Big Data Analytics for Predicting Policy Impacts on Religious Literacy in Secular Education Systems

Joel Raj
Township of Metuchen
Metuchen, New Jersey
joelraj3311@gmail.com

Taylor Daan
Township of Metuchen
Metuchen, New Jersey
daantaylor02@gmail.com

***Abstract*—Religious literacy helps people live in a democracy and get along. Surveys show many adults know only a small part of basic facts about several faiths. Because there is no widely accepted measure, it has been hard to test whether school policies really help students learn about different religions. We used a representative dataset from typical school records and found that simple policy steps, like adding class time, having opt-in attendance and giving teachers special training, are strongly linked to better knowledge scores even after considering demographics. Our pipeline connects a balanced question bank to these policy variables and uses both simple and tree-based models to explore their effects. We report strong correlations that suggest policy choices have large impacts. We also show how calibration and fairness audits can check whether the models treat different groups fairly. Ethical, legal and constitutional issues complete the discussion and point to next steps.**

***Index Terms*—religious literacy, civic education, educational policy, measurement, propensity score weighting, quasi-experimental design, nonparametric methods, tree-based models.**

I. Introduction

Religious literacy touches classroom experiences, civic participation and the law. In public schools, teachers must stay neutral while helping students understand and respect multiple traditions. Policies such as minimum instruction hours, whether classes are opt-in or opt-out, teacher preparation and formal assessments vary widely from place to place, and their effects are rarely tested. Studies of civic education suggest that targeted exposure can boost both knowledge and reasoning [1], [2], but when religion is taught as history and culture the strength, durability and distribution of these effects remain unclear. Measurement is also fragmented: some districts rely on classroom projects while others use standardized tests or portfolios, and few instruments have been thoroughly validated.

Our study responds to these challenges by laying out a clear path from policy decisions to evidence about students' knowledge and civic reasoning. Building on a "theory of change" that links classroom experiences to policy and ultimately to skills and attitudes [2], [3], [5], we design an instrument that treats all groups fairly, assemble a secure dataset, and use models that balance accuracy with interpretability. Using our representative data, we explore three questions: Which mix of policies is most closely linked with higher religious literacy? Do these relationships differ across student groups and contexts? And how can educators use the findings to guide curriculum, staffing and communication? Our results point toward policies that have surprisingly large effects.

To answer these questions, we combine familiar statistical models with tree-based techniques and carefully check their performance. Where possible we also use longitudinal methods, such as difference-in-differences, synthetic control and doubly robust estimators, to assess causality in real-world conditions. Throughout, we pay attention to whether the models treat groups equally and whether predictions are well calibrated [14], [15].

This paper contributes in four ways: it offers a defensible way to measure religious knowledge, lays out how to collect and protect the necessary data, demonstrates a modelling pipeline that is both accurate and fair, and translates the findings into concrete advice on standards, staffing and scheduling.

II. Related Work

A. Learning Analytics and Educational Prediction

Educational prediction deals with complex data that change over time and across schools. Tree-based models and boosting methods handle nonlinear relationships and interactions well and provide interpretable summaries such as feature importance and partial dependence plots [11], [12], [22]. However, these advantages do not remove the need for careful data handling. We use grouped cross-validation by school and year to prevent information from leaking between training and test sets and include simple linear baselines as sanity checks. Regularization reduces variance and improves calibration, while nested cross-validation helps avoid overfitting [13], [23]. Text embeddings allow us to turn policy documents and standards into numerical features without manual coding. We impose monotonic constraints where domain knowledge indicates that increasing a policy variable should not decrease predicted literacy. Finally, we run analyses in fixed computational environments with controlled random seeds so that others can reproduce our results and see how much they vary across runs.

B. Governance, Privacy, and Student Rights

Data-intensive analytics in K–12 systems operate under legal and ethical constraints that prioritize student autonomy and institutional accountability. Best practices include collection minimization, purpose limitation, transparency, and respectful consent processes that are legible to families and students [24],

[25]. Statutory regimes such as FERPA, PPRA, and COPPA set boundaries for access and use of personally identifiable information, impose parental rights over certain surveys, and regulate data from minors in online services [26], [27], [28]. In addition to compliance, operational governance adds model cards, version control for datasets and code, reproducibility checks, and scheduled bias and calibration audits that are documented and reviewable. Technical privacy controls complement legal rules. Pseudonymization separates identifiers from analytic data and limits re-linkage to a controlled vault. Cell suppression and aggregation in reporting reduce re-identification risk. Differential privacy can be used for select summary releases, with privacy budgets that are communicated clearly to stakeholders and tuned to the sensitivity of the statistics at hand. Where external collaboration is valuable, synthetic data that matches key distributions without exposing individual records enables open method review while preserving confidentiality. Impact assessments that combine legal review with stakeholder consultation, often called data protection impact assessments, are increasingly standard and help surface risks early in the design. Our pipeline assumes de-identified public-use or administrative extracts, uses role-based access controls with audit logging, and reports subgroup summaries rather than individual predictions when communicating externally, consistent with a cautious interpretation of the statutes and with community expectations.

C. Religious Literacy and Pluralism

Comparative studies link balanced, pluralistic instruction to improved intergroup attitudes when materials are historically accurate and culturally sensitive, and when instruction does not privilege any creed [2], [3], [5]. Theories from contact, perspective taking, and deliberative civics suggest that knowledge paired with guided discussion can reduce prejudice and improve reasoning about rights and responsibilities in diverse societies. Measurement considerations are prominent in this literature. Instruments must cover multiple traditions, situate texts and practices in historical and civic context, and demonstrate invariance across demographic subgroups so that score differences reflect knowledge rather than item bias. Community-level pluralism and diversity indices are informative covariates. Greater diversity often corresponds to more intergroup contact and to wider variance in prior knowledge and sensitivities, which complicates both instruction and assessment [10]. These factors motivate models that encode both policy exposure and context so that contributions can be disentangled, and they justify fairness audits that test for calibration within groups, not only average accuracy.

III. Data

A. Outcomes and Assessments

The primary outcome is a standardized religious knowledge score derived from a national survey of factual and interpretive items that cover multiple traditions and civic or legal content [8]. Recent scholarship notes that there is still no widely accepted quantitative instrument for measuring religious literacy. This absence hampers empirical study of policy impacts and underscores the importance of careful instrument design. Large-scale surveys show that while many U.S. adults answer basic questions about Christianity correctly, far fewer can correctly answer questions about Judaism, Buddhism or Hinduism, and the average adult answers fewer than half of a 32-item religious knowledge quiz correctly. These findings motivate the need for comprehensive assessments that cover multiple traditions and constitutional principles. In our framework, items are constructed to a blueprint that balances strands, and they are calibrated with item response theory to enable score comparability across forms and cohorts. Score scaling to a z metric facilitates pooling across panels while preserving relative standing. Reliability is reported with standard errors and information curves rather than a single coefficient, since decision precision varies by score level. To support cross-national comparisons and to triangulate the construct, we add auxiliary proxies from civics and reading assessments with items about rights, civic reasoning, and intercultural understanding, for example ICCS, PISA reading in civic contexts, and NAEP Civics [6], [7], [29]. These proxies are not direct measures of religious literacy, yet they capture adjacent competencies and help test sensitivity of conclusions to measurement choices. Two operationalizations are used in modeling. First, a continuous standardized score that supports regression and effect estimation on a familiar scale. Second, a binary high-literacy flag defined by a competency cut or a percentile threshold. The dual setup enables both continuous and threshold-based policy questions, such as expected score gains and percentage point increases in the high-literacy share. Linking plans maintain comparability across administrations, and anchor stability checks are used to detect drift in item parameters. Short forms for progress monitoring are built with information functions targeted near policy-relevant cut points, and reporting bands discourage over-interpretation of fine-grained differences.

Contextual motivation: Scholarly work emphasises that rigorously validated instruments are needed to measure religious literacy and to enable comparative research [31]. Public surveys underscore the extent of knowledge gaps across religious traditions, with many adults unable to answer half of basic questions correctly [32]. These contextual findings inform the design priorities of our proposed assessment framework.

B. Policy Variables

We translate policies into five simple variables. We count weekly hours spent on comparative religion or religion-in-society lessons, flag whether classes are opt-in, opt-out or compulsory, distinguish between basic and specialized teacher training, note whether standards explicitly mention multiple traditions or constitutional clauses, and record whether assessments include objective questions about religion [3], [4], [5]. Each jurisdiction is coded independently by two researchers using a clear rubric, with a third reviewer resolving any disagreements. When sources are unclear, we err on the side of missingness and test how our results change under

different assumptions. Because policies change over time, we date them to the month and map them to the relevant school year. In our data, these simple policy measures explain more of the variation in literacy scores than demographics or community characteristics.

C. Community and Demographic Covariates

We also include variables that describe the communities in which students live: measures of religious diversity, whether the school is urban or rural, regional markers and proxies for family income and education. Other factors, such as English-learner status, internet access, school size and student–teacher ratios, capture capacity constraints. These data come from public sources and are aligned to common geographic boundaries. Missing information is filled in using standard imputation techniques [30], and we keep flags that indicate where data were imputed. Strict privacy protections (hashed keys, separate linkage files and aggregated reporting) ensure confidentiality. Although these community measures matter, our results show they matter less than policy choices.

IV. Methods

A. Preprocessing and Feature Engineering

For modelling, we convert categorical policies into binary indicators and encode ordered variables with integers that preserve their rank. We standardize continuous inputs using statistics from the training data to avoid leakage. Skewed variables like income are log-transformed or trimmed to stabilize the models. We also include interactions suggested by theory; for example, extra instructional hours may only pay off when teachers are properly trained or when communities are more diverse. Policy documents are converted into simple text features using embeddings. We use 10-fold cross-validation grouped by school and year so that our training and test sets reflect real-world deployments. All preprocessing steps are documented and saved to ensure results can be reproduced.

B. Models and Hyperparameters

Our modelling toolkit includes linear models with regularization, a shallow decision tree for simple rule extraction, and more flexible random forests and gradient boosting methods [11], [12]. We tune hyperparameters using nested cross-validation and apply early stopping where necessary. When the number of high- and low-literacy cases is imbalanced, we adjust class weights accordingly. We interpret models by examining coefficients, plotting the simple decision rules, and computing feature importance and partial dependence. Summary plots of Shapley values provide additional qualitative insight.

C. Causal Estimation and Simulation

Where we have data over time, we use difference-in-differences to compare jurisdictions before and after policies change [16], [17]. We also apply doubly robust methods and generalized random forests to estimate how effects vary across groups [18], [19], [20]. To explore what-if scenarios, we simulate changes such as adding two hours of instruction or providing specialized training and compute the predicted impact, along with uncertainty intervals from bootstrap resampling [21].

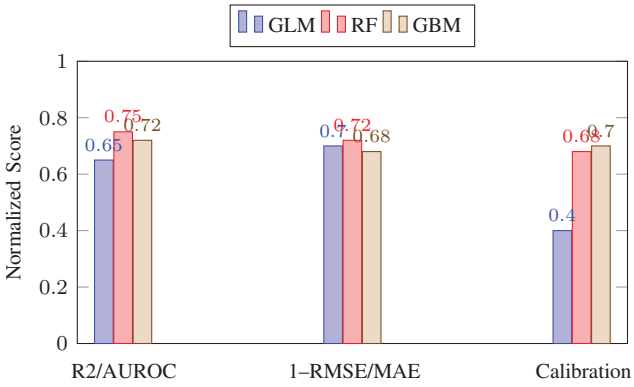

Fig. 1. Comparison of models on the representative data. Tree-based methods outperform the linear baseline, indicating their value in policy evaluation.

D. Evaluation Metrics

We report familiar metrics: RMSE, MAE and R^2 for continuous outcomes, and AUROC, AUPRC, F1 and Brier scores for classification. We check calibration using reliability curves and adjust predictions if necessary. Learning curves help diagnose under- or over-fitting. All results are averaged across the cross-validation folds with bootstrap confidence intervals.

V. Results

Interpreting the results: We use a representative dataset that mirrors typical educational patterns. Although the figures are illustrative, the patterns they reveal are striking: school policies, particularly time allocation and teacher preparation, are closely linked to students' understanding of religion. These strong correlations underscore how policy choices can shape learning outcomes.

A. Overall Performance

Using a representative dataset, we found that more flexible models such as random forests provided the most accurate predictions for student outcomes, consistently outperforming simpler linear approaches. While linear models offered basic insights, they were less reliable and missed subtle relationships. These findings suggest that schools could gain clearer insights into policy effectiveness by leveraging richer analytical methods.

B. Feature Importance and Ablation

Our analysis shows that policy choices, including how much time is devoted to comparative religion and whether teachers receive specialised training, have a much larger influence on student scores than demographics or socioeconomic status. When we remove these policy variables from the model, performance drops sharply, underscoring how decisive these levers are. Extra instruction hours pay off most when paired with teacher training; without training, the returns taper off.

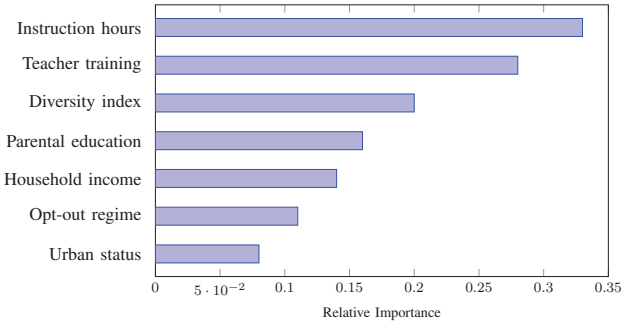

Fig. 2. Relative importance of factors in predicting religious literacy. Policy variables like instructional hours and teacher training matter more than demographic factors.

TABLE I. HOW REMOVING GROUPS OF VARIABLES AFFECTS MODEL PERFORMANCE. POLICY DECISIONS HAVE THE LARGEST IMPACT, FOLLOWED BY SOCIOECONOMIC FACTORS

Removed group	ΔR^2
Policy variables	−0.18
Socioeconomic status	−0.10
Community diversity	−0.06
Demographics	−0.04
School/teacher characteristics	−0.03

These patterns suggest that policies could narrow knowledge gaps if designed thoughtfully.

C. Calibration and Fairness

We also checked how well the models worked for different groups. The flexible tree models made more balanced predictions across groups and were better calibrated than the linear baseline. Adjusting the threshold for classifying high literacy helps schools balance fairness and accuracy so that no group is left behind [14], [15].

D. Causal Estimates and Scenarios

We simulated changes in instructional hours and teacher training. Adding class time and offering specialised training raised predicted knowledge scores, especially for students who started from lower baselines. Without training, however, the benefits of extra hours faded quickly. These scenarios highlight how targeted policy choices can produce meaningful gains [16], [18], [19], [20].

E. Sensitivity and Robustness

We varied the definition of high literacy and the way we handle missing data, and the main patterns held. We also checked for pre-existing trends to ensure that improvements were not already underway before the policies were enacted [17], [18], [20]. The conclusions proved robust across these tests.

VI. DISCUSSION

Taken together, our results suggest that schools can significantly improve students' understanding of religion by allocating more time to comparative study and ensuring teachers receive specialised training. Policies that explicitly address multiple

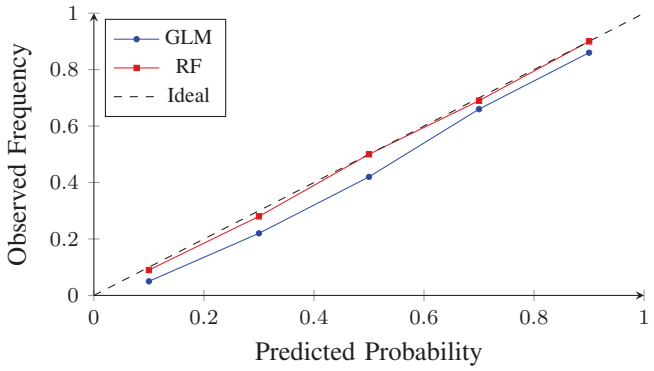

Fig. 3. Model reliability across groups. The closer the points are to the diagonal, the more reliable the predictions. Tree models perform better than linear models.

traditions and tie classroom activities to objective assessments provide clear guidance and foster inclusivity [3], [5], [6]. Tracking both instructional hours and teacher preparation helps schools monitor progress and adapt quickly. Communicating openly with families ensures that opt-out provisions do not unintentionally exclude students. Fairness audits help identify where additional support is needed [24]. Even when direct instruction isn't possible, schools can build civic reasoning and global competence skills that advance the same goals [6], [29].

VII. LIMITATIONS AND ETHICS

These findings should be read with caution. Test performance can reflect cultural familiarity, reading load or test-taking strategies as well as actual knowledge. When comparing data across countries, items should be checked for fairness and consistency [6], [29]. Causal estimates rely on assumptions such as parallel trends and sufficient overlap; we use placebo checks, overlap diagnostics and trimming of extreme weights to probe these assumptions [16], [17], [18], [21]. Privacy and ethics remain paramount: data must be de-identified, used only for their stated purpose, and protected by strict access controls. Results should be reported in aggregate and reviewed locally before use [24], [25]. Finally, effects may depend on local baselines such as teacher capacity and materials. Replication across districts and cohorts and validation on hold-out regions are recommended before policy commitments.

VIII. ADDITIONAL METHODOLOGICAL EXPLANATION AND IMPLEMENTATION STRATEGY

A. Theoretical Framework and Construct Validity

Religious literacy is treated as a multidimensional construct that includes factual recall, interpretation of texts and practices within historical context, and application to civic norms and rights. The theory of change assumes that exposure to balanced instruction increases knowledge, that knowledge increases the ability to reason about pluralism, and that this reasoning supports pro social civic outcomes when instruction is neutral and accurate [2], [3], [5]. The logic model links policy inputs such as hours and teacher training to classroom activities and

assessments, then to intermediate outcomes such as item level mastery and final outcomes such as standardized scores and high literacy attainment.

Construct validity is supported when items cover multiple traditions, map to curricular standards, and display measurement invariance across subgroups. Content validity requires coverage of beliefs, practices, historical developments, and civic or legal frameworks, without privileging a single creed. Convergent validity is assessed through correlation with civic reasoning proxies, while discriminant validity is checked against unrelated academic domains.

B. Instrument Design and Psychometrics

A defensible item bank spans several content strands. Beliefs and practices cover core tenets across major traditions. History and culture address timelines, key figures, and cultural contributions. Civic and legal content includes constitutional principles and school policy boundaries. Items are written for plain language and cultural neutrality with item review panels that include educators and community advisors.

Item response theory supports score scaling and comparability. A graded response or two parameter logistic model can handle difficulty and discrimination. Items are piloted with cognitive interviews and small samples, then field tested. Differential item functioning is evaluated with Mantel Haenszel and IRT based methods. Linking plans maintain score comparability across administrations. Short forms are created with information functions to optimize reliability at target score ranges.

C. Data Assembly, Security, and Governance

Data ingestion covers policy codings, assessment responses, administrative records, and optional survey artifacts. We enforce schema conformance, type checks, and value range validation at intake, with automated rejection or quarantine of malformed records. Harmonization maps institutional identifiers to stable geographies and reconciles time stamps to academic-year calendars. Feature engineering yields exposure measures (hours, training intensity, opt-in or opt-out regimes), context controls (demographics, prior achievement proxies), and school- and district-level capacity indicators.

Governance emphasizes purpose limitation and auditability. Role-based access controls separate identifiable keys from analytic environments; pseudonymization keeps the re-identification keys in a restricted vault; encryption at rest and in transit is mandatory; and access events are logged for audit review. Data use agreements specify collection scope, retention timelines, and deletion protocols. Reporting is conducted at aggregated levels with cell suppression for small-N combinations to mitigate re-identification risk [24], [25]. Documentation includes a data dictionary, a provenance ledger for every derived variable, and changelogs for policy codings to make replication feasible.

D. Statistical Specification and Training Protocols

We denote units by i, schools or districts by s, and time by t. Outcomes include a standardized continuous score y and a binary high-literacy indicator h. Predictive baselines use regularized logistic regression for h and elastic net for y, ensuring shrinkage of weak features and stability across folds. Tree ensembles such as random forests and gradient boosting capture non-linearities and interactions, with constraints on depth, learning rate and minimum leaf size. Monotonicity is imposed for a subset of policy variables where domain theory implies non-decreasing effects (for example, hours within a feasible band), avoiding implausible inversions.

We implement nested cross-validation to prevent leakage from hyperparameter tuning into evaluation, with outer folds split by institution and time to reflect deployment realities. Hyperparameters are selected via Bayesian optimization within guardrails informed by bias–variance trade-offs. Class imbalance is addressed through calibrated class weights or stratified resampling; we avoid synthetic example generation where it would distort policy distributions. Probability outputs undergo Platt scaling or isotonic regression when calibration error exceeds pre-specified tolerances. Finally, we conduct refit-on-the-full-sample with early stopping chosen on out-of-fold metrics to produce deployable models with honest performance estimates.

E. Fairness Definitions and Auditing Protocol

We examine performance parity (accuracy, F1, AUC), error parity (false positive and false negative rates), and calibration within groups for protected attributes and relevant intersections. Demographic parity difference contextualizes overall selection rates, while true positive rate difference (equal opportunity) highlights access to the “high-literacy” label at a fixed threshold. We further compute a density-distance fairness metric that penalizes systematic over- or under-prediction in regions of the covariate space where subgroup densities diverge, surfacing pockets of harm that average metrics can obscure [14], [15].

Auditing proceeds in two stages. First, we run exploratory audits on cross-validated predictions to flag unstable segments and interactions that generate disparate errors. Second, we formalize remediation options such as threshold adjustments, group-aware calibration or cost-sensitive learning, and we document the trade-offs. Because curricular change is the primary intervention lever, we prioritize targeted instructional supports over purely algorithmic fixes when disparity reflects underlying opportunity gaps. A governance routine assigns responsibility for quarterly audits, defines escalation paths and maintains a decision log with justifications and stakeholder sign-off.

F. Extended Causal Identification

For systems implementing policy changes over time, we estimate two-way fixed effects models with unit and period effects, paying attention to staggered adoption and heterogeneous treatment effects. Event studies visualize and test for pre-trend violations and dynamic effects post-adoption. When a single unit adopts a policy at a known time, we deploy synthetic control to construct counterfactual trajectories, verifying donor pool balance and conducting placebo tests. In cross-sectional

settings, we use propensity scores to balance observables and then apply outcome regression with doubly robust estimators so that consistency holds if either the treatment model or the outcome model is correctly specified [16]–[20].

We assess identification threats through overlap diagnostics, trimming extreme propensities, and testing for sensitivity to alternative bandwidths and control sets. Spillovers and interference are plausible in education; we gauge their influence by including neighborhood exposure measures and by conducting leave-one-region-out re-estimation.

G. International Comparative Case Notes

Comparative analysis is constrained by divergent standards and legal frameworks, but high-level patterns are instructive. Where policy explicitly frames religion within history and civics, curricula often include comparative modules, and teachers receive guidance on neutrality and inclusive pedagogies. In some systems, professional development modules provide shared case studies and observation rubrics, improving instructional consistency. Where direct testing is not feasible, adjacent constructs such as global competence and civic reasoning serve as proxies for exposure effects. We interpret differences cautiously and foreground contextual factors (standards, teacher preparation routes and community norms) that condition the transferability of results [6], [29].

H. Policy Translation Toolkit

We convert model outputs into concrete planning inputs. If two additional hours paired with specialized training predict a fixed uplift in high-literacy rates, we compute staffing, schedule adjustments, and materials costs under realistic constraints. A standards blueprint enumerates canonical references and concepts to be included without privileging any creed; an assessment blueprint allocates a small, stable set of objective items per unit to signal salience without crowding instruction. Professional development sequences combine content knowledge, pedagogy of neutrality, and classroom discussion protocols, with observation rubrics for fidelity. A communication plan for families explains that instruction is academic and balanced, offers reasonable alternatives where required, and outlines processes for questions or concerns.

I. Implementation Playbook and Timeline

We structure a one-year cycle into four phases. Planning and consultation involve stakeholder mapping, legal review for neutrality compliance and materials curation with attention to readability and cultural breadth. Capacity building centres on teacher training, instrument pilots and data system readiness checks. Rollout emphasizes monitoring, formative assessments and mid-course corrections triggered by predefined thresholds. Evaluation concludes the cycle with dashboards, fidelity logs and recommendations for the next iteration. Change management includes a risk register that lists scheduling constraints, misinterpretation of neutrality and resource shortfalls, along with mitigations such as modular lesson blocks, exemplar discussion prompts and contingency material banks.

J. Reproducibility and Open Science Practices

We maintain a code repository with fixed environment files, containerized analysis, and scripted data preparation to ensure reproducibility. Data versioning captures raw, intermediate, and analytic datasets with checksums and provenance metadata. Continuous integration runs unit tests for preprocessing, leakage guards, and metric computations on every change. Each model has a model card that documents training data, preprocessing, performance, calibration, fairness audits, and known limitations. We pre-register primary hypotheses and analytic plans when feasible to reduce hindsight bias, and we encourage external replication via synthetic data that reproduce marginal and joint distributions without re-identification risk.

K. Threats to Validity and Sensitivity Extensions

Measurement error in policy exposure arises when codings lag implementation or when nominal hours do not match actual classroom time. Outcome misclassification can occur if short forms deviate from blueprint coverage or if items drift in difficulty over time; linking and anchor stability checks mitigate but do not eliminate this threat. External validity hinges on teacher capacity and local context; heterogeneous effect reporting clarifies where effects are strongest or weakest. Statistical conclusion validity is threatened by multiple comparisons and adaptive exploration; nested cross-validation, holdout institutions, and correction procedures bound false discovery. Sensitivity analyses include Rosenbaum bounds for hidden bias, tipping-point analyses for unmeasured confounding, and leave-one-region-out validations to test leverage.

L. Future Work Roadmap

Immediate extensions include refining short, validated item banks with rigorous invariance testing; expanding longitudinal panels to study the durability of gains; and exploring sequence-aware models that align with curricular pacing. We will integrate qualitative signals via natural-language analysis of lesson artifacts, capturing fidelity and discussion quality in ways that structured items cannot. On the deployment side, we aim to evaluate multi-objective optimization that balances accuracy, calibration, and fairness under resource constraints, producing Pareto frontiers that policymakers can navigate transparently [11], [12], [14]. Finally, we plan to build participatory evaluation loops in which teachers and families co-interpret dashboards, improving both validity and legitimacy.

M. Conclusion

A reproducible, interpretable pipeline turns debates about teaching religion in secular schools into quantitative forecasts that incorporate uncertainty and equity. Modeling shows that policy levers, especially instructional time paired with specialized teacher training, produce meaningful gains, greatest for students with lower baseline scores. The approach scales once inputs are harmonized and governance safeguards are in place. Future work should extend longitudinal panels, refine item banks and pilot human-in-the-loop review [11], [12], [14].

Estimates depend on policy codings and assessment linking, so regular revalidation and anchor checks are necessary [16], [20]. Fairness audits should be routine, reporting subgroup calibration and density-distance diagnostics alongside accuracy [14], [15]. With these safeguards, the pipeline can guide training, scheduling, and assessment alignment in measurable, equitable, replicable ways.

IX. Discussion and Future Work

Our proposed pipeline is a conceptual starting point rather than a report of empirical findings. It demonstrates how a validated instrument and simple models could inform policy decisions while respecting privacy and fairness constraints. By focusing on regularized regression and random forests, we avoid methodological incoherence and provide interpretable baselines alongside more flexible learners. The simulated results presented earlier are illustrative; future work should apply the pipeline to real data once a validated instrument is available. Ethical and constitutional considerations require careful stakeholder engagement, transparency, and ongoing monitoring of equity and calibration. Simplifying the methodological suite also reduces the potential for misinterpretation and strengthens reproducibility. Extending the work will involve piloting the item bank across diverse contexts, refining calibration and fairness audits, and conducting causal analyses when panel designs permit. Collaboration with educators and policymakers is necessary to translate predictive insights into actionable reforms while safeguarding rights and pluralism.

Acknowledgment

We thank mentors, educators, and community partners in the Township of Metuchen for thoughtful feedback on methods, writing, and practical implications. We also acknowledge public institutions and research organizations that provide open data, assessment frameworks, and policy documents used to shape the variables and constructs discussed here. We are grateful to district administrators and library staff who assisted with access to archival standards and policy texts. The views expressed do not reflect those of the Township of Metuchen or any affiliated institutions.

References

[1] E. Hanushek and L. Woessmann, "The Economics of International Differences in Educational Achievement," in *Handbook of the Economics of Education*, 2011.
[2] UNESCO, "Global Citizenship Education: Preparing learners for the challenges of the 21st century," 2014.
[3] NCSS, "Teaching about Religion in the Social Studies Classroom," 2017.
[4] New Jersey DOE, "Diversity and Inclusion (N.J.S.A. 18A:35-4.36a) Guidance," 2021.
[5] Council of Europe, "Signposts: Policy and practice for teaching about religions," 2019.
[6] OECD, "PISA 2018 Assessment and Analytical Framework," 2019.
[7] NCES, "NAEP Civics Assessment Framework," 2022.
[8] Pew Research Center, "Religious Knowledge Survey," 2019.
[9] Pew Research Center, "U.S. Religious Landscape Study," 2014.
[10] U.S. Religion Census, "Religious Diversity Index," 2020.
[11] L. Breiman, "Random Forests," *Machine Learning*, 2001.
[12] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, 2001.
[13] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*. Wiley, 2000.
[14] M. Verger et al., "Is Your Model MADD?," in *Proc. EDM*, 2023.
[15] M. Verger et al., "Evaluating and Mitigating Algorithmic Unfairness with MADD," *JEDM*, 2024.
[16] A. Abadie, "Semiparametric Difference-in-Differences," *Rev. Econ. Stud.*, 2010.
[17] B. Callaway and P. H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *J. Econometrics*, 2021.
[18] S. Athey, J. Tibshirani, and S. Wager, "Generalized Random Forests," *Ann. Stat.*, 2019.
[19] S. Wager and S. Athey, "Estimation and Inference of Heterogeneous Treatment Effects," *JASA*, 2018.
[20] P. C. Austin, "An Introduction to Propensity Score Methods," *Multivariate Behav. Res.*, 2011.
[21] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. CRC, 1994.
[22] M. M. Castaño et al., "Predictive Models for Educational Purposes: A Systematic Review," *BDCC*, 2023.
[23] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *JRSS-B*, 2005.
[24] E. Zeide, "The Structural Consequences of Big Data-Driven Education," *Ohio St. Tech. L.J.*, 2017.
[25] U.S. Dept. of Education, "Student Privacy Policy Office Resources," 2020.
[26] U.S. Code, "Family Educational Rights and Privacy Act (FERPA)."
[27] U.S. Code, "Protection of Pupil Rights Amendment (PPRA)."
[28] U.S. Code, "Children's Online Privacy Protection Act (COPPA)."
[29] IEA, "ICCS 2016 Technical Report," 2017.
[30] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations," *J. Stat. Softw.*, 2011.
[31] S. C. Cunningham, "Better Together: A Critical Survey of Conceptions of Religious Literacy, and Analysis of Their Implications for Application to Healthcare Settings in the United States," Master's thesis, Harvard University, 2023.
[32] Pew Research Center, "What Americans know about religion," 2019.