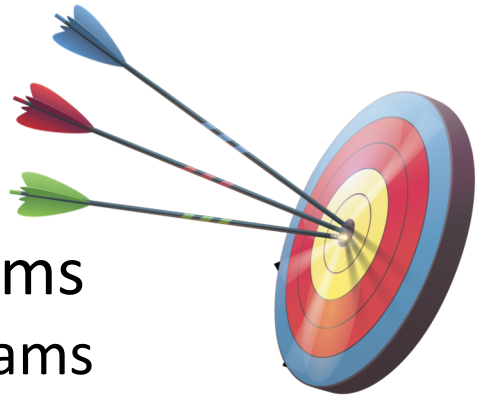
Are They Learning or Guessing?
Investigating Trial-and-Error Behavior with Limited Test Attempts

Bowen Hui

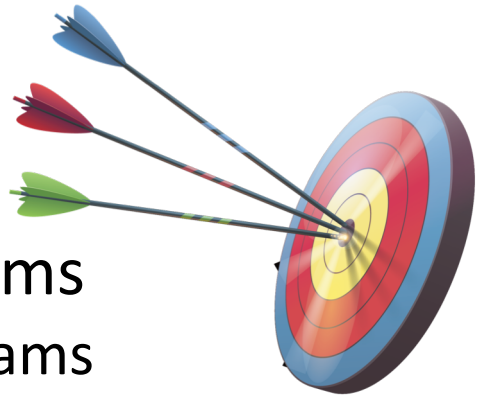
Computer Science
University of British Columbia
Okanagan Campus

Motivation



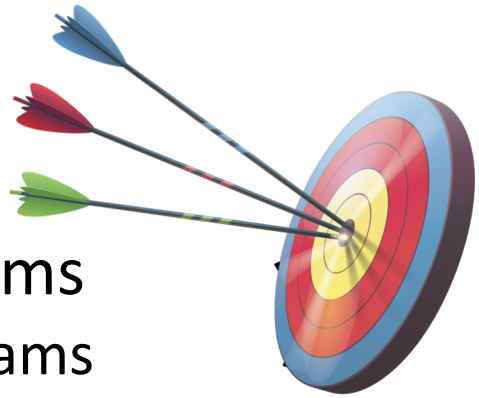
- Prevalent use of mastery learning platforms
  - Gives multiple chances on assignments/exams
  - Focus on deliberate practice until mastery

Motivation



- Prevalent use of mastery learning platforms
  - Gives multiple chances on assignments/exams
  - Focus on deliberate practice until mastery
- Varying implementations of resubmission policy
  - When unlimited attempts allowed, studies found students over-submit and engage in trial-and-error behavior
  - Application of regression penalties have found less guessing, but negatively creates exam anxiety

Motivation



- Prevalent use of mastery learning platforms
  - Gives multiple chances on assignments/exams
  - Focus on deliberate practice until mastery
- Varying implementations of resubmission policy
  - When unlimited attempts allowed, studies found students over-submit and engage in trial-and-error behavior
  - Application of regression penalties have found less guessing, but negatively creates exam anxiety
- Open questions:
  - How many attempts should be given?
  - How much guessing is actually there?
    - Our focus: Tests, max 3 attempts over 3 weeks, best score only

Our Research Questions

1. What are the general test-taking patterns and performance levels in this mastery learning environment?
   - Learning gains between pre vs. post-test?
   - When are test attempts made?

2. What can we observe about the behavior surrounding subsequent attempts?
   - Relationship to performance?
   - When are subsequent attempts made?

3. How might we model guessing behavior using attempt sequences and what are the implications?
   - Dynamic model?
   - How much guessing happened?

Goal: Improve course design

Related Work

- Large body of literature on designing assessments with MCQs [Fellenz, 2004; Harper, 2003]

- Low-stakes assessments refer to non-credit exams that typically measure student aptitude for cross-institutional comparison
  - Students not motivated and do not take them seriously [Noorbehbahani et al., 2022; Silm et al., 2013; Silm et al., 2020; Wise & DeMars, 2005]
  - Increased guessing behavior over the years [Must & Must, 2013]

Related Work on Guessing Behavior

- Solution behavior vs. rapid guessing behavior [Schnipke, 1995]

Related Work on Guessing Behavior

- Solution behavior vs. rapid guessing behavior [Schnipke, 1995]
- Rapid guessing behavior in low-stakes tests [Wise & Kong, 2005]

Related Work on Guessing Behavior

- Solution behavior vs. rapid guessing behavior [Schnipke, 1995]
- Rapid guessing behavior in low-stakes tests [Wise & Kong, 2005]
- Lots of work on measuring time to estimate when students are guessing [Wise, 2017; Kong et al., 2007]
  - Post-hoc analysis of visual inspection of response time distribution
  - Calculation of surface features of test item
  - Pre-defined threshold (3-5 seconds per item)
  - Mixture model of response times and accuracy
  - *All involve item analyses and comparable to threshold method*

Related Work on Guessing Behavior

- Solution behavior vs. rapid guessing behavior [Schnipke, 1995]
- Rapid guessing behavior in low-stakes tests [Wise & Kong, 2005]
- Lots of work on measuring time to estimate when students are guessing [Wise, 2017; Kong et al., 2007]
  - Post-hoc analysis of visual inspection of response time distribution
  - Calculation of surface features of test item
  - Pre-defined threshold (3-5 seconds per item)
  - Mixture model of response times and accuracy
  - *All involve item analyses and comparable to threshold method*

- Guessing behavior associated with specific items
  - Longer text or occurring later in the test [Wise et al., 2009; Demars, 2007]
  - Less guessing when item has table or image [Wise et al., 2009]

Related Work on Guessing Behavior

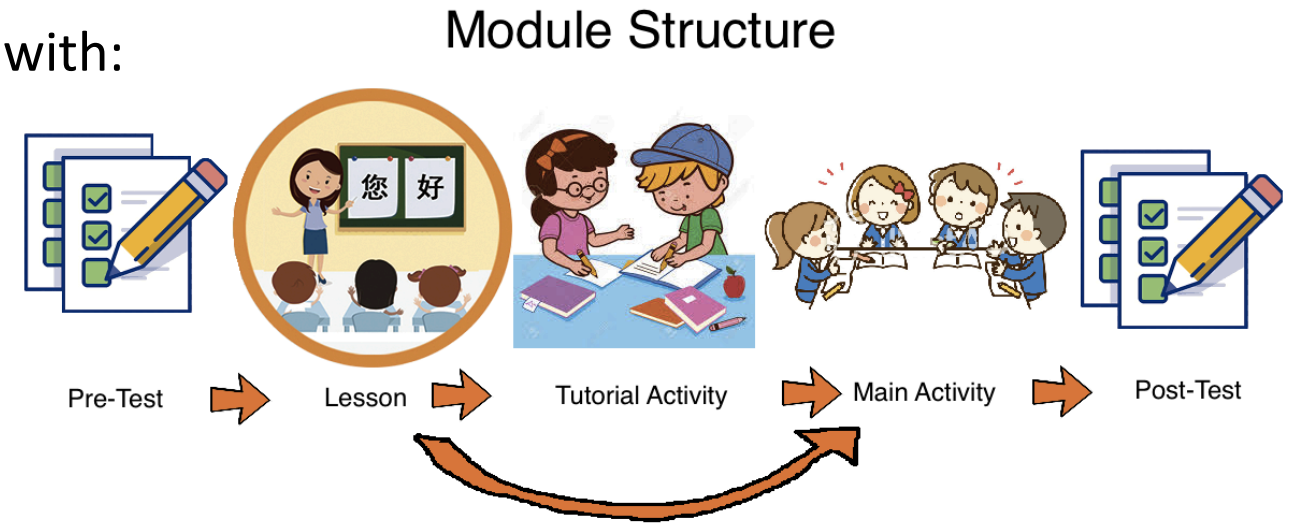
- Solution behavior vs. rapid guessing behavior [Schnipke, 1995]
- Rapid guessing behavior in low-stakes tests [Wise & Kong, 2005]
- Lots of work on measuring time to estimate when students are guessing [Wise, 2017; Kong et al., 2007]
  - Post-hoc analysis of visual inspection of response time distribution
  - Calculation of surface features of test item
  - Pre-defined threshold (3-5 seconds per item)
  - Mixture model of response times and accuracy
  - *All involve item analyses and comparable to threshold method*

- Guessing behavior associated with specific items
  - Longer text or occurring later in the test [Wise et al., 2009; Demars, 2007]
  - Less guessing when item has table or image [Wise et al., 2009]
- Unaware of work on modeling guessing behavior when tests have repetition or longer test-taking windows (common to mastery learning)

Course Context

- Third-year undergrad Human-Computer Interaction (HCI) course with diverse student backgrounds

- 10 modules, each with:
  - Pre-test
  - Content
  - Tutorial activity
  - Group activity
  - Post-test



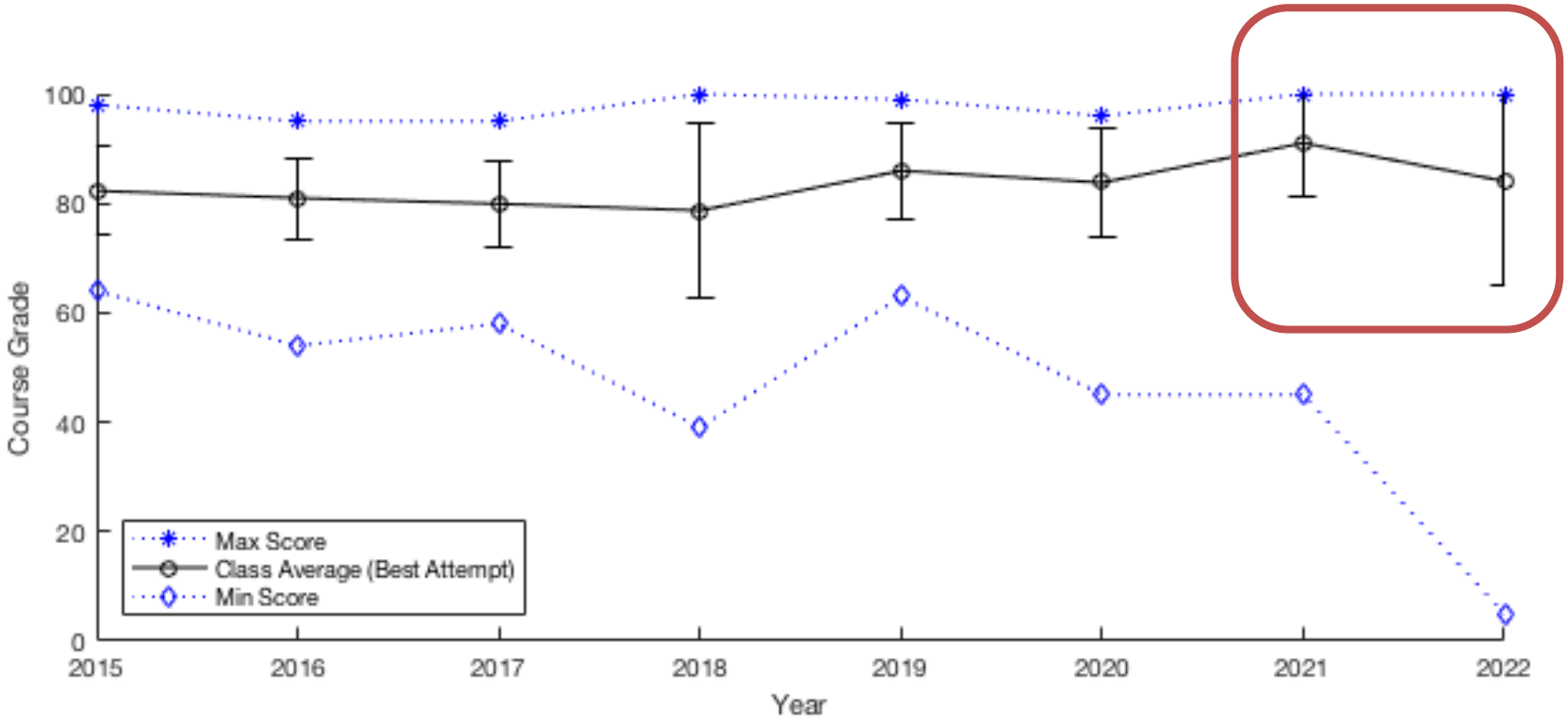
- Participants:
  - Winter 2021: 160 students (29 females; 131 males)
  - Winter 2022: 199 students (29 females, 170 males)
  - Total: 359 students

Data

- Total 104 questions in 20 tests
  - Average 5 questions per test
    - Most questions were MCQ (~70 words)
    - Most questions had 4 response options (~33 words)
  - Among these, 37 questions had images and 4 had tables
  - *Guessing is likely due to knowledge gaps rather than boredom*
- Delivered with all questions at once on Canvas LMS
  - *Could not get item-level statistics*

RQ1: General Test-Taking Patterns and Performance

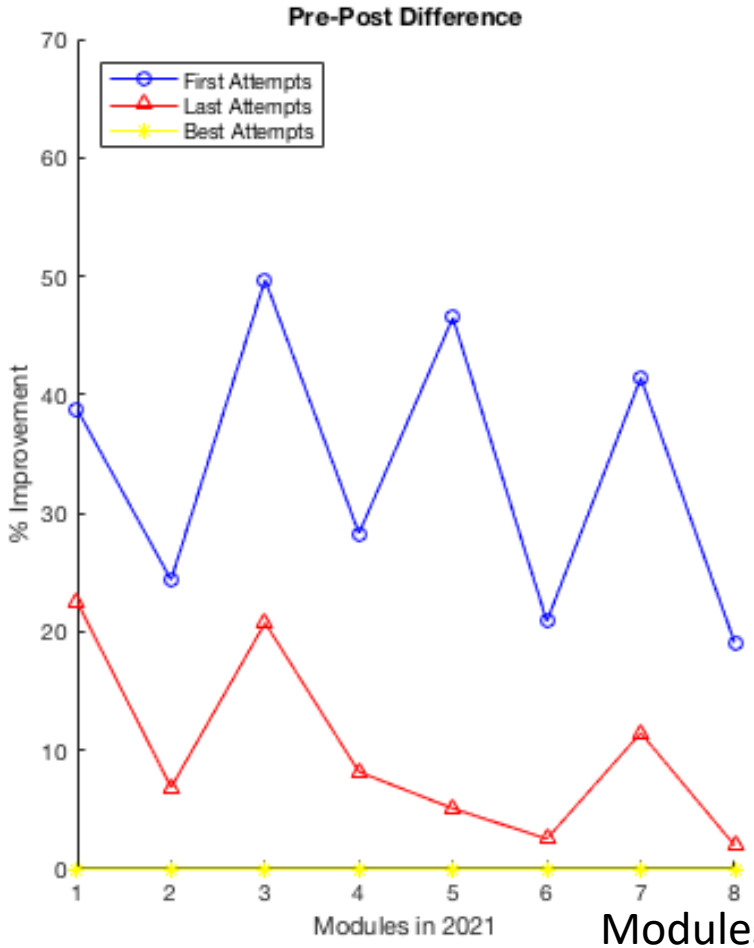
- Overall performance improvement?

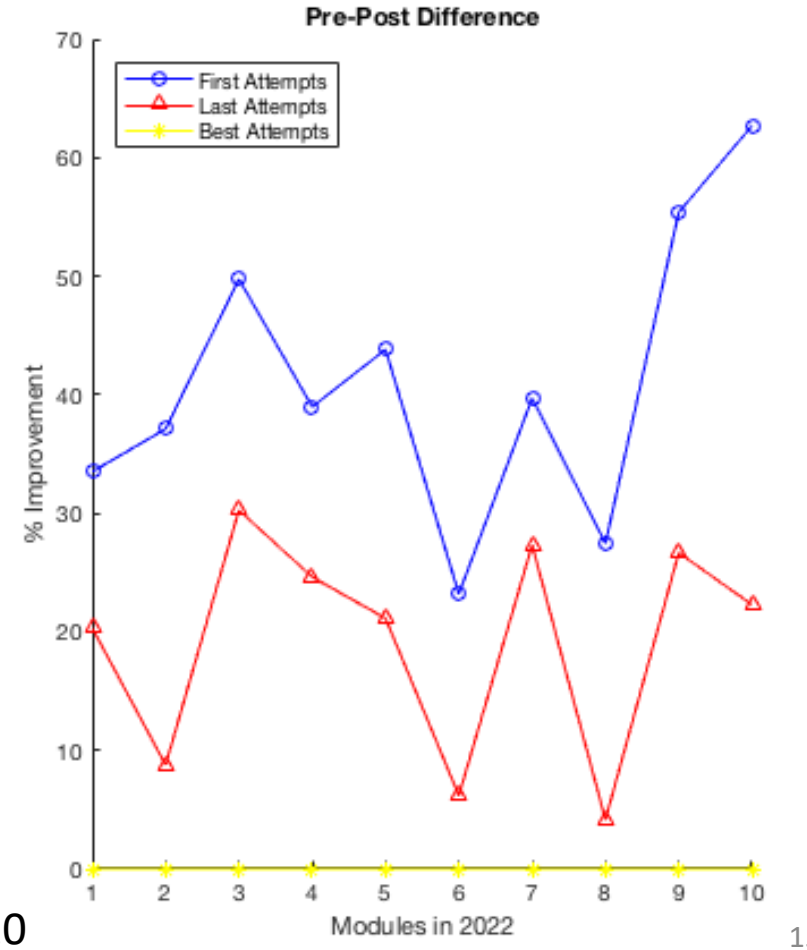


Significant improvement in 2021

RQ1: General Test-Taking Patterns and Performance

- Learning gains per module?

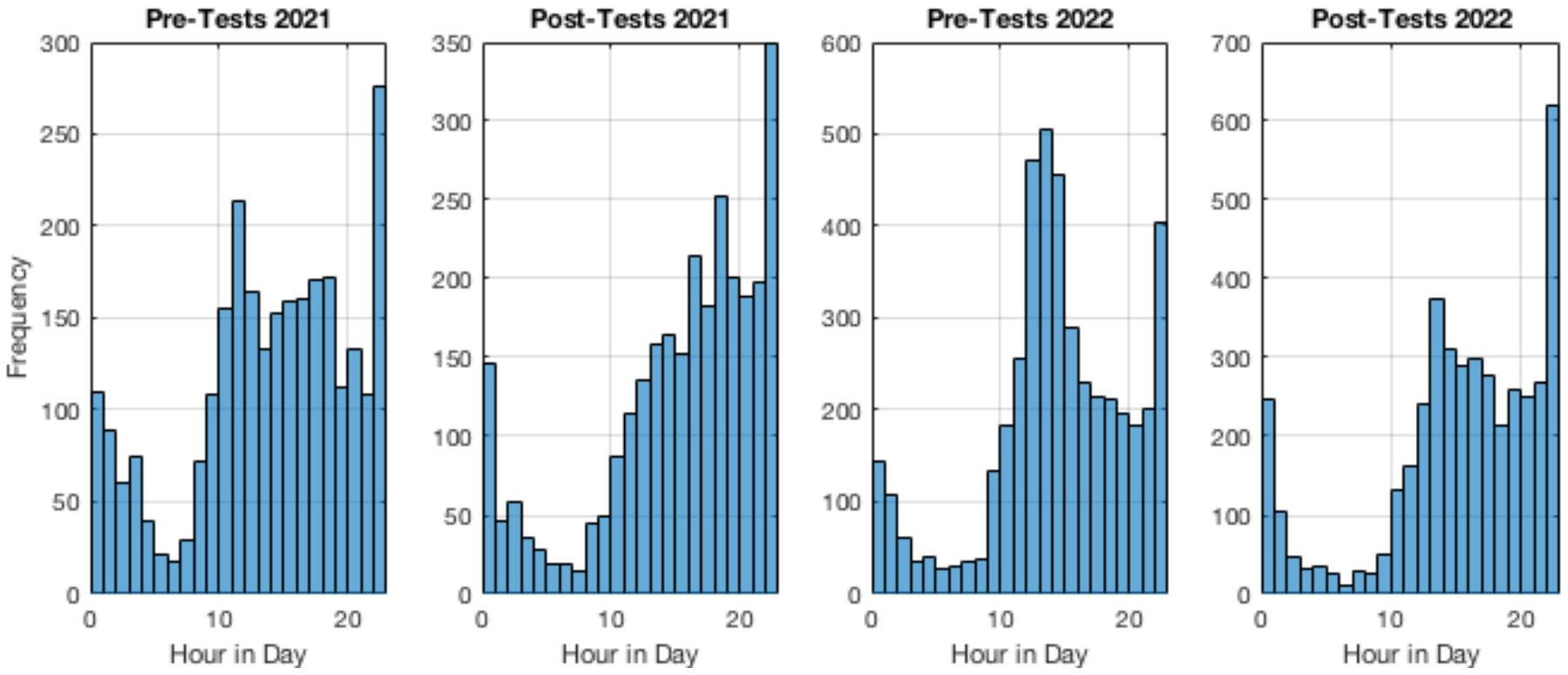




Modules 9, 10 were cancelled

RQ1: General Test-Taking Patterns and Performance

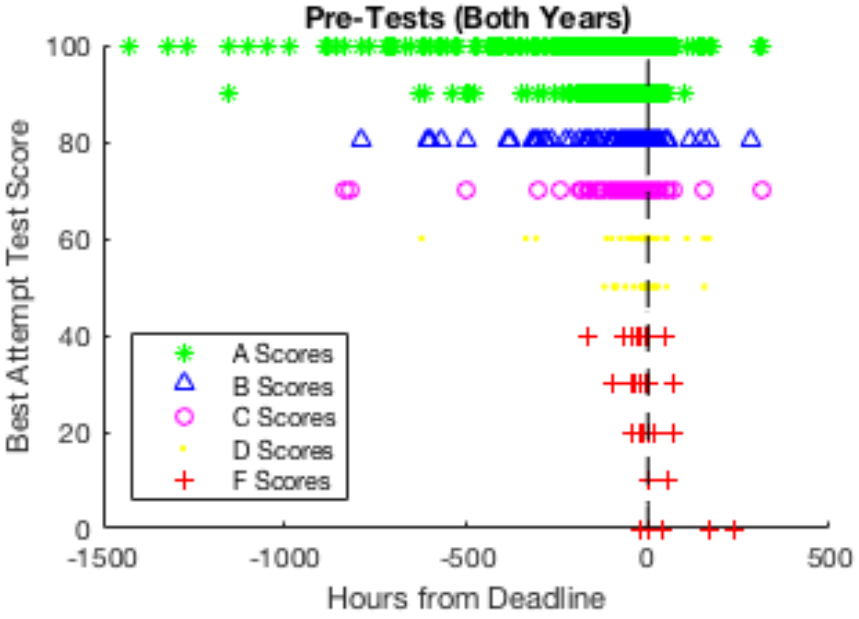
- When do students take tests?

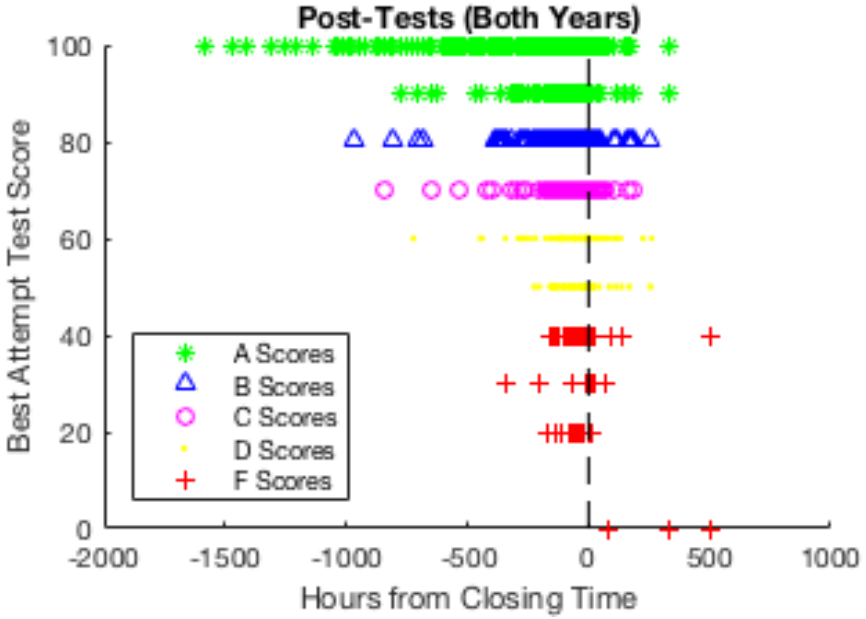


10:30 deadline | Closes 23:59 | 14:00 deadline | Closes 23:59

RQ1: General Test-Taking Patterns and Performance

- Submission time relative to deadline?

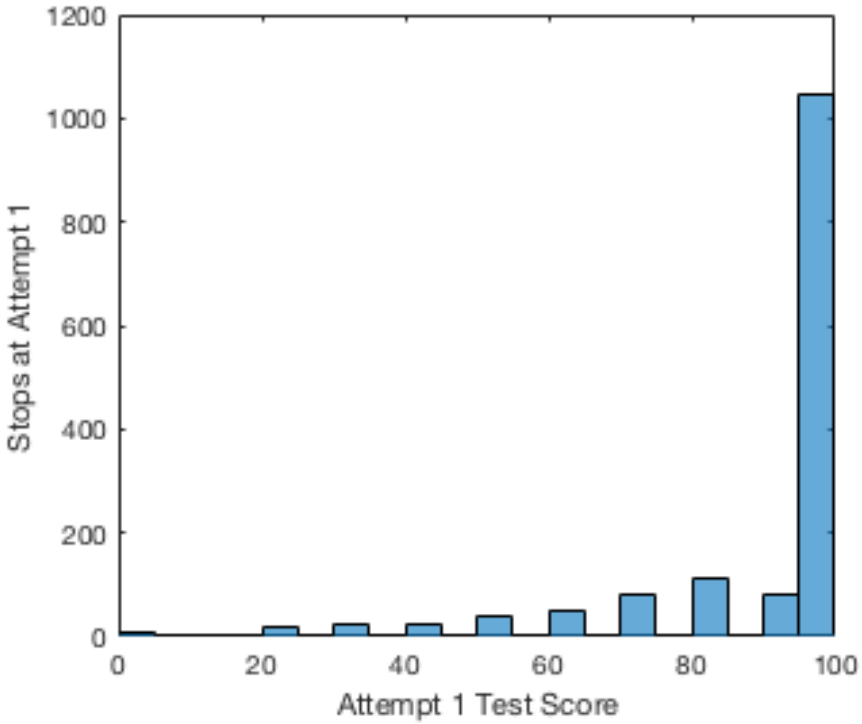


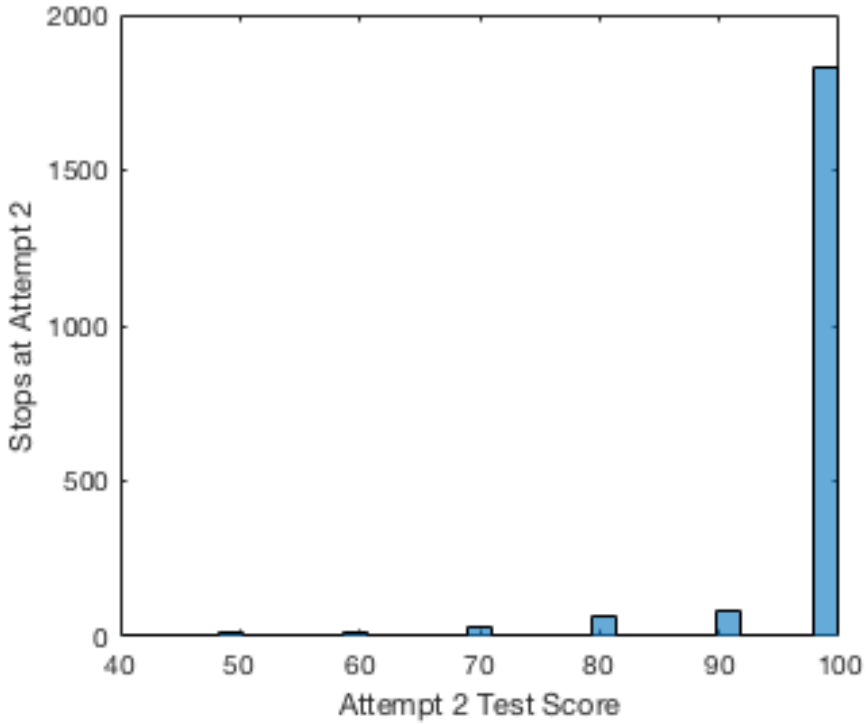


Many A students start early, but not all.
Most low-performing students start close to the deadline.

RQ2: Behavior Surrounding Subsequent Attempts

- Are subsequent attempts dependent on performance?

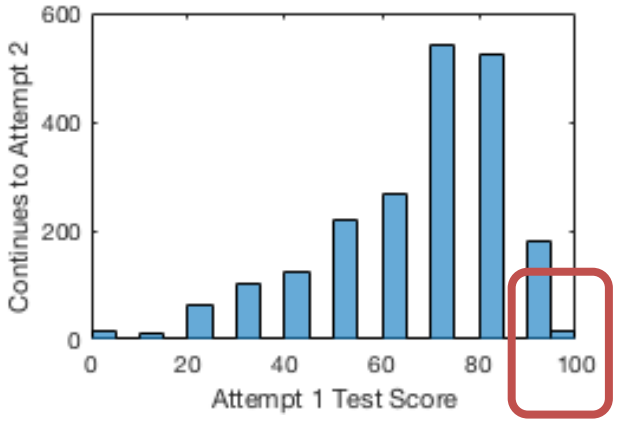


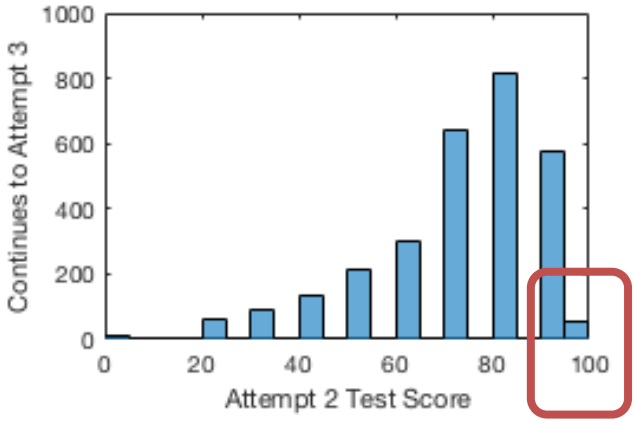


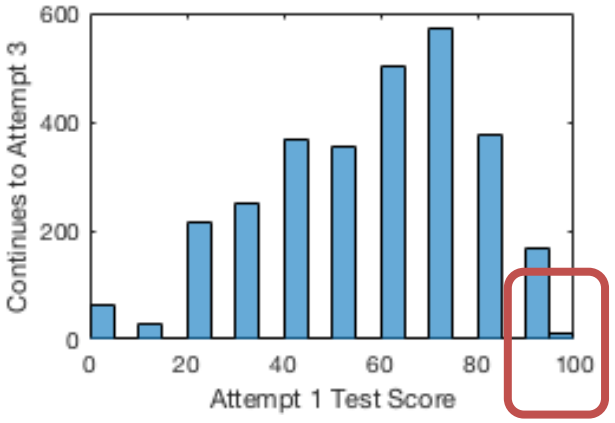
Most students who get 100% don't bother taking another attempt after. Students who only make one attempt tend to submit closer to the deadlines than students who stop early.

RQ2: Behavior Surrounding Subsequent Attempts

- If subsequent attempt is taken, is it due to an imperfect score?



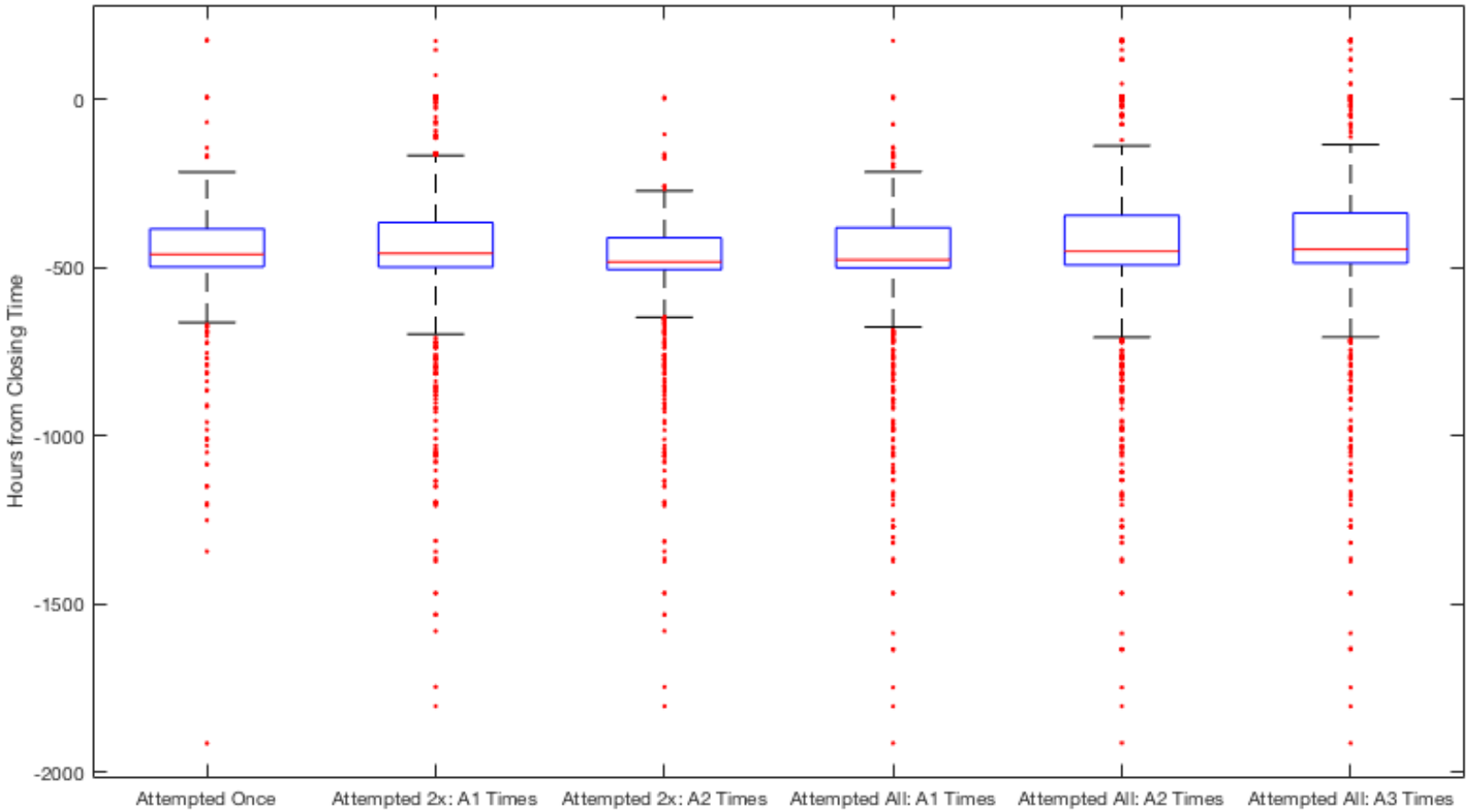




Some students who get perfect still make a subsequent attempt.
18-100% of the instances result in a lower mark on the future attempt.
Suggests exploratory learning behavior.

RQ2: Behavior Surrounding Subsequent Attempts

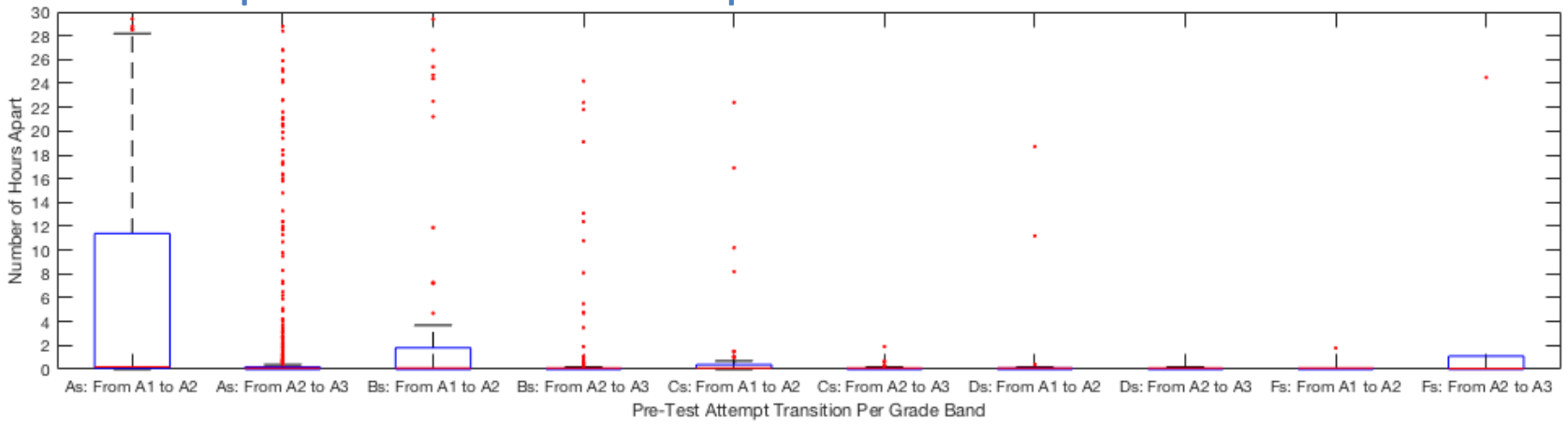
- When are subsequent attempts made?



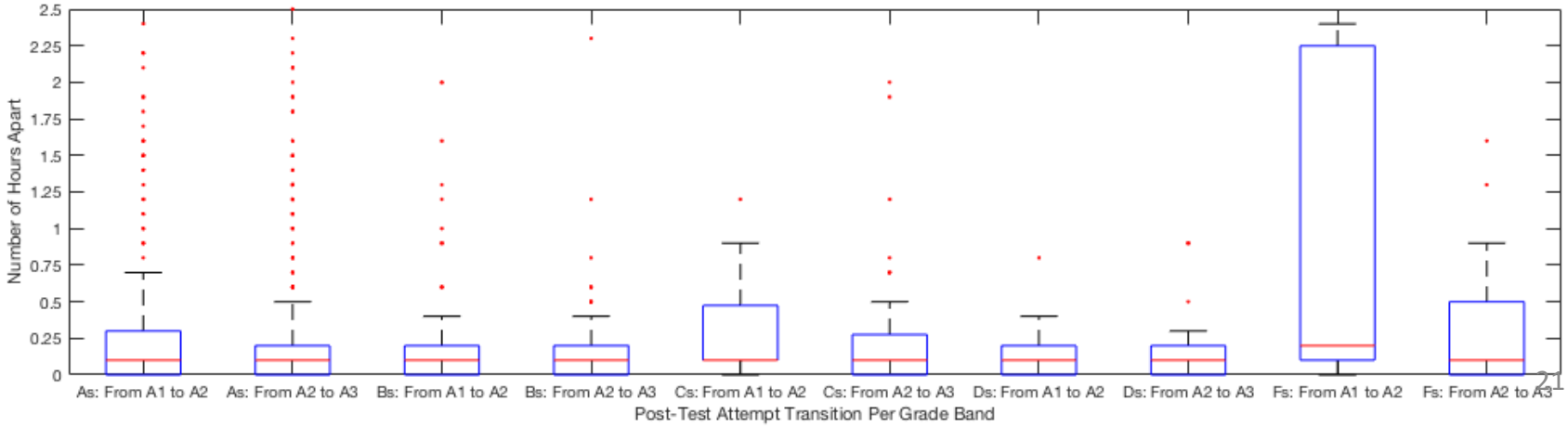
Students are not making full use of the 3-week window.

RQ2: Behavior Surrounding Subsequent Attempts

- Hours apart between attempts?

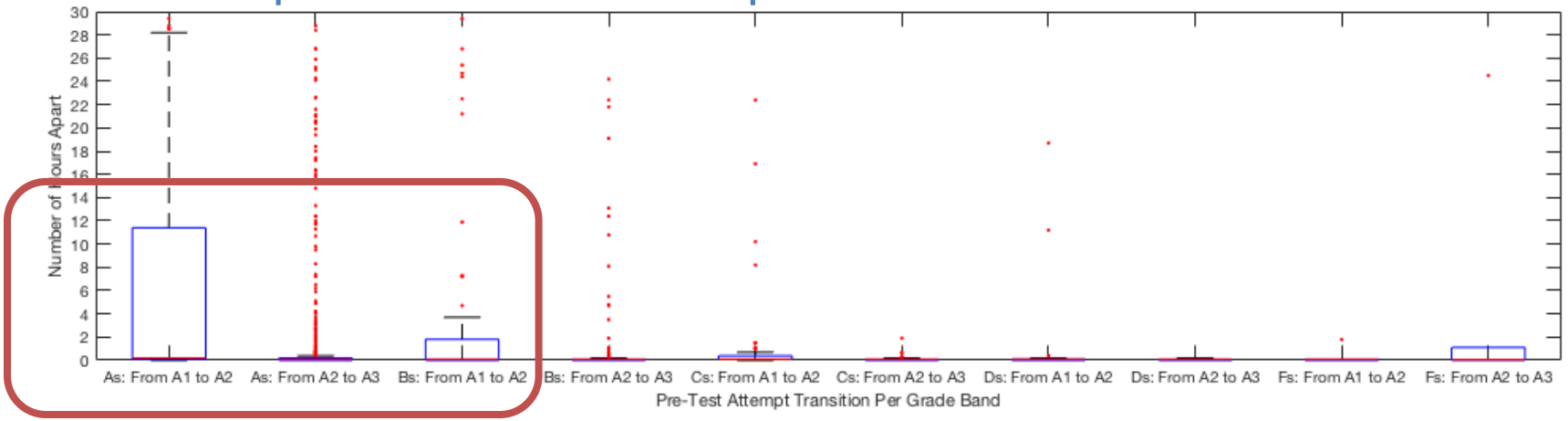


Median of 6 minutes between attempts.

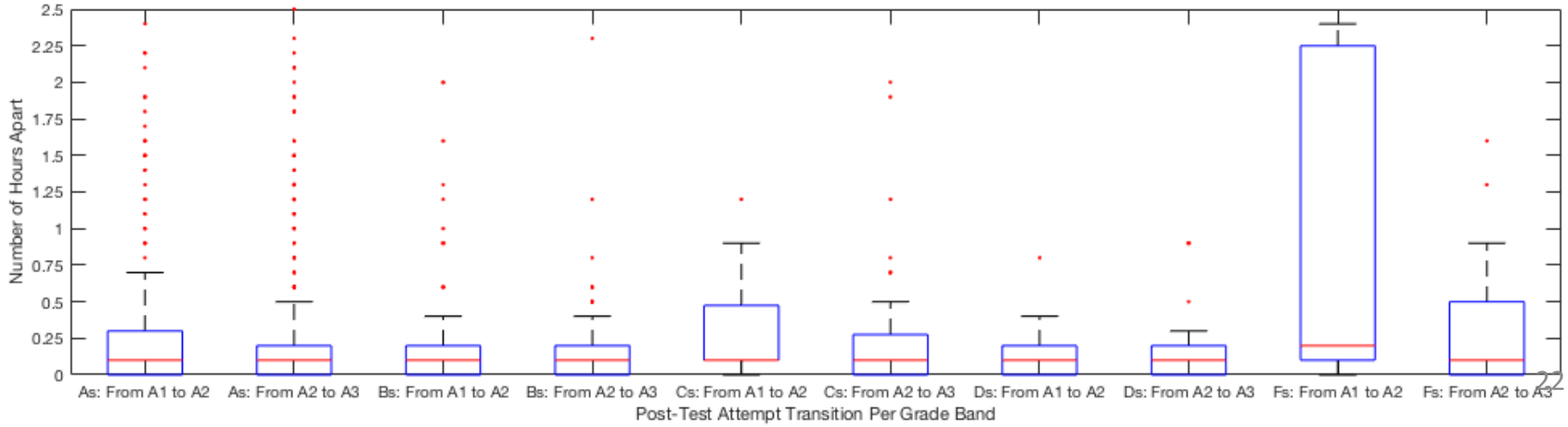


RQ2: Behavior Surrounding Subsequent Attempts

- Hours apart between attempts?

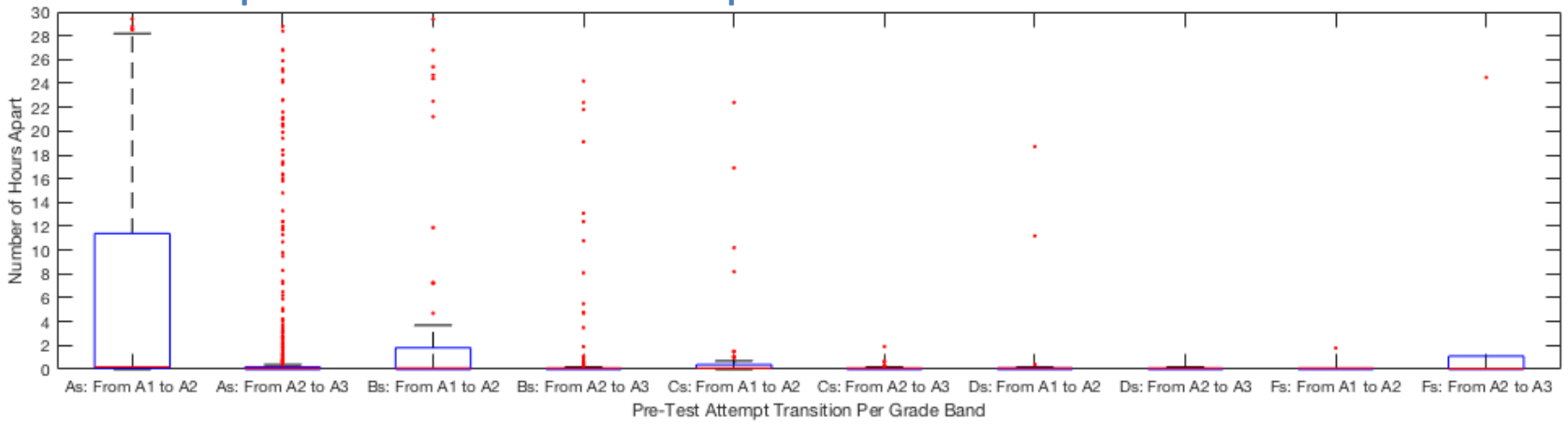


High-performing students reflect between A1 and A2 on pre-tests.

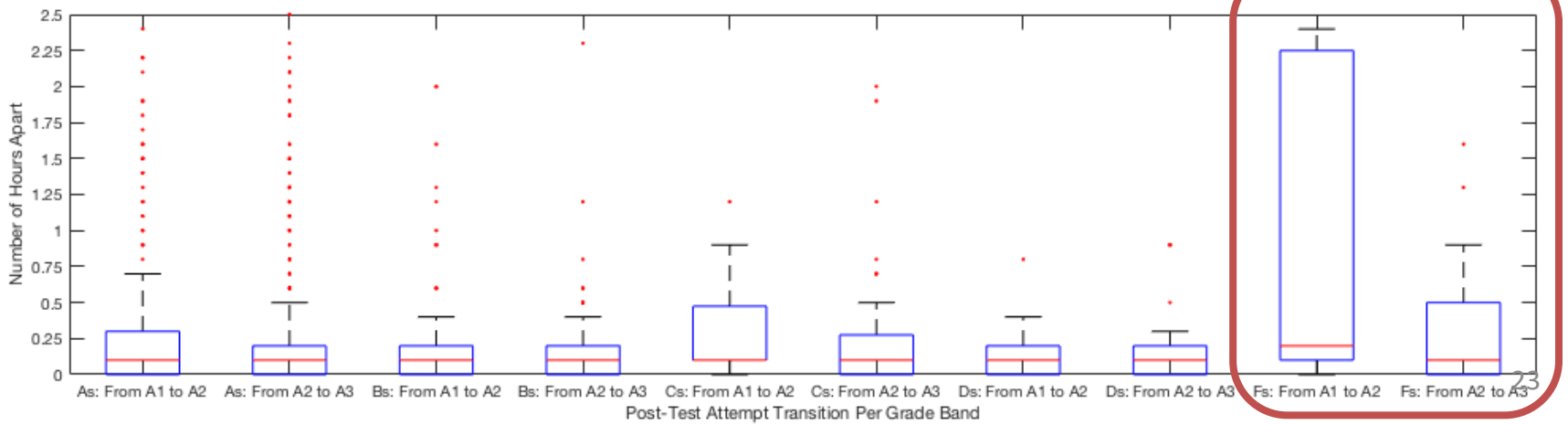


RQ2: Behavior Surrounding Subsequent Attempts

- Hours apart between attempts?

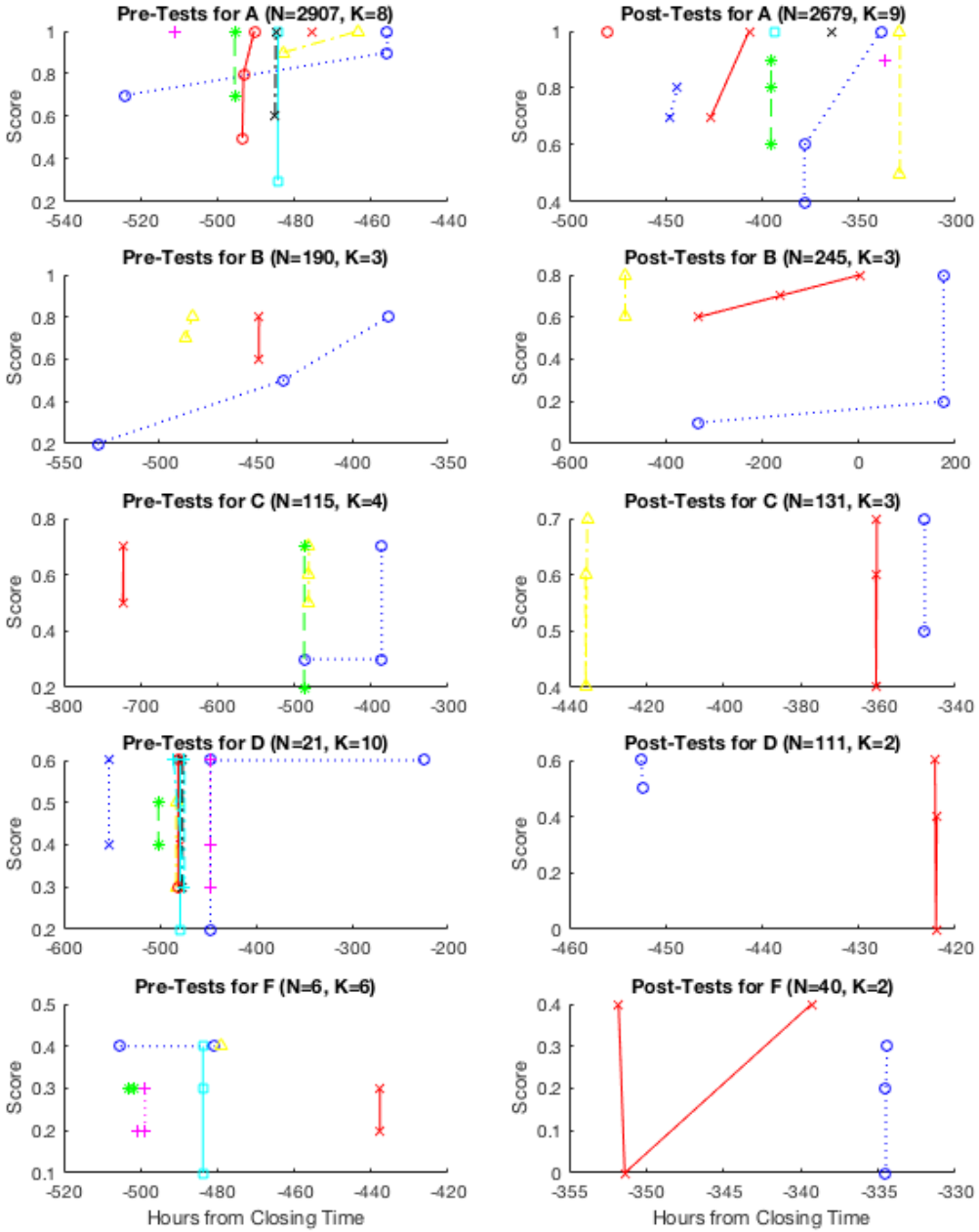


F-students are wheel spinning on post-tests.



RQ3: Analyzing Guessing Behavior Using Attempt Sequences

- How might we model trial-and-error sequences?



K-medoids clustering by test type and grade based on 1,000 trials.
Model: Threshold of 6-minutes and subsequent decrease in performance.

RQ3: Analyzing Guessing Behavior Using Attempt Sequences

- How much guessing is present?

Sample counts from two modules

Module	Test	Grade	Guesses	Improvements	Regression	Trial-and-Errors	Totals
3	Pre-Test	A	34	88	13.8%	31	318 (89.1%)
		B	20	13	10.0%	19	33 (9.2%)
		C	0	1	0%	0	2 (0.6%)
		D	2	1	30.0%	2	2 (0.6%)
		F	1	3	0%	1	2 (0.6%)
		All	57 (16.0%)	106 (29.7%)		53 (14.8%)	357
3	Post-Test	A	22	112	14.3%	20	307 (86.7%)
		B	16	9	14.5%	13	25 (7.1%)
		C	11	7	20.0%	10	16 (4.5%)
		D	2	1	10.0%	1	4 (1.1%)
		F	1	1	10.0%	1	2 (0.6%)
		All	52 (14.7%)	130 (36.7%)		45 (12.7%)	354
10	Pre-Test	A	10	51	12.9%	10	160 (82.1%)
		B	12	3	10.0%	9	11 (5.6%)
		C	19	11	20.0%	13	24 (12.31%)
		D	0	0	0%	0	0
		F	0	0	0%	0	0
		All	41 (21.0%)	65 (33.3%)		32 (16.4%)	195
10	Post-Test	A	10	53	12.5%	10	176 (91.2%)
		B	6	6	10.0%	5	8 (4.2%)
		C	6	3	40.0%	5	6 (3.1%)
		D	4	0	20.0%	2	2 (1.0%)
		F	2	0	0.0%	1	1 (0.5%)
		All	28 (14.5%)	62 (32.1%)		23 (11.9%)	193

RQ3: Analyzing Guessing Behavior Using Attempt Sequences

- How much guessing is present?

Most eventually get an A

Module	Test	Grade	Guesses	Improvements	Regression	Trial-and-Errors	Totals
3	Pre-Test	A	34	88	13.8%	31	318 (89.1%)
		B	20	13	10.0%	19	33 (9.2%)
		C	0	1	0%	0	2 (0.6%)
		D	2	1	30.0%	2	2 (0.6%)
		F	1	3	0%	1	2 (0.6%)
		All	57 (16.0%)	106 (29.7%)		53 (14.8%)	357
3	Post-Test	A	22	112	14.3%	20	307 (86.7%)
		B	16	9	14.5%	13	25 (7.1%)
		C	11	7	20.0%	10	16 (4.5%)
		D	2	1	10.0%	1	4 (1.1%)
		F	1	1	10.0%	1	2 (0.6%)
		All	52 (14.7%)	130 (36.7%)		45 (12.7%)	354
10	Pre-Test	A	10	51	12.9%	10	160 (82.1%)
		B	12	3	10.0%	9	11 (5.6%)
		C	19	11	20.0%	13	24 (12.31%)
		D	0	0	0%	0	0
		F	0	0	0%	0	0
		All	41 (21.0%)	65 (33.3%)		32 (16.4%)	195
10	Post-Test	A	10	53	12.5%	10	176 (91.2%)
		B	6	6	10.0%	5	8 (4.2%)
		C	6	3	40.0%	5	6 (3.1%)
		D	4	0	20.0%	2	2 (1.0%)
		F	2	0	0.0%	1	1 (0.5%)
		All	28 (14.5%)	62 (32.1%)		23 (11.9%)	193

RQ3: Analyzing Guessing Behavior Using Attempt Sequences

- How much guessing is present?

Module	Test	Grade	Guesses	Improvements	Regression	Trial-and-Errors	Totals
3	Pre-Test	A	34	88	13.8%	31	318 (89.1%)
		B	20	13	10.0%	19	33 (9.2%)
		C	0	1	0%	0	2 (0.6%)
		D	2	1	30.0%	2	2 (0.6%)
		F	1	3	0%	1	2 (0.6%)
		All	57 (16.0%)	106 (29.7%)		53 (14.8%)	357
3	Post-Test	A	22	112	14.3%	20	307 (86.7%)
		B	16	9	14.5%	13	25 (7.1%)
		C	11	7	20.0%	10	16 (4.5%)
		D	2	1	10.0%	1	4 (1.1%)
		F	1	1	10.0%	1	2 (0.6%)
		All	52 (14.7%)	130 (36.7%)		45 (12.7%)	354
10	Pre-Test	A	10	51	12.9%	10	160 (82.1%)
		B	12	3	10.0%	9	11 (5.6%)
		C	19	11	20.0%	13	24 (12.31%)
		D	0	0	0%	0	0
		F	0	0	0%	0	0
		All	41 (21.0%)	65 (33.3%)		32 (16.4%)	195
10	Post-Test	A	10	53	12.5%	10	176 (91.2%)
		B	6	6	10.0%	5	8 (4.2%)
		C	6	3	40.0%	5	6 (3.1%)
		D	4	0	20.0%	2	2 (1.0%)
		F	2	0	0.0%	1	1 (0.5%)
		All	28 (14.5%)	62 (32.1%)		23 (11.9%)	193

Lower proportion of guesses in A's

Overall: 8% in A's vs. 48-61% in others

Total: 13.8% guesses

RQ3: Analyzing Guessing Behavior Using Attempt Sequences

- How much guessing is present?

Multiple guesses in one attempt

Module	Test	Grade	Guesses	Improvements	Regression	Trial-and-Errors	Totals
3	Pre-Test	A	34	88	13.8%	31	318 (89.1%)
		B	20	13	10.0%	19	33 (9.2%)
		C	0	1	0%	0	2 (0.6%)
		D	2	1	30.0%	2	2 (0.6%)
		F	1	3	0%	1	2 (0.6%)
		All	57 (16.0%)	106 (29.7%)		53 (14.8%)	357
3	Post-Test	A	22	112	14.3%	20	307 (86.7%)
		B	16	9	14.5%	13	25 (7.1%)
		C	11	7	20.0%	10	16 (4.5%)
		D	2	1	10.0%	1	4 (1.1%)
		F	1	1	10.0%	1	2 (0.6%)
		All	52 (14.7%)	130 (36.7%)		45 (12.7%)	354
10	Pre-Test	A	10	51	12.9%	10	160 (82.1%)
		B	12	3	10.0%	9	11 (5.6%)
		C	19	11	20.0%	13	24 (12.31%)
		D	0	0	0%	0	0
		F	0	0	0%	0	0
		All	41 (21.0%)	65 (33.3%)		32 (16.4%)	195
10	Post-Test	A	10	53	12.5%	10	176 (91.2%)
		B	6	6	10.0%	5	8 (4.2%)
		C	6	3	40.0%	5	6 (3.1%)
		D	4	0	20.0%	2	2 (1.0%)
		F	2	0	0.0%	1	1 (0.5%)
		All	28 (14.5%)	62 (32.1%)		23 (11.9%)	193

Learning Indicators

Statistically significant improvement on overall course grade over previous years

- Pre-test/post-test learning gains in certain modules
- Repeated attempts to get full marks

Learning Indicators

- ✓ Statistically significant improvement on overall course grade over previous years
  - Pre-test/post-test learning gains in certain modules
  - Repeated attempts to get full marks
- ✓ Exploratory learning behavior observed in use of subsequent attempts

Learning Indicators

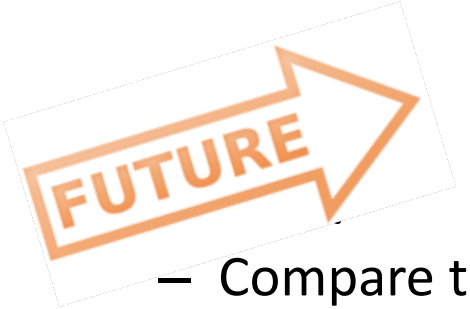
- ✓ Statistically significant improvement on overall course grade over previous years
  - Pre-test/post-test learning gains in certain modules
  - Repeated attempts to get full marks
- ✓ Exploratory learning behavior observed in use of subsequent attempts
- ? Not exhibiting rapid guessing behavior, but students are not taking full advantage of the 3-week window

Learning Indicators

- ✓ Statistically significant improvement on overall course grade over previous years
  - Pre-test/post-test learning gains in certain modules
  - Repeated attempts to get full marks
- ✓ Exploratory learning behavior observed in use of subsequent attempts
- ? Not exhibiting rapid guessing behavior, but students are not taking full advantage of the 3-week window
- ✓ Possible to develop dynamic model to detect guessing behavior
  - When combined with performance prediction, model can also detect wheel spinning behavior and offer adaptive help
  - ? However, false positives and false negatives can occur in the model
    Same limitation in the literature

Conclusions & Future Work

- Summary:
  - Analyzed learning behavior in test-taking context with multiple attempts (max = 3)
  - Proposed threshold model to quantify guessing behavior



  - Compare to other models of guessing behavior
  - Analyze idiosyncratic behavior (per person, per test, per concept)
  - Improve module and assessment design
  - Extend to contexts with other test types and assignment work