Reducing Student-Instructor Mismatches in Driving Schools Through Compatibility Analysis

Aki Takahashi[1]
Takahiro Yonekawa[2]
[1]Musashi-sakai Driving School Co., Ltd., 2-6-43 Sakai, Musashino-shi, Tokyo 180-0022, Japan
[2]Brain Signal, Inc., 27F, Shiroyama Trust Tower, 4-3-1 Toranomon, Minato-ku, Tokyo 105-6027, Japan
aki@musasisakai-ds.co.jp, yonekawa@bsgnl.com

Abstract

Matching students with suitable instructors is critical for learning effectiveness in driving schools, yet most schools rely on manual, experience-based allocation. This paper presents a real-world case study at a Japanese driving school analyzing student-instructor compatibility using operational data. We integrate 5-point and 2-point evaluation data, perform exploratory analysis connecting aptitude-test personality patterns to instructor ratings, and build a LightGBM classification model to predict mismatches. Rather than recommending optimal pairs, we focus on reducing mismatches by avoiding low-compatibility pairings while respecting equal distribution, enabling instructors to maximize their strengths and improving student outcomes.

Introduction

Driving schools play a central role in road-safety education. In Japan, nearly all citizens obtain their driver's license through driving schools. Instructors are typically assigned to students using a combination of manual reservation rules, instructor availability, and local heuristics based on experience. The fundamental allocation rule in most driving schools is to distribute students evenly among instructors to ensure fairness. However, mismatches in *compatibility*—for example, differences in communication style, pace, or attitude—can lower satisfaction, slow learning progress, and increase rescheduling costs. Beyond the immediate learning process, improving compatibility may also contribute to long-term road safety outcomes, such as reducing traffic violations and accidents after license acquisition.

Driving schools commonly use standardized aptitude tests such as the OD-style safety aptitude test, developed by Denno Co., to assess learners across multiple domains—driving function, health and maturity, personality traits, and driving manners—before instruction begins; administration of such aptitude tests is mandated by the National Police Agency of Japan (Denno Co., Ltd. n.d.; Fujihira 2002).

Rather than replacing human judgment with AI-driven optimal matching, we propose using AI to *reduce mismatches* while preserving the human-centered allocation practice. This approach allows instructors to maximize their strengths and enables students to learn more effectively, while balancing fairness (equal distribution) with mismatch avoidance—a constraint that distinguishes our problem from conventional recommendation tasks.

Our contribution is to formalize the problem of identifying student-instructor mismatches from a practical scheduling perspective.

Problem Setting and Dataset

Operational Setting

We collaborate with Musashi-sakai Driving School, Japan's largest driving school by enrollment, which operates multiple cars and instructors in parallel. Each student books lessons over several weeks and completes an evaluation questionnaire about the instructor after each lesson. We treat these post-lesson ratings as an operational proxy for student-instructor compatibility. The school maintains extensive records of these satisfaction surveys, providing a rich dataset for analysis. Each student also takes an aptitude test before starting, which yields personality pattern codes and various aptitude scores. These two data sources—post-lesson evaluations and aptitude-test profiles—form the foundation of our compatibility analysis.

Our goal is not to replace the school's equal-distribution rule but to *reduce mismatches* by filtering out student-instructor pairs that are likely to result in low compatibility ratings. This approach respects human judgment while using AI to reduce risks.

Data Representation and Integration

The analysis consolidates two data sources: (1) aptitude-test data for students, and (2) post-lesson evaluation data (student ratings of instructors). All personal identifiers are removed and replaced by internal IDs. The aptitude-test data contains 46,221 records (after deduplication) with personality pattern codes and various aptitude scores. The post-lesson evaluation data consists of two formats: 5-point star ratings (224,759 records) and 2-point evaluations (194,939 records). Table 1 provides an overview of the dataset, showing the number of records and the time period for each data type.

Data Integration Method. We map 5-point ratings to binary: stars 4–5 become "good" (1) and stars 1–3 become "mismatch" (0). This mapping reflects the heavy skew in

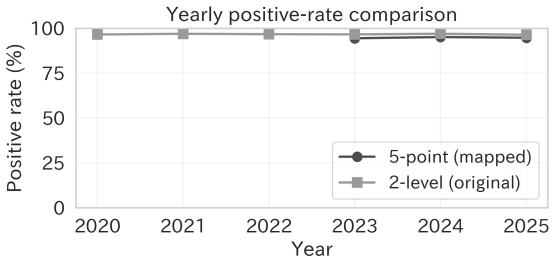

Figure 1: Year-by-year positive-rate comparison between the two data sources (5-point mapped and original 2-point). The consistent trends support the integration method.

Data Type	Records	Period
Aptitude test data	46,221	2020–2025
5-point evaluations	224,759	2020–2025
2-point evaluations	194,939	2020–2025
Integrated dataset	345,908	2020–2025

Table 1: Dataset overview.

5-point data (stars 4–5 account for 94.85% (213,176 of 224,759 records), making the distribution effectively binary). We integrate the mapped 5-point data with the original 2-point data (converting its 1=good, 2=mismatch format to 1/0), yielding a unified binary dataset.

Validation. We validate the integration by comparing the year-by-year positive rate (share of "good" evaluations) for both sources. Figure 1 shows that the trends are consistent across years after mapping. The final integrated dataset contains 345,908 evaluation records spanning 2020–2025.

Aptitude Test Details

The aptitude test used in this study corresponds to an OD-style safety aptitude test administered in Japanese driving schools (Denno Co., Ltd. n.d.). The test provides multiple sub-scores (e.g., attention, judgment, flexibility, decision-making, precision, health/maturity indicators, and driving manners) and a categorical personality-pattern item recorded as a pattern number together with student gender. The OD test has also been examined in prior traffic-psychology research in relation to the propensity for traffic accidents and violations (Fujihira 2002).

Gender-specific personality patterns and BasicType mapping. In the raw data, the personality-pattern IDs are gender-specific and non-overlapping (88 distinct IDs in total across male and female examinees). To avoid gender-dependent coding and reduce sparsity, we map each pair (gender, pattern ID) to a gender-agnostic *BasicType* with 44 categories using a conversion table provided with the test materials. All analyses in this paper use BasicType rather than the original gender-specific pattern numbers.

Data Preprocessing. Before integration, we impute missing values, handle outliers via the IQR method, and encode categorical variables, yielding 23 features.

Compatibility Analysis

Exploratory Data Analysis

We first visualize the distribution of evaluations to understand patterns in student-instructor compatibility. Cross-tabulation between student BasicTypes and instructors reveals combinations that frequently yield low compatibility ratings. Figure 2 shows a heatmap of positive rates for BasicType-instructor pairs. Notably, some instructors show consistently high compatibility across BasicTypes, while others exhibit more variable patterns, suggesting that instructor-specific factors play a significant role in compatibility.

Personality pattern analysis. Using gender-agnostic *BasicType* categories, we observe systematic differences in mismatch rates across BasicTypes. Figure 3 shows the positive rate by BasicType, revealing that some personality patterns are more prone to low-compatibility outcomes and thus require careful instructor assignment.

Identifying Mismatches

We identify mismatches as student-instructor pairs where the evaluation indicates low compatibility. After joining evaluation records with aptitude-test features and removing incomplete rows, the modeling dataset contains 328,456 records and is highly imbalanced: mismatches (0) account for 4.50% (14,768 records) while good matches (1) account for 95.50% (313,688 records). This imbalance reflects the operational reality that most lessons are satisfactory, making detection of the minority of problematic pairs the central modeling challenge.

Prediction Model

To predict mismatches for new students, we build a LightGBM classification model that predicts whether a student-instructor pair will result in a low compatibility evaluation. Input features include: (1) personality pattern codes, (2) aptitude test scores (attention, judgment, flexibility, etc.), and (3) instructor IDs.

The model is trained with class-imbalance correction using SMOTE (Chawla et al. 2002). Data from all collection years (2020–2025) are pooled and split by stratified random sampling (75% training, 25% test) to preserve the class ratio. Model selection was performed via 10-fold stratified cross-validation, in which LightGBM achieved the highest F1-score among nine candidate algorithms. The year-by-year positive-rate comparison (Figure 1) confirms that annual cohorts exhibit consistent compatibility patterns, supporting the validity of pooled random splitting. At the default decision threshold (0.5), mismatch detection is extremely conservative (mismatch precision 0.458, recall 0.007), which misses most mismatches.

Threshold optimization for mismatch recall. Because our goal is to *avoid* low-compatibility pairs, we tune the decision threshold instead of relying on the default 0.5. Figure 4 reports ROC/PR curves (top) and shows how mismatch recall/precision/F1 and overall accuracy change with the threshold (bottom). We select the threshold by grid search

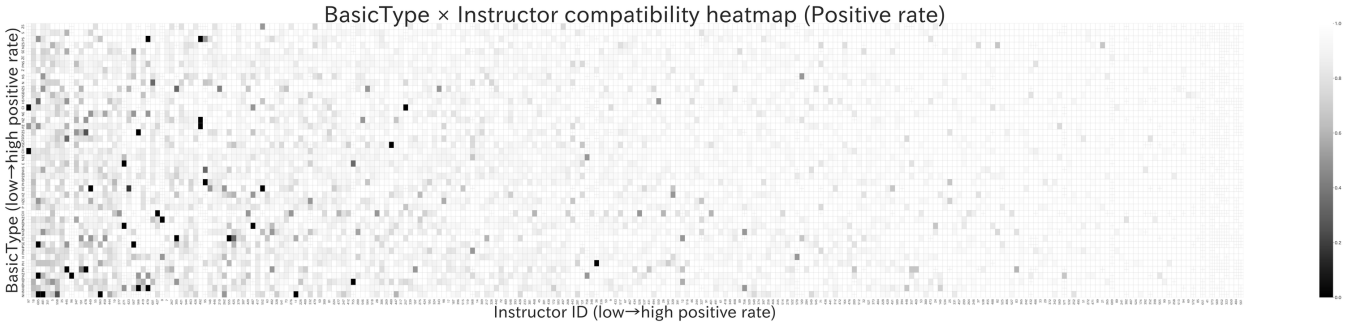

Figure 2: Compatibility heatmap showing positive rates for BasicType-instructor pairs. Darker regions indicate more mismatches; lighter regions indicate better compatibility. Columns with predominantly dark cells correspond to instructors who exhibit low compatibility across multiple BasicTypes.

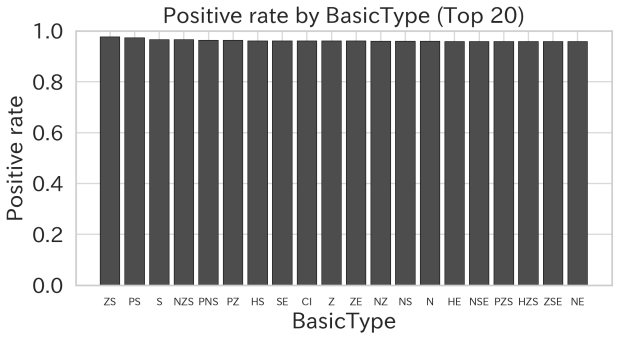

Figure 3: Positive rate by BasicType (top 20 by frequency).

Threshold 0.5 (default)		
	Pred mismatch (0)	**Pred good (1)**
True mismatch (0)	33	4,397
True good (1)	39	94,068
Threshold 0.925 (selected)		
	Pred mismatch (0)	**Pred good (1)**
True mismatch (0)	3,163	1,267
True good (1)	46,241	47,866

Table 2: Confusion matrices on the test split at the default threshold (0.5) and the selected threshold (0.925).

over 0.05–0.95 (step 0.005) using the strategy "maximize F1 subject to mismatch recall ≥ 0.70", yielding 0.925. On the test split, mismatch recall improves to 0.714 (precision 0.064; overall accuracy 0.52). Table 2 compares confusion matrices at thresholds 0.5 and 0.925.

We observed that instructor identity (InstructorID_enc) is the single most influential feature for mismatch prediction, followed by BasicType-related variables and student attributes such as age and aptitude-test sub-scores (Figure 5). This finding confirms that instructor-specific factors dominate the compatibility relationship.

Discussion

Our approach identifies potentially low-compatibility pairs while preserving the school's fairness rule of equal distribution. The threshold can be tuned to prioritize avoiding missed mismatches, with staff review as a safety net. Limitations include the inherent class imbalance of satisfaction data, the use of data from a single driving school, and the simplification of multi-level ratings to binary labels; future work should validate the approach across multiple schools and explore richer outcome measures such as learning progression speed and post-licensing safety records.

From the instructor's perspective, our analysis reveals that instructor identity is the most influential feature for mismatch prediction (Figure 5), and the compatibility heatmap (Figure 2) shows that some instructors exhibit consistently low compatibility across many BasicTypes. This indicates that the mismatch problem cannot be fully resolved through reassignment alone. The heatmap serves as a diagnostic tool for instructor development: instructors with broad mismatch patterns can receive targeted training in communication adaptability and pedagogical flexibility for specific personality types. This dual approach—reducing mismatches through intelligent assignment while simultaneously improving instructor capabilities through data-driven feedback—addresses root causes rather than merely managing symptoms.

Expected impact and threshold sensitivity. While direct measurement of rating changes after intervention remains future work, the confusion matrix at threshold 0.925 (Table 2) provides an estimate of expected impact: the system would correctly flag 71.4% (3,163/4,430) of actual mismatches for staff review, potentially preventing the majority of low-compatibility assignments. Regarding threshold sensitivity, Figure 4 shows that mismatch recall increases sharply as the threshold rises above 0.8, while precision decreases; the selected threshold of 0.925 balances recall ($\geq$0.70) with manageable alert volume. The system operates as an advisory tool—staff review flagged pairs and retain final decision authority—which mitigates the cost of false positives. Prospective validation of actual rating improvements after deployment is planned.

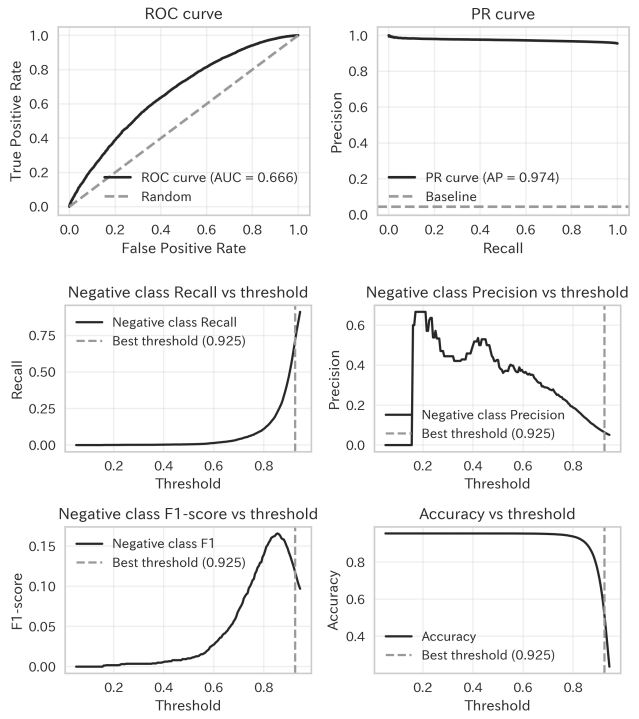

Figure 4: (Top) ROC and PR curves for the improved evaluation (ROC AUC=0.666, PR AP=0.974). (Bottom) Threshold sensitivity of mismatch detection: mismatch recall, precision, F1-score, and overall accuracy as functions of the decision threshold; the selected threshold is 0.925.

Human-AI co-evolution and Well-Being. Our system illustrates a concrete path toward co-evolving human and machine intelligence in an educational setting. On the machine side, staff decisions on flagged pairs and instructor feedback continuously refine the model's predictions. On the human side—and more fundamentally—the AI acts as a mirror: the compatibility heatmap reveals to each instructor their own interaction patterns across personality types, enabling a level of self-awareness that manual observation alone cannot provide. As instructors adapt their strategies, their improved practices generate new data that allow the model to discover higher-order patterns—exemplifying co-evolution rather than mere automation. From a well-being perspective, all stakeholders benefit: students receive better-matched instruction, instructors grow professionally, and better-trained drivers contribute to road safety. Looking forward, integrating post-licensing outcome data from the National Police Agency can further deepen this cycle.

Conclusion

We presented a real-world case study of using operational data to reduce student-instructor mismatches in a driving school via compatibility analysis and mismatch prediction. Threshold optimization enables a practical trade-off between missed mismatches and false alarms for risk-aware deployment. Our analysis reveals that instructor-specific factors are the primary drivers of compatibility, underscoring the value of combining algorithmic assignment support with targeted instructor development. Future work will validate the approach across multiple schools and evaluate actual rating improvements after deploying the mismatch-avoidance system.

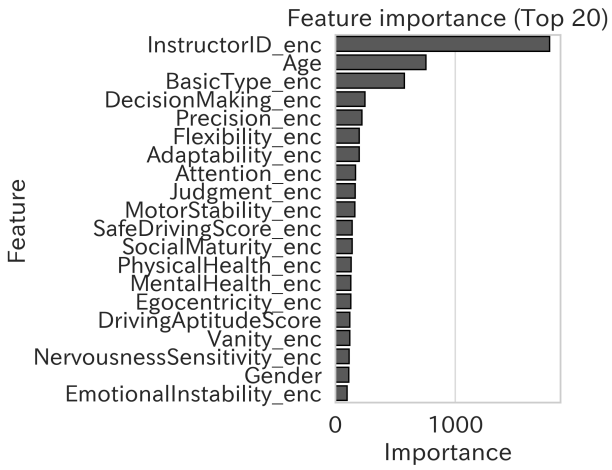

Figure 5: Top features for mismatch prediction (LightGBM). Instructor identity and BasicType are the most influential, followed by aptitude-test sub-scores.

Acknowledgments

We are grateful to the management and staff of Musashisakai Driving School for granting access to anonymized operational data and for their ongoing feedback and support throughout the project. We also thank the students who participated in the lessons and provided evaluations that made this study possible.

References

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Denno Co., Ltd. n.d. OD-Style Safety Aptitude Test for Driving School Students. https://www.dennoo.co.jp/S_introduction001.html. Accessed: 2026-01-21.

Fujihira, J. 2002. On the Evaluation of the OD-Style Safety Aptitude Test (1): Relationship with Traffic Accidents and Violations. *Research in Traffic Psychology*, 18: 44. Available at https://cir.nii.ac.jp/crid/1570009750814420352.