Can Teachers and Machines Predict the Long-Run Academic Performance of Young Children?

Max Coveney Sacha Kapoor Dinand Webbink

Erasmus University Rotterdam

December 2025

Motivation

- Teachers routinely assess students' long-run potential
- These assessments shape consequential decisions:
  - Time and effort allocation
  - Grades and recommendations
  - Track placement
- **Key question:** How accurate are these assessments?
- **Policy relevance:** Can algorithms improve or complement teacher judgment?

This Paper

Research Questions:

1. How well do teachers predict students' academic tracks?
2. Do prediction errors vary by student characteristics (SES, gender)?
3. Can machine learning algorithms outperform teachers?
4. Are human and algorithmic predictions complementary?

Contribution:

- First large-scale comparison of teacher vs. algorithmic predictions
- Unique data: teacher predictions at multiple time horizons
- Framework for understanding human-AI complementarity in education

Context: Dutch Education System

- Students allocated to academic tracks at age 12
- Four track levels (ordered by academic intensity):
  - Track 0: Vocational (VMBO-basis)
  - Track 1: Vocational-theoretical (VMBO-kader/gemengd/theoretisch)
  - Track 2: Professional (HAVO)
  - Track 3: Academic (VWO) $\rightarrow$ University
- Track assignment based on:
  - Teacher recommendation (highly influential)
  - Standardized test scores (CITO)

Data: PRIMA Cohort Study

Source: Primary Education Cohort Study (PRIMA), 1994–2005

Key features:

- 250,000+ teacher predictions about student track placement
- Predictions made at four time horizons:
  - Age 6 (6 years before track placement)
  - Age 8 (4 years before)
  - Age 10 (2 years before)
  - Age 12 (months before)
- Linked to realized track outcomes
- Rich student covariates: test scores, SES, gender, ethnicity

Sample: 600 schools, nationally representative + low-SES oversample

Empirical Framework

Setup:

- Teacher i predicts track $t_{ij}^{(y)*}$ for student j, y years ahead
- We observe realized track $t_{ij}^{(y)}$

Assumption: Teachers minimize expected loss from misclassification

$$t_j^{(y)*}(I_i) = \arg\min_t \mathbb{E}[L(t_{ij}^{(y)}, t)]$$

Loss functions:

- Binary (0-1): $L = \mathbf{1}[t_{ij}^{(y)} \neq t_j^{(y)*}]$
- Absolute: $L = |t_{ij}^{(y)} - t_j^{(y)*}|$
- Quadratic: $L = |t_{ij}^{(y)} - t_j^{(y)*}|^2$

Measuring Teacher Accuracy

Step 1: Compute loss per student

$$L_{ij}^{(y)} = L(t_{ij}^{(y)}, t_j^{(y)*}(I_i))$$

Step 2: Average loss per teacher

$$\bar{L}_i^{(y)} = \frac{1}{S(i)} \sum_{j \in C(i)} L_{ij}^{(y)}$$

Step 3: Analyze distribution of teacher accuracy

- How does accuracy vary across teachers?
- How does accuracy improve as prediction horizon shortens?

Classification Errors

Two types of errors:

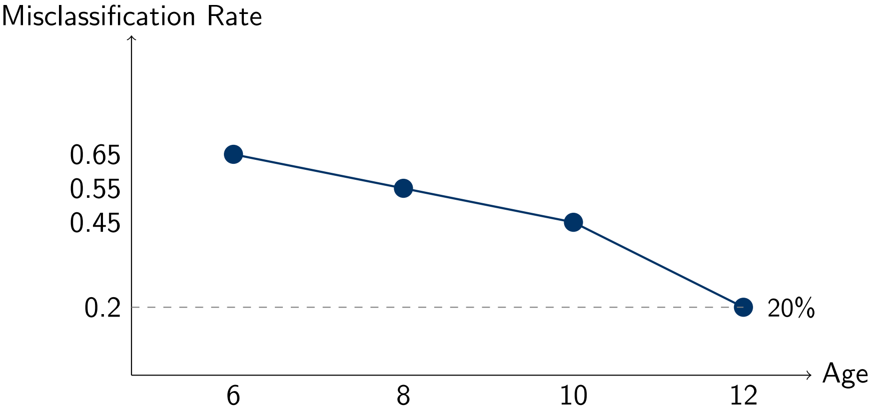
- **False Positive (Type I):** Teacher predicts higher track than realized
  - Overestimation of student potential
- **False Negative (Type II):** Teacher predicts lower track than realized
  - Underestimation of student potential

Rates:

$$FPR_i^{(y)} = \frac{FP_i^{(y)}}{TN_i^{(y)} + FP_i^{(y)}} \qquad FNR_i^{(y)} = \frac{FN_i^{(y)}}{TP_i^{(y)} + FN_i^{(y)}}$$

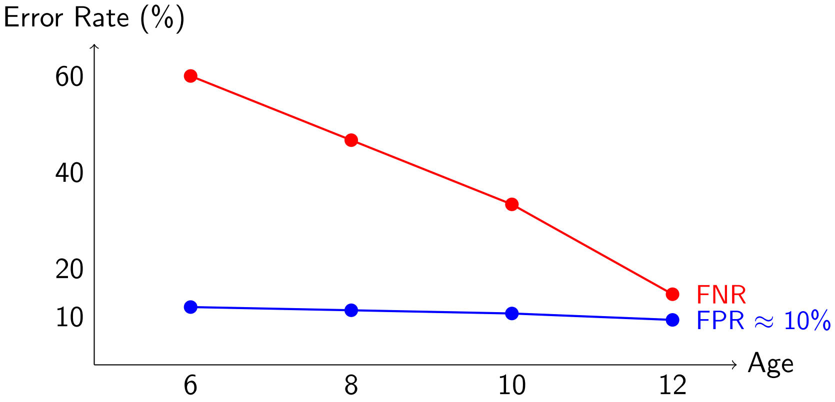
Result 1: Prediction Accuracy Improves with Age

Average Misclassification Rate (Binary Loss)



- Teachers misclassify 65% of 6-year-olds, 20% of 12-year-olds
- Largest improvement: ages 10 to 12 (recommendation already formed)

Result 2: Error Types Differ by Age



- **False positives** (overestimation): stable at ≈10%
- **False negatives** (underestimation): 60% at age 6 → 15% at age 12
- System is pessimistic about young students' potential

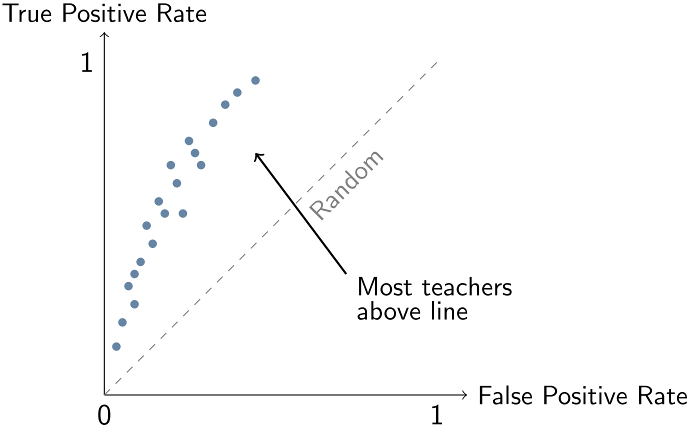
Result 3: Errors Differ by SES

SES Group	Age 6	Age 12
False Positive Rate (Overestimation)		
High SES Natives	19%	10%
Low SES Natives & Migrants	10%	5%
False Negative Rate (Underestimation)		
High SES Natives	55%	10%
Low SES Natives & Migrants	75%	20%

- Teachers **overestimate** high-SES students (2$\times$ higher FPR)
- Teachers **underestimate** low-SES students (FNR gap persists at age 12)

Result 4: Variation Across Teachers

ROC Analysis: Teacher Performance vs. Random Guessing



- Most teachers perform better than random guessing

Can Machines Do Better?

Approach:

- Train ML model to predict track using student observables X_{ij}
- Multinomial logistic regression with elastic net regularization
- Cross-validation to select penalty parameters

Features (X_{ij}):

- Test scores (language, math, reading comprehension)
- Student demographics (gender, ethnicity)
- Family background (parental education, SES)
- School characteristics

Key question: Does $t_j^{(y)*}(X_{ij})$ outperform $t_j^{(y)*}(I_i)$?

Machine vs. Teacher: Preliminary Results

	Misclassification Rate			
Predictor	Age 6	Age 8	Age 10	Age 12
Teacher	0.65	0.55	0.45	0.20
Machine (all features)	–	–	–	–
Machine (no test scores)	–	–	–	–
Complementarity				
Teacher + Machine	–	–	–	–

Analysis in progress

- Comparing out-of-sample prediction accuracy
- Testing whether combining teacher and machine predictions improves accuracy

Interpretation and Implications

What we learn:

- Teachers are informative but imperfect predictors
- Systematic biases: overestimate high-SES, underestimate low-SES
- System "learns" over time but biases persist

Policy implications:

- Algorithms could reduce bias in track recommendations
- But: Should we *want* accurate predictions?
  - Overestimating disadvantaged students might help them
  - Accurate early prediction could become self-fulfilling
- Human-AI complementarity: algorithms flag, humans decide

Broader Relevance: Human-AI Collaboration

This paper speaks to:

- When should organizations trust human vs. algorithmic judgment?
- How do we design systems that combine both effectively?

Related applications:

- Hiring decisions (Hoffman, Kahn & Li 2018)
- Bail decisions (Kleinberg et al. 2018)
- Medical diagnosis (Mullainathan & Obermeyer 2022)
- School quality assessment (Dutch Inspectorate)

Core insight: Neither humans nor machines dominate; optimal configuration is context-dependent

Conclusion

Summary:

1. Teachers misclassify 65% of 6-year-olds, 20% of 12-year-olds
2. Systematic bias: underestimate low-SES, overestimate high-SES
3. Bias persists even at age 12
4. Machine learning offers potential for improvement

Next steps:

- Complete ML comparison (in progress)
- Test complementarity of human + machine predictions
- Explore implications for track assignment policy

Thank you

kapoor@ese.eur.nl