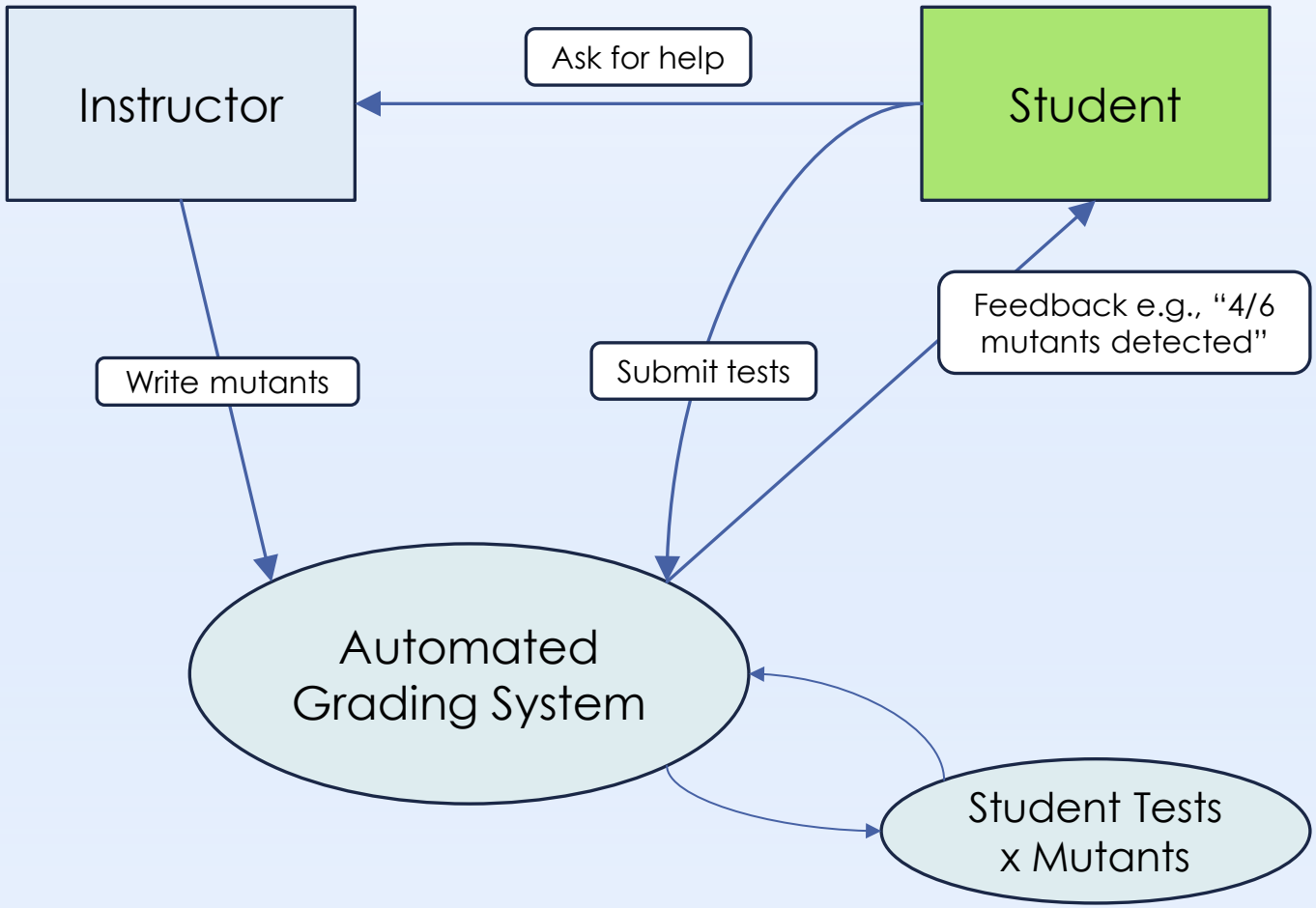
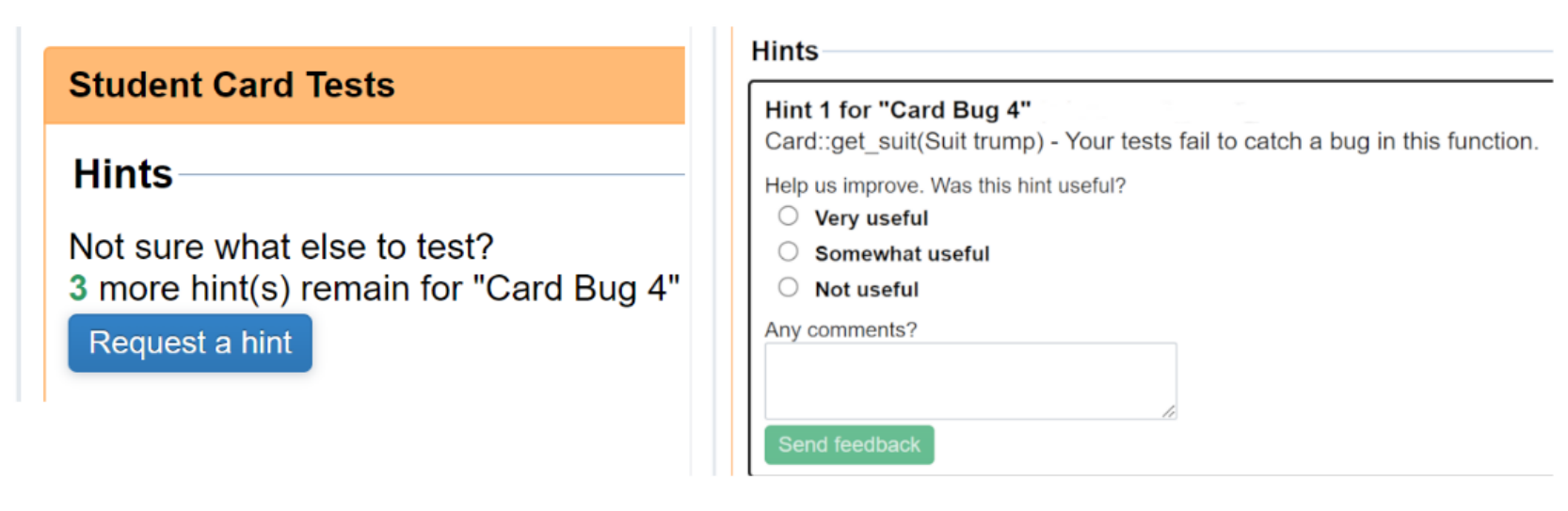
Instructor-Written Hints as Automated Test Suite Quality Feedback

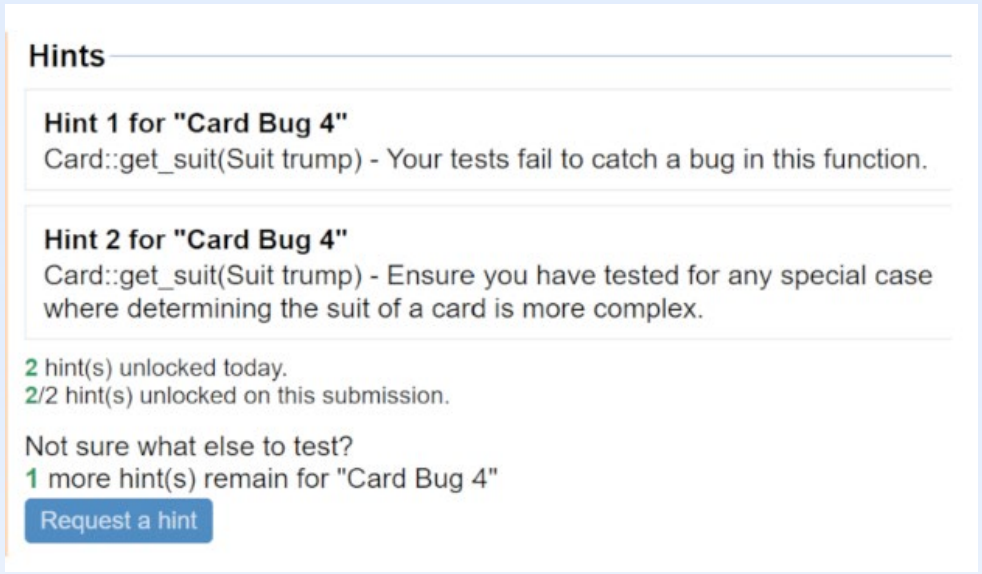
James Perretta (Northeastern University)
Andrew DeOrio (University of Michigan)
Arjun Guha (Northeastern University)
Jonathan Bell (Northeastern University)

Automated Test Quality Feedback



Hint System (Autograder.io)





Example of Hints

- Mutants hand-written by instructors

```
List<T>& List<T>::operator= (const List &l) {
  if (this == &l) return *this;
  removeAll();
  copyAll(l);
  return *this;
}
```

- **Hint 1:** Your tests fail to catch a bug in the Assignment Operator overload. Double check that you have tests for the assignment operator specifically (etc.)
- **Hint 2:** Ensure you test the assignment operator on a variety of lists, including any relevant special cases.
- **Hint 3:** Double check that you have tests for self-assignment, which is a special case. The list should contain the same values after self-assignment.

Study Overview

- **RQ1:** Does access to an automated hint system increase student test suite quality?
  - Controlled experiment: Record **# of mutants detected** with & without hints
- **RQ2:** What is the relationship between hints and student test suite revision?
  - Controlled experiment: Record **# of revisions** to completion
  - Quantitative analysis of "hint outcomes"
- **RQ3:** What kinds of hints do students perceive as helpful?
  - Mixed-methods analysis of hint ratings & comments

Assignment Summary

Assignment	# Students		Hint Access?	
	Fall '23	Spring '24	Control (Fall '23)	Experiment (Spring '24)
CS1 P1	1040	634	No	Yes
CS2 P1	777	746	No	Yes
CS2 P2	777	746	No	Yes
CS2 P3		756		
CS2 P4	762	698	No	No
PL P1		72		
PL P2		67		
SE P1		253		

RQ1: Controlled Experiment

	Number of mutants detected							
	CS1 P1		CS2 P1		CS2 P2		CS2 P4	
	Ctrl	Expr*	Ctrl	Expr*	Ctrl	Expr*	Ctrl	Expr
N	1015.00	607.00	735.00	693.00	714.00	667.00	721.00	665.00
Mean	7.30	7.23	12.64	13.00	11.67	12.05	13.59	14.19
Stdev	1.27	1.42	3.66	3.42	6.50	6.52	4.36	3.67
Min	0.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
Q1	7.00	7.00	12.00	13.00	7.00	7.00	12.00	13.00
Median	8.00	8.00	14.00	14.00	14.00	14.00	15.00	15.00
Q3	8.00	8.00	15.00	15.00	17.00	17.00	16.00	16.00
Max	8.00	8.00	16.00	16.00	20.00	20.00	19.00	19.00
U	307039.5		239414.5		227482.0		223368.0	
p-value	0.900		0.045		0.149		0.026	
r	0.003		0.060		0.045		0.068	

RQ2: Controlled Experiment

	Number of revisions							
	CS1 P1		CS2 P1		CS2 P2		CS2 P4	
	Ctrl	Expr*	Ctrl	Expr*	Ctrl	Expr*	Ctrl	Expr
N	1015.00	607.00	735.00	693.00	714.00	667.00	721.00	665.00
Mean	2.76	2.98	2.76	3.21	3.63	3.86	3.61	3.93
Stdev	1.76	1.93	1.86	2.09	2.32	2.47	2.56	2.65
Min	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
Q1	1.00	2.00	1.00	2.00	2.00	2.00	2.00	2.00
Median	2.00	3.00	2.00	3.00	3.00	4.00	3.00	3.00
Q3	4.00	4.00	4.00	4.00	5.00	5.00	5.00	5.00
Max	12.00	13.00	13.00	14.00	14.00	23.00	21.00	21.00
U	289282.0		222460.5		225205.5		220772.0	
p-value	0.035		<0.001		0.078		0.010	
r	0.061		0.127		0.054		0.079	

RQ2: Quantitative Analysis

Assignment	# Hints Shown	Mutant Detected					Hint Requested	Nothing
		Total	$NR = 1$	$NR = 2$	$2 < NR \leq 4$	$NR > 4$		
CS1 P1	535	47%	74%	20%	5%	1%	40%	14%
CS2 P1	1138	47%	85%	11%	4%	0%	36%	18%
CS2 P2	1897	31%	84%	12%	4%	1%	51%	18%
CS2 P3	4196	35%	77%	14%	7%	2%	54%	10%
PL P1	552	50%	65%	14%	13%	7%	35%	14%
PL P2	543	42%	57%	21%	13%	9%	45%	13%
SE P1	3149	37%	42%	24%	21%	12%	51%	13%
All		38%					49%	13%

- "Mutant Detected" (Mutant detected before requesting another hint)
  - Up to **85%** (CS2 P1) of "mutant detected" outcomes happened after a **single revision**
- "Hint requested" (Next hint requested before detecting mutant or not)
  - **65%** happened on the **same revision**
  - **25%** happened after **one revision**

"Nothing" = Mutant not detected, no more hints requested. Usually means student ran out of time

RQ3: Qualitative Analysis

- "Very Useful" Hints
  - CS2 P3 example:
    - Iterator::operator==() Bug #1 - Consider adding a test case that **compares a default-constructed iterator with iterators from a list**. The default-constructed iterator should not be equal to any of these iterators, not even an end() iterator.
  - PL P1 example:
    - This bug affects **error checking** in **quote**.
    - Student comment: "This is very useful, without it, I wouldn't think of do[ing] error checking for quote."
- Observations:
  - Hints that **reveal more** tend to have **higher ratings**
  - **Usefulness** can be **contextual**: PL P1 hint **addresses gap** in some students' testing approach

RQ3: Qualitative Analysis

- "Somewhat Useful" Hints:
  - PL P1 example:
    - This bug affects **division procedures**.
    - Student comment: "Shows a **general location to look**, so I would say it's sufficient."
  - SE P1 example:
    - This bug results in an incorrect turn order.
- Observations:
  - These hints give the **general location** of the mutation
  - Contextually useful if student just needs a nudge in the right direction

RQ3: Qualitative Analysis

- "Not Useful" Hints:
  - CS1 P1 example: Check that your test functions are called from within startTests
    - (This is the first hint for every mutant on this assignment)
  - PL P1 example: This bug affects **division procedures.**
    - Student comment: "I did test zero division but didn't catch bugs. This is just a random guessing game."
  - CS2 P3 example: Iterator::operator==() Bug #1 –
    - Consider adding a test case that **compares a default-constructed iterator with iterators from a list**. The default-constructed iterator should not be equal to any of these iterators, not even an end() iterator.
    - Student comment: "I have a test that does exactly this."
- Observations:
  - Hints with **redundant information** are less likely to be useful
  - If students think they've already tested a behavior, they **may need additional guidance** as to why their tests are insufficient or incorrect

Takeaways

- Automated hints can help **equalize access** to instructor feedback
  - All students have access to the same hints
  - Students with access to hints submitted more revisions
- "Too much information" **may be what students need**
  - Students in the experiment group (who had access to the most revealing hints) still detected **more mutants on average, even after access to hints was removed**
- Hints can help students make **more productive revisions**
  - Limits on hints/submission & submissions/day may help **encourage reflection**, discourages spamming

Thank You!