not a formal test of between-group differences. Full held-out counts, eligibility flags, and intervals are provided in the reproducibility materials. These findings describe diagnostic heterogeneity; they do not certify algorithmic fairness or establish equal performance or suitability across demographic groups.

Table V. Racial/ethnic calibration diagnostics for the selected expanded HGB model in 2023

Race or ethnicity	n	Events	Non-events	Slope	Intercept	Adequate
American Indian/Alaskan Native	270	136	134	1.375	0.206	Yes
Asian	203	57	146	—	—	No
Black or African American	331	120	211	1.379	−0.174	Yes
Native Hawaiian or Other Pacific Islander	17	4	13	—	—	No
White	1,929	779	1,150	0.930	0.048	Yes
Hispanic/Latino	243	90	153	—	—	No
Multiple–Hispanic/Latino	545	252	293	0.842	0.045	Yes
Multiple–Non-Hispanic/Latino	364	161	203	0.958	0.020	Yes
Missing	71	23	48	—	—	No

Adequate: at least 100 events and 100 non-events. Dashes indicate suppressed estimates. Categories follow CDC race/ethnicity coding [6]. Estimates are held-out diagnostic summaries, not a fairness assessment; subgroup intervals are provided in the reproducibility materials.

4.7 Cross-wave external validation

The frozen 2023-trained pipelines were applied to 16,961 eligible observations from the 2021 National YRBS. The 2021 baseline model achieved ROC AUC 0.652, average precision 0.524, Brier score 0.222, calibration slope 1.077, and calibration intercept 0.059. The frozen expanded model achieved ROC AUC 0.716, average precision 0.629, Brier score 0.205, calibration slope 1.030, and calibration intercept 0.072.

The incremental predictive pattern was retained on the independent wave. ROC AUC increased by 0.064 (conditional percentile 95% CI 0.058–0.070), average precision increased by 0.105 (0.095–0.115), and Brier score improved by 0.017 (0.016–0.019). CDC-weighted external evaluation produced ROC AUC 0.720, average precision 0.656, and Brier score 0.209 for the expanded model. Table VI compares the matched Platt-calibrated pipelines across waves.

Table VI. Matched 2023 internal and 2021 cross-wave performance for frozen Platt-calibrated HGB pipelines

Wave	Feature set	ROC AUC	Avg. precision	Brier	Slope	Intercept
2023 internal	Baseline	0.650	0.538	0.225	1.033	0.024
2023 internal	Expanded	0.716	0.644	0.206	1.005	0.014
2021 external	Baseline	0.652	0.524	0.222	1.077	0.059
2021 external	Expanded	0.716	0.629	0.205	1.030	0.072

The external result strengthens the evidence for transport of the feature-set comparison, but its interpretation remains bounded. The 2021 YRBS was collected in fall during the COVID-19 period rather than under the usual spring schedule [13], and the 2021 and 2023 national surveys differ in administration and sampling details [5], [13]. The external analysis therefore provides reverse-time cross-wave transportability evidence, not prospective temporal validation.

4.8 Methodological interpretation and scope

This YRBS analysis illustrates an applied-computing approach to assessing a bounded group of predictors: define a baseline, hold model-development choices constant, fit calibration

© 2026 The Author(s) Licensed under CC BY 4.0 | Published by Engineering, Science and Environmental Research Publications | e-ISSN 2839-5360 | ijese.in

within training data, quantify paired performance changes, and evaluate frozen pipelines on compatible independent data. The evidence concerns this survey application and its specific outcome.

The same evaluation principles may inform studies in engineering or other scientific domains, but their usefulness in a new setting must be established with domain-specific data and validation. No engineering measurements, systems, or outcomes were analyzed here, so the reported performance cannot be transferred to those settings.

The same distinction applies to decision-support research. Good calibration within an evaluated dataset does not establish causality, clinical validity, fairness, or deployment readiness. Operational use would require validation under the intended setting, explicit consequence and threshold analysis, governance controls, and monitoring for distribution and calibration shift.

4.9 Limitations

Several limitations define the scope of the findings. First, both YRBS waves are cross-sectional and self-reported. The outcome is a survey indicator of persistent sadness or hopelessness, not a clinical diagnosis. The bullying variables are reported experiences rather than verified longitudinal exposure histories. The results therefore support predictive association, not causal inference.

Second, the demographic baseline is deliberately narrow. This improves the interpretability of the feature-set comparison but limits claims about incremental value relative to richer behavioral, family, school, or contextual predictors. Third, the 2021 analysis is an earlier-wave transportability test rather than prospective temporal validation. Differences in collection timing, pandemic context, administration mode, and sampling remain possible sources of transport shift.

Survey design and cluster dependence limit the internal performance estimates. The individual-level split shares PSUs across training and evaluation, allowing respondents in both partitions to have common school-level or environmental characteristics. Internal performance may therefore be optimistic relative to evaluation on separated clusters. Primary unweighted training targets controlled respondent-level prediction, not population prevalence or fully design-weighted predictive performance. Respondent-weight and PSU-within-stratum sensitivities support the direction of the comparison but neither remove this dependence nor provide complete multistage design-based inference. Observation-bootstrap intervals condition on fixed predictions and event counts and omit fitting, selection, and full design uncertainty.

Model-development choices further constrain interpretation. HGB hyperparameters were tuned on the demographic baseline and transferred unchanged to preserve comparability; globally optimal expanded-model settings are not established. Retrospective training-side selection followed consultation of the 2023 held-out results, limiting claims of an untouched internal test. The 2021 evaluation provides separate reverse-time cross-wave evidence. The marginal Platt Brier reduction does not establish practical benefit by itself.

Subgroup calibration is heterogeneous. The racial/ethnic slopes of 0.842–1.379, wide intervals in smaller eligible categories, and categories failing the 100/100 rule limit subgroup interpretation. Average model calibration does not imply equal subgroup calibration. These diagnostics do not establish fairness, equal performance, or equal suitability across demographic groups.

The model is research-only and not validated for screening, diagnosis, clinical triage, intervention assignment, school or student ranking, disciplinary decisions, resource allocation, individual risk labeling, operational decision-making, or deployment. Concurrent self-reported survey associations do not determine causes or identify who would benefit from an intervention.

5. CONCLUSION

We evaluated whether two bullying indicators add predictive information for persistent sadness or hopelessness beyond a demographic baseline under a calibration-aware feature-set comparison. On the 2023 internal evaluation set, the indicators improved ROC AUC, average precision, and Brier score for matched HGB models, with conditional paired confidence intervals that excluded zero improvement.

The same incremental pattern was observed when the frozen 2023-trained pipelines were applied without tuning or recalibration to the independent 2021 National YRBS. The expanded external model achieved ROC AUC 0.716, average precision 0.629, Brier score 0.205, calibration slope 1.030, and calibration intercept 0.072. Respondent-weight and public-design sensitivities retained positive increments. Subgroup diagnostics documented variation in calibration and limited the interpretation.

The methodological contribution is the evaluation structure illustrated in this survey application. The evidence supports a reproducible comparison of a bounded feature group while assessing calibration, conditional uncertainty, weighting sensitivity, subgroup diagnostics, and cross-wave transportability. The findings remain predictive, non-causal, and research-only.

DECLARATIONS

Acknowledgement

The authors acknowledge the Centers for Disease Control and Prevention for making the 2021 and 2023 Youth Risk Behavior Survey public-use data and documentation available. This article develops work originating in Gabriel Long's 2026 undergraduate senior thesis [21], with subsequent revisions and extensions to the evaluation, sensitivity analyses, and cross-wave validation.

Author Contributions

Gabriel Long conducted the study development, software implementation, formal analysis, visualization, reproducibility work, and manuscript drafting and revision. Sridhar Ramachandran supervised the work, provided research and methodological guidance, and reviewed and edited the manuscript.

Data Availability

The 2021 and 2023 National YRBS public-use data are available from the Centers for Disease Control and Prevention at https://www.cdc.gov/yrbs/data/index.html. Scripts, aggregate outputs, subgroup diagnostics, sensitivity analyses, fitted pipelines, fixed split and fold definitions, dependency specifications, and reproduction instructions are available at https://github.com/gblongiu/yrbs-ijese-reproducibility (publication snapshot v1.0.0: https://github.com/gblongiu/yrbs-ijese-reproducibility/tree/v1.0.0). Raw CDC respondent-level microdata is not redistributed.

Funding

The authors received no specific funding for this work.

Ethics Statement

This study used publicly available, de-identified data from the 2021 and 2023 National Youth Risk Behavior Surveys. The authors had no direct interaction with participants and did not access identifiable private information. Separate institutional review board (IRB) approval and additional informed consent were not required for this secondary analysis of public-use data.

Competing Interests

The authors declare no conflict of interest.

REFERENCES

[1] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, "Developing clinical prediction models: A step-by-step guide," BMJ, vol. 386, p. e078276, 2024, doi: 10.1136/bmj-2023-078276.

[2] A. Tharwat, "Classification assessment methods," Applied Computing and Informatics, vol. 17, no. 1, pp. 168–192, 2021, doi: 10.1016/j.aci.2018.08.003.

[3] Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado, "A tutorial on calibration measurements and calibration models for clinical prediction models," Journal of the American Medical Informatics Association, vol. 27, no. 4, pp. 621–633, 2020, doi: 10.1093/jamia/ocz228.

[4] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: A survey on how to assess and improve predicted class probabilities," Machine Learning, vol. 112, no. 9, pp. 3211–3260, 2023, doi: 10.1007/s10994-023-06336-7.

[5] N. D. Brener, J. J. Mpofu, K. H. Krause, S. E. Jones, J. E. Thornton, Z. Myles, et al., "Overview and methods for the Youth Risk Behavior Surveillance System—United States, 2023," MMWR Supplements, vol. 73, no. 4, pp. 1–12, 2024, doi: 10.15585/mmwr.su7304a1.

[6] Centers for Disease Control and Prevention, "2023 YRBS Data User's Guide," 2024. Available: https://www.cdc.gov/yrbs/media/pdf/2023/2023_National_YRBS_Data_Users_Guide508.pdf (Accessed: Aug. 7, 2026).

[7] E. Young, J. L. McCain, M. C. Mercado, M. F. Ballesteros, S. Moore, L. Licitis, et al., "Frequent social media use and experiences with bullying victimization, persistent feelings of sadness or hopelessness, and suicide risk among high school students: Youth Risk Behavior Survey, United States, 2023," MMWR Supplements, vol. 73, no. 4, pp. 23–30, 2024, doi: 10.15585/mmwr.su7304a3.

[8] Z. Ye, D. Wu, X. He, Q. Ma, J. Peng, G. Mao, et al., "Meta-analysis of the relationship between bullying and depressive symptoms in children and adolescents," BMC Psychiatry, vol. 23, p. 215, 2023, doi: 10.1186/s12888-023-04681-4.

[9] M. Grimland, Y. Mori, S. Lesinskiene, L. Li, S. H. Ong, S. K. Praharaj, et al., "Cyberbullying victimization and suicide attempt among adolescents: A cross-national comparison," International Journal of Environmental Research and Public Health, vol. 22, no. 3, p. 385, 2025, doi: 10.3390/ijerph22030385.

[10] S. Hinduja and J. W. Patchin, "Cyberbullying through the lens of trauma: An empirical examination of US youth," BMC Public Health, vol. 25, p. 1709, 2025, doi: 10.1186/s12889-025-22692-6.

[11] K. R. Agarwal, S. A. Fleary, P. Kreniske, and C. A. Teasdale, "Bullying and suicide attempts among US high school students," JAMA Network Open, vol. 9, no. 1, p. e2552089, 2026, doi: 10.1001/jamanetworkopen.2025.52089.

[12] Centers for Disease Control and Prevention, "2021 YRBS Data User's Guide," Apr. 2023. Available: https://www.cdc.gov/yrbs/files/2021/pdf/2021_YRBS_Data_Users_Guide_508.pdf (Accessed: Sep. 15, 2026).

[13] J. J. Mpofu, J. M. Underwood, J. E. Thornton, N. D. Brener, A. Rico, G. Kilmer, et al., "Overview and methods for the Youth Risk Behavior Surveillance System—United States, 2021," MMWR Supplements, vol. 72, Suppl. 1, pp. 1–12, 2023, doi: 10.15585/mmwr.su7201a1.

[14] N. H. Somé, P. Noormohammadpour, and S. Lange, "The use of machine learning on administrative and survey data to predict suicidal thoughts and behaviors: A systematic review," Frontiers in Psychiatry, vol. 15, p. 1291362, 2024, doi: 10.3389/fpsyt.2024.1291362.

[15] M. R. Gohari, A. Doggett, K. A. Patte, M. A. Ferro, J. A. Dubin, C. Hilario, et al., "Using random forest to identify correlates of depression symptoms among adolescents," Social Psychiatry and Psychiatric Epidemiology, vol. 59, no. 11, pp. 2063–2071, 2024, doi: 10.1007/s00127-024-02695-1.

[16] W. A. Rothenberg, A. Bizzego, G. Esposito, J. E. Lansford, S. M. Al-Hassan, D. Bacchini, et al., "Predicting adolescent mental health outcomes across cultures: A machine learning approach," Journal of Youth and Adolescence, vol. 52, no. 8, pp. 1595–1619, 2023, doi: 10.1007/s10964-023-01767-w.

[17] S. Kim, H. Kim, S. Kim, H. Lee, A. Hammoodi, Y. Choi, et al., "Machine learning-based prediction of substance use in adolescents in three independent worldwide cohorts: Algorithm development and validation study," Journal of Medical Internet Research, vol. 27, p. e62805, 2025, doi: 10.2196/62805.

[18] E. D. Hill, P. Kashyap, E. Raffanello, Y. Wang, T. E. Moffitt, A. Caspi, et al., "Prediction of mental health risk in adolescents," Nature Medicine, vol. 31, no. 6, pp. 1840–1846, 2025, doi: 10.1038/s41591-025-03560-7.

[19] Centers for Disease Control and Prevention, "2021 National Youth Risk Behavior Survey Questionnaire," 2021. Available: https://www.cdc.gov/yrbs/files/2021/pdf/2021-YRBS-National-HS-Questionnaire.pdf (Accessed: Sep. 17, 2026).

[20] Centers for Disease Control and Prevention, "2023 National Youth Risk Behavior Survey Questionnaire," 2023. Available: https://www.cdc.gov/yrbs/media/pdf/2023/2023_YRBS_National_HS_Questionnaire.pdf (Accessed: Sep. 17, 2026).

[21] G. Long, "Incremental Predictive Value of Bullying Exposure Indicators for Predicting Persistent Sadness or Hopelessness in the 2023 National Youth Risk Behavior Survey," undergraduate senior thesis, Indiana University Indianapolis, Indianapolis, IN, USA, 2026. Available: https://github.com/gblongiu/yrbs-thesis.

© 2026 The Author(s) Licensed under CC BY 4.0 | Published by Engineering, Science and Environmental Research Publications | e-ISSN 2839-5360 | ijese.in