

# The End of One-Size-Fits-All Testing:

## Personalizing Test Schedules with Recommender Systems

**Okan Bulut & Damien C. Cormier**

Centre for Research in Applied Measurement and Evaluation  
University of Alberta

**Jinnie Shin**

University of Florida

# Computer-Based Assessments in K-12



PISA



# Computer-Based **[Formative]** Assessments in K-12



Screening Assessments



Progress Monitoring Measures

# Flexibility Comes at a Price



# One-Size-Fits-All Test Scheduling

**Over-testing students**



**Loss of instructional time**

(Bulut, Cormier, & Shin, 2020)



# Our Study

- Can we build **a recommender system** that would guide teachers on finding an optimal testing schedule for each student?
- **Goal:** To create an optimal testing schedule for each student
  - The number of assessments
  - The frequency of assessments
  - Time of test administration



# Data Source



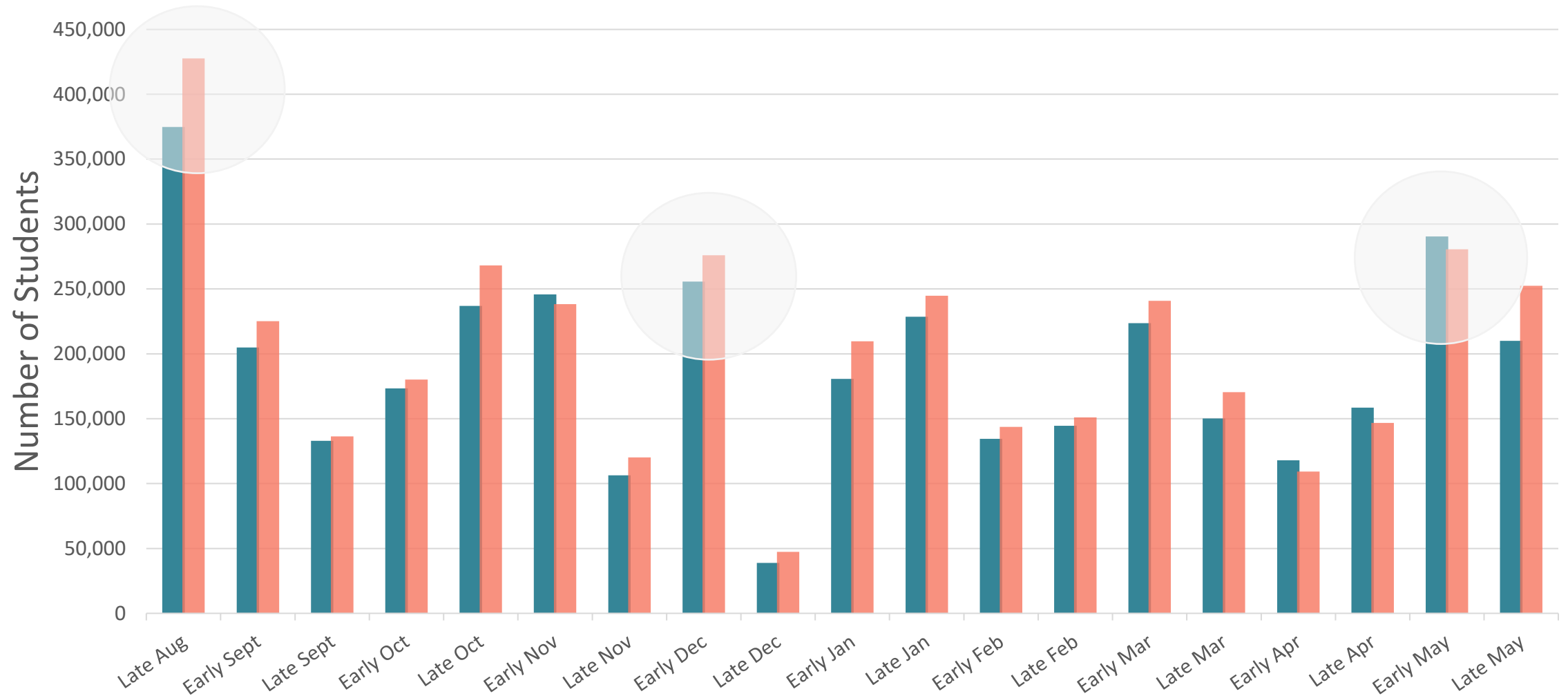
Star Reading by  
Renaissance



Grade 2 ( $n = 668,324$ )  
Grade 4 ( $n = 727,147$ )

# Test Administration Windows (Mid-August to Late May)

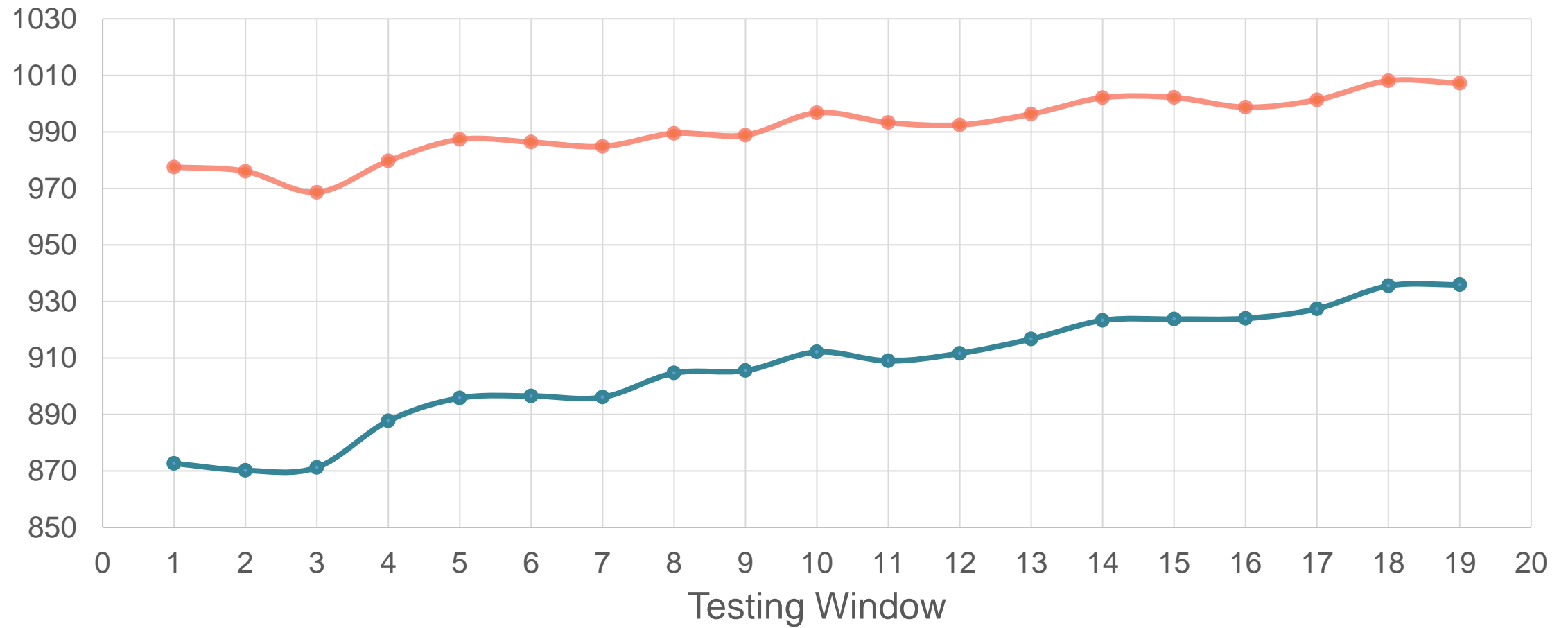
■ Grade 2 ■ Grade 4





## Average Scores in Star Reading

Grade 2 Grade 4

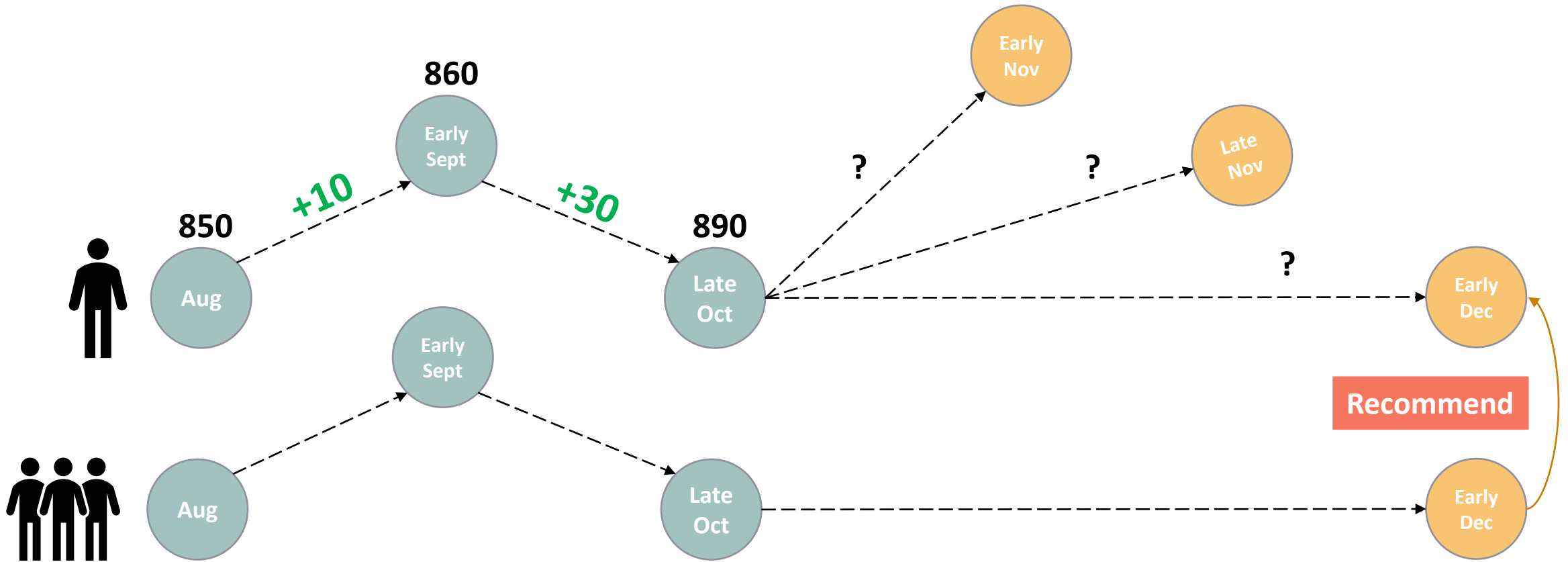


# Data Preparation

1. Find students in the data set whose:
  - Growth trajectories are positive
  - Last Star Reading scores  $>$  the cut-off score for the 25<sup>th</sup> percentile
  - Slope estimates (i.e., daily score increase)  $>$  the median slope for the sample
2. Split the data into training (students from Step #1) and test (remaining students) sets:
  - **Grade 2:** 276,087 (Training); 392,237 (Test)
  - **Grade 4:** 278,442 (Training); 448,705 (Test)

# Recommender System

- Dijkstra's Shortest Path First (SPF) algorithm
  - **Maximize** the (positive and absolute) score change between test administrations
  - **Minimize** the number of test administrations



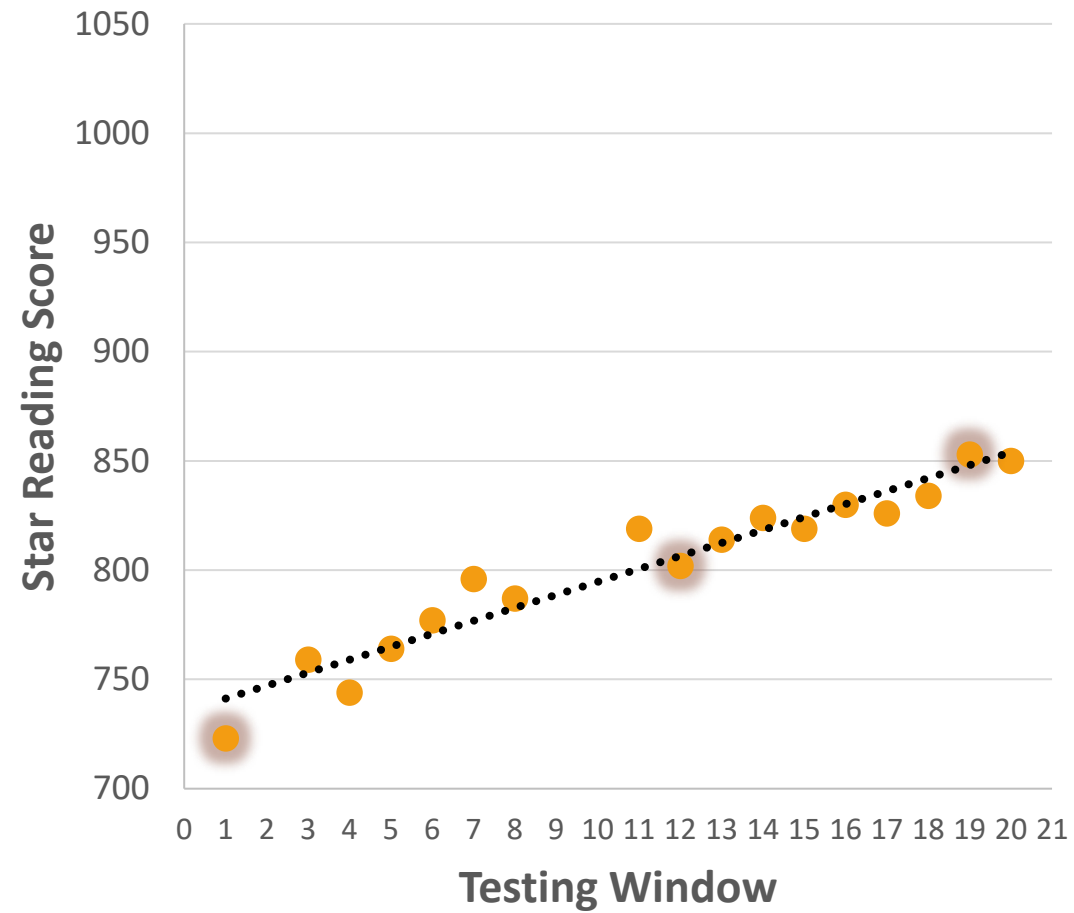
- Find similar students (with max score change + fewest tests)
- Select the most similar student based on Euclidean distance

# Results

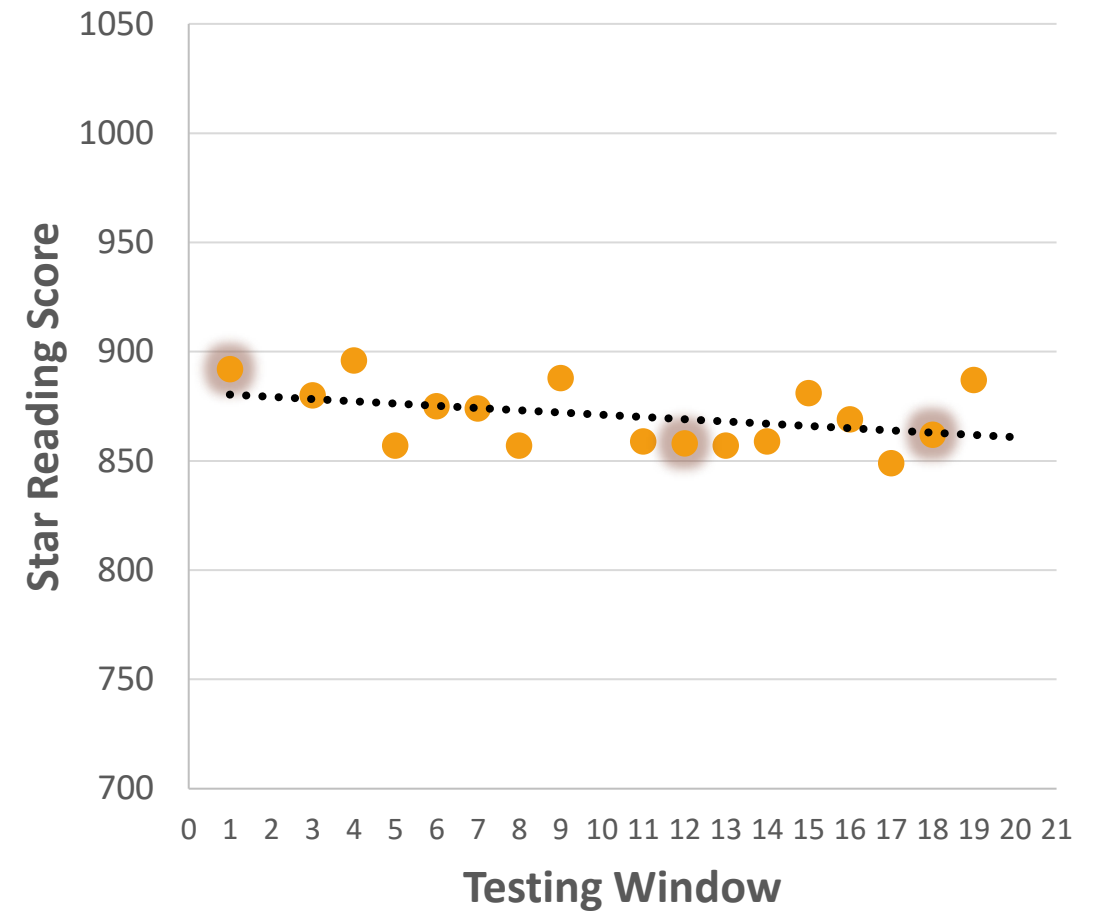
**Standard Practice** = Schedules Determined by Teachers    **RS** = Recommender System

Evaluation Criteria	Grade 2		Grade 4	
	Standard Practice	RS	Standard Practice	RS
Average number of tests administered	5.42	→ <b>3.51</b>	5.37	→ <b>3.84</b>
Average score change between tests	8.32	→ <b>12.25</b>	3.49	→ <b>4.63</b>
Range of tests required	(1, 18)	→ (1, 5)	(1, 17)	→ (1, 6)
Non-recommendable cases	-	0.05%		0.10%

## Student 1



## Student 2



● Actual test administrations

● Recommended test administrations

# Summary

- The **one-size-fits-all** approach → Students are being over-tested with computerized formative assessments (up to 18 tests per student).
- Our recommender system can pinpoint the optimal time, number, and frequency of test administrations for each student.
- Individualized test administration → personalized learning in schools

# Thank You!

For further information, please contact:

**Okan Bulut**

**bulut@ualberta.ca**