Peer and Self Assessment in Massive Online Classes

Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, Scott R. Klemmer

Stanford University, Coursera Inc., and UC San Diego

Peer assessment

Classmates assess each other
Provide summative evaluation, and constructive criticism

Peer assessment used in 100+ classes

Human-computer Interaction

Design

Programming in Python

Code

Introduction to Philosophy

Essays

Teaching character

Management

Child Nutrition

Recipes

Social Psychology

Essays

Constitutional law

Arguments

World Music

Music

Our peer assessment process

1) Practice 2) Assess 5 Peers 3) Self-Assess

Provides students grades and improvement-oriented feedback
59% of student submissions get grades within 10% of staff grades

In this talk

- *Why peer assessment is necessary*
- *How peers assess open-ended work*
  Peer assessment compares well with staff grade (r=0.73)
- *How peers improve open-ended work*
  Qualitative feedback + opportunities to reflect
- *How data improves peer assessment*

Open-ended assignments are pedagogically valuable

- Closed ended questions
  - constrain choices
  - Recognizing a correct solution does not mean you can generate it
- Open ended assignments
  - Require students to generate solutions, not only recognize them
  - Can assess on more realistic tasks
  - Embrace multiple solutions

Veloski et al 1999
Thompson et al 2000

Challenges of Open-ended assessment

- Realistic, open-ended assignments require lots of grading time
  - staff grading takes prohibitive labor (400+ hours/ week)
  - Machine grading reliant on lexical and syntactical features not robust enough
- Though sharing work provides inspiration and encourages discussion, students don't see others' work

The paradox of peer processes

Non-experts performing expert work

In-person classes

- Peer grades correlate well with staff
- Peers can provide constructive criticism

Does this scale to global online classes?

Falchikov and Goldfinch (2000)
Sluijsmans et al. (2002)
Kulkarni&Klemmer (2012)
Tinapple et al (2013)

Our peer assessment process

1) Practice 2) Assess 5 Peers 3) Self-Assess

Similar to CPR but calibrate students, not the algorithm.

Final peer grade is a simple median, not a calibrated, weighted mean.

Carlson & Berry(2003)

coursera COURSES UNIVERSITIES ABOUT ▼ SCOTT R KLEMMER ▼

STANFORD UNIVERSITY

Human-Computer Interaction

Scott Klemmer, Associate Professor

Helping you build human-centered design skills, so that you have the principles and methods to create excellent interfaces with any technology.

Sign Up Preview

Started on: Sep 24th 2012 (9 weeks long)
Workload: 8-10 hours/week
Information, Technology, and Design
Computer Science: Programming & Software Engineering

3,709 Tweet 5.9k +1 235k Like

About the Course

In this course, you will learn how to design technologies that bring people joy, rather than frustration. You'll learn several techniques for rapidly prototyping and evaluating multiple interface alternatives -- and why rapid prototyping and comparative evaluation are essential to excellent interaction design. You'll learn how to conduct fieldwork with people to help you get design

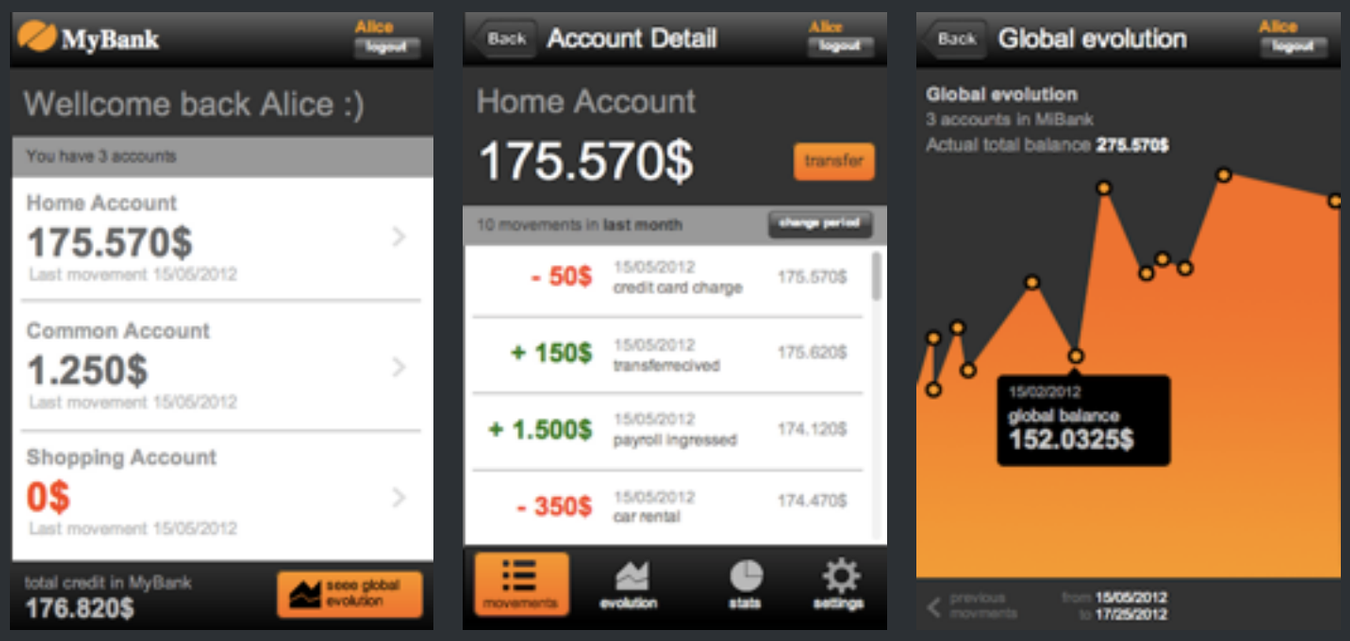
Free; anyone can enroll. Open-ended assignments central to class, 6-7 hours/week
In all, 65711 students watched videos, 5,876 students submitted open-ended assignments

Assignments step through a human-centered design process

Needfinding and low-fi prototypes

Task	Deadline	Status	Comm
Create the project in Axure aplication	6/19/0212		
Set up the project online (using dropbox) to access everyone	6/19/0212		
Create a demo page in the project, whith some links and upload it to check the online version is ok	6/20/0212		it cost time th to uplo page
Test that the web can run in a mobile	6/20/2012		finally to impl functio mobile decide implem simula
Create the navigational flow and main structure of the app in axure	6/21/2012		
Code up screen place holders for the app navigation and main functionalities (login, home, navigation between accounts, different views of an account, transfer functionality, retrieve pass)	6/22/2012		
Upload and test	6/22/2012		
Developing static main screens of the app I: List of accounts and general evolution	6/22/2012		
			change setting interac

Implementation plan



Functional prototypes

Movements screen

- Difficult to understand the "change period" button, no user know how to see more movements (3/3 users have this problem)

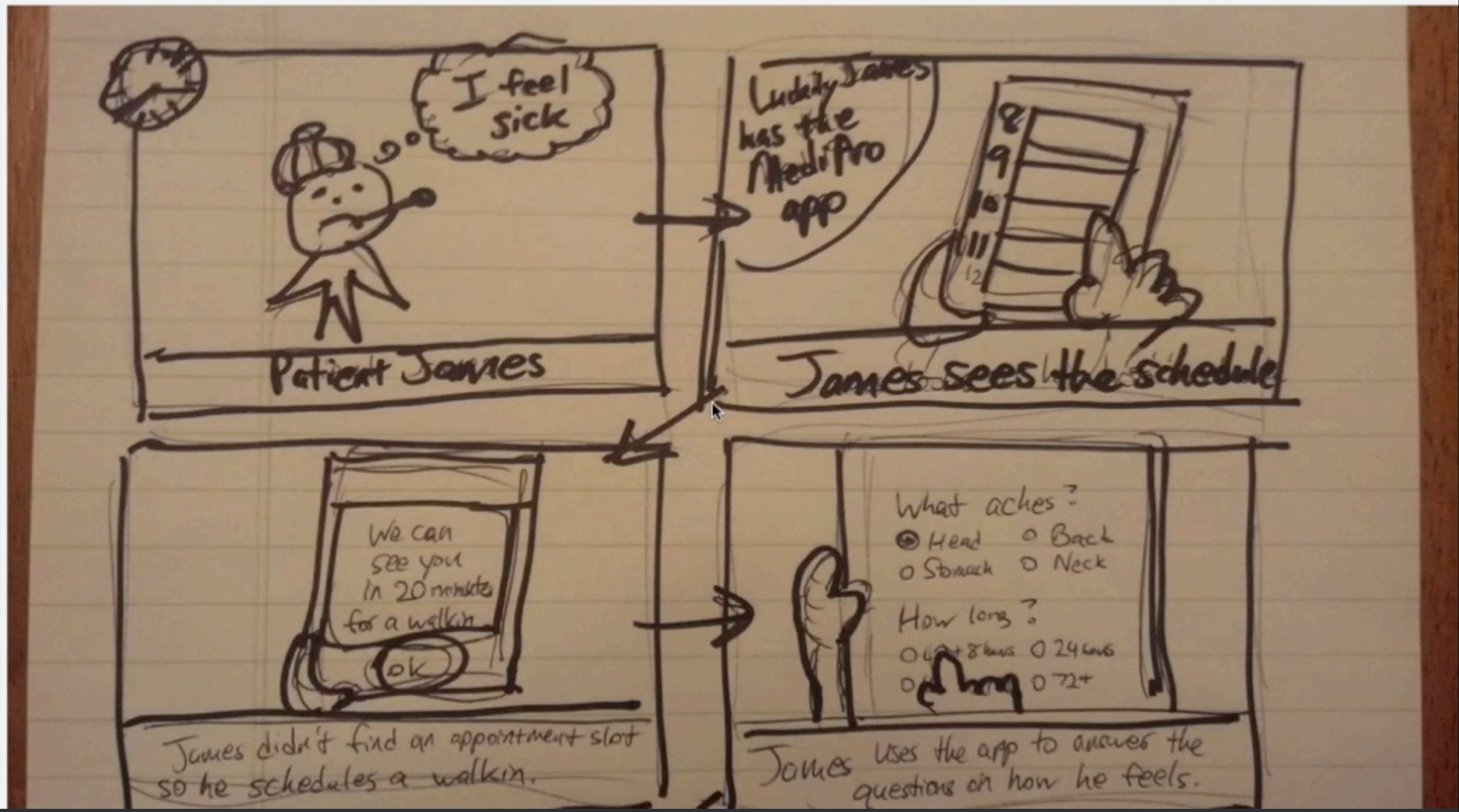
- It's no very clear at first glance if transfers button is to make a transfer from this account or to this account. (1/3 users have this problem)

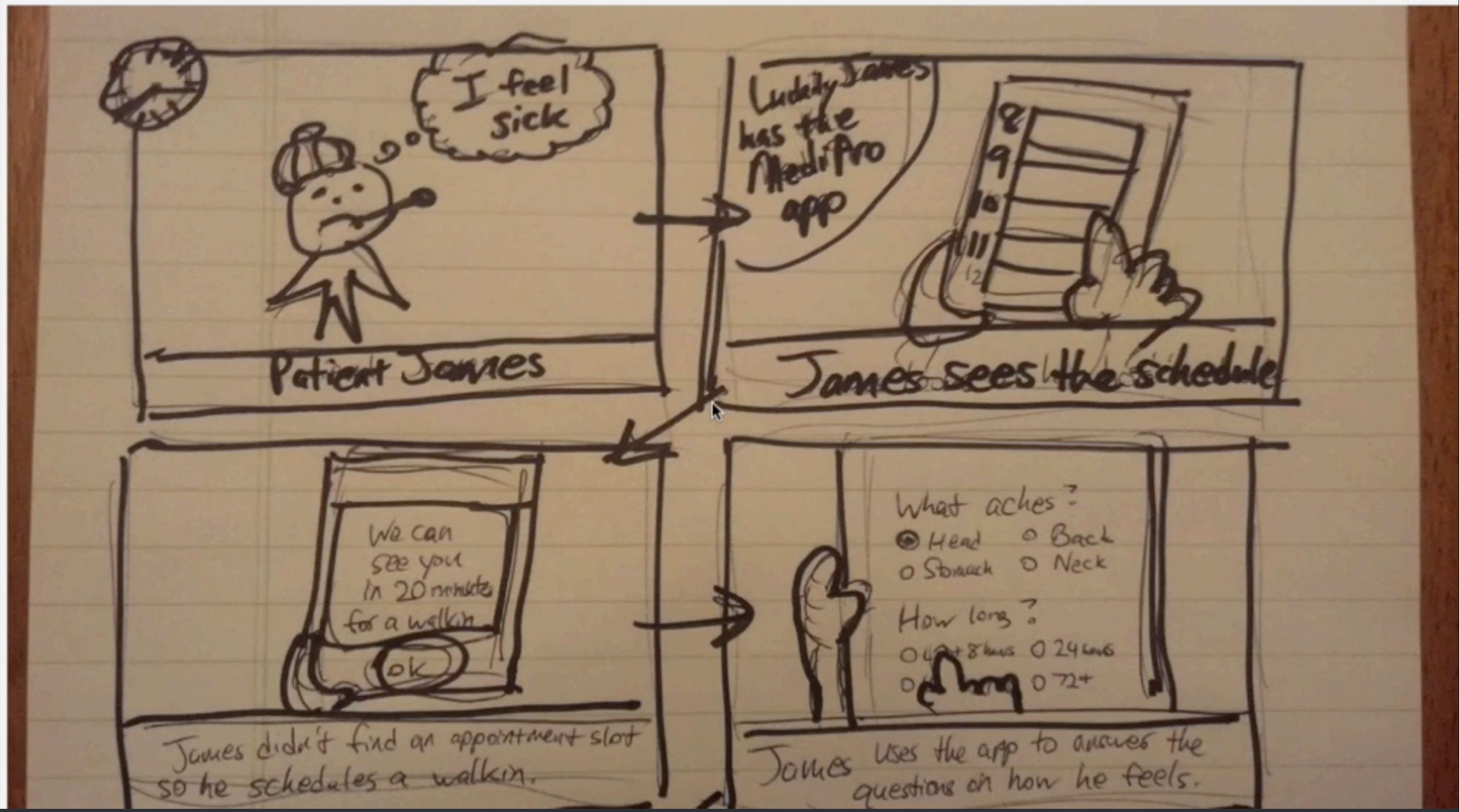
Menu

- It's difficult to users understand the difference between stats and evolution (2/3 users don't understand it)

- Users don't guess what the stats option contains before visit the screen (3/3 users don't know it)

User testing and iteration


I feel sick
Patient James
Luckily James has the MediPro app
James sees the schedule
We can see you in 20 minute for a walkin
ok
James didn't find an appointment slot so he schedules a walkin.
What aches?
Head
Back
Stomach
Neck
How long?
72+
James uses the app to answer the questions on how he feels.

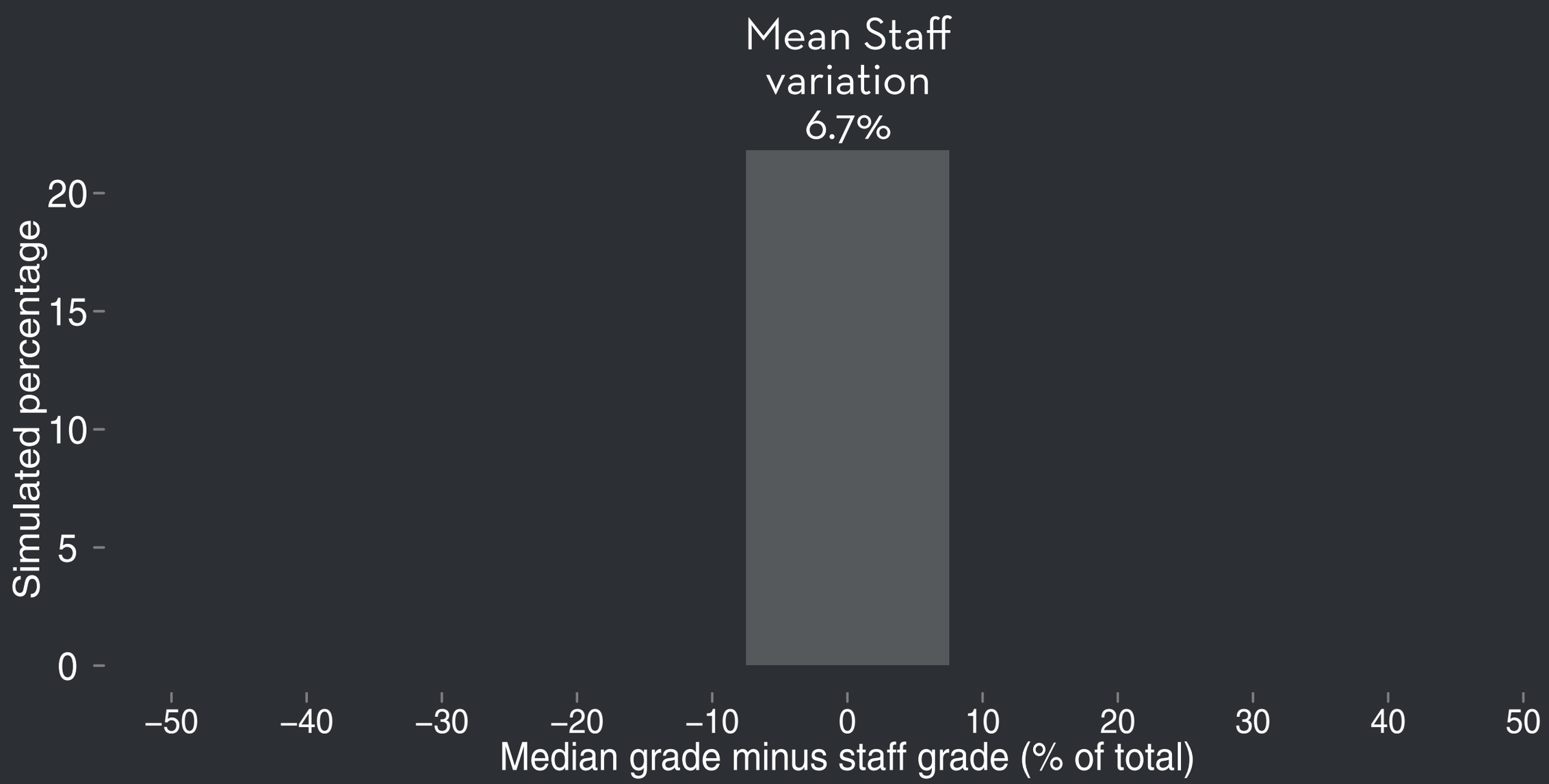

I feel sick
Patient James
Luckily James has the MediPro app
James sees the schedule
We can see you in 20 minute for a walkin
ok
James didn't find an appointment slot so he schedules a walkin.
What aches?
Head
Back
Stomach
Neck
How long?
72+
James uses the app to answer the questions on how he feels.

1) Practice

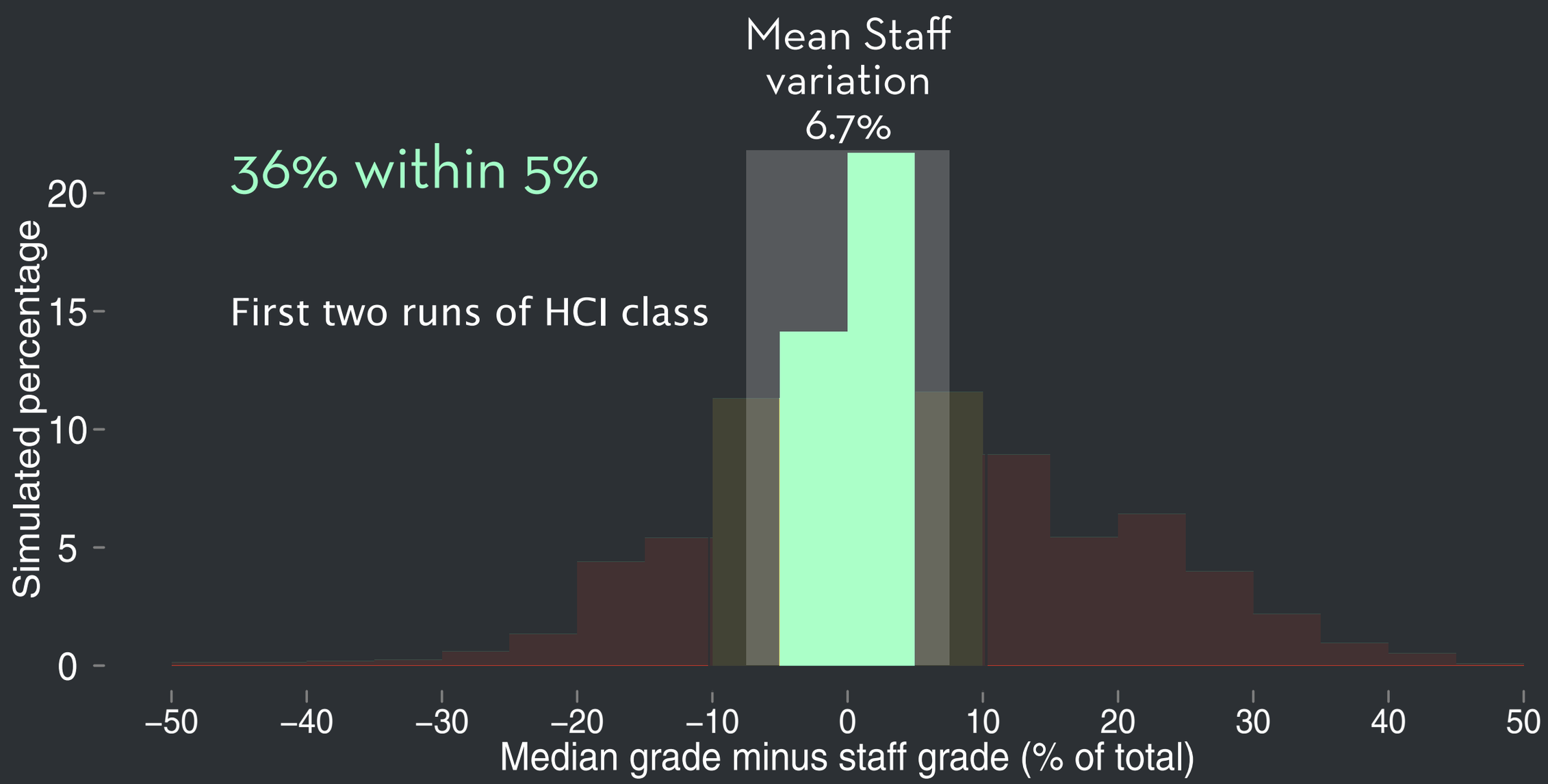
2) Assess 5 Peers

3) Self-Assess

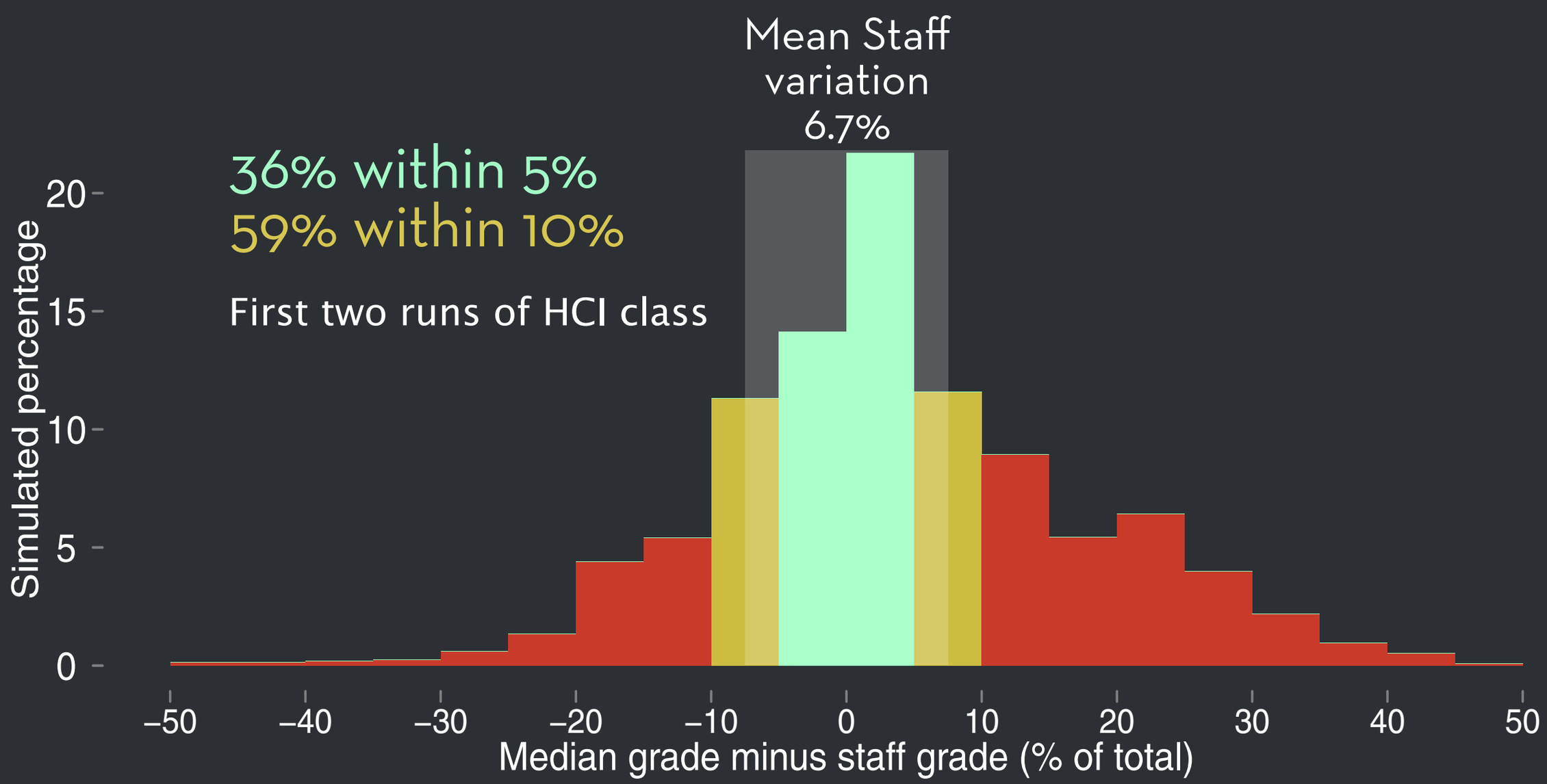
Median peer grades correlate with staff



Median peer grades correlate with staff



Median peer grades correlate with staff



Grading adequate for pass-fail class

- Errors are nearly symmetric around staff grades

Summing grades on all assignments in a simulation

Assessing others yields perspective and inspiration

*"... **seeing how others** tackled the assignments sometimes helped me point out things I missed and offered some perspective..."*

114 similar responses

*"... Giving me a wider **point of view** based on the others work..."*

36 similar responses

Consistent with in-person effects

Chinn (2005); Tinapple et al. (2013)

Assessing yourself *after* peers yields reflection and comparison

*"...Was nice to evaluate **my own work** AFTER evaluat[ing] others because I could compare my work and effort..."*

175 similar responses

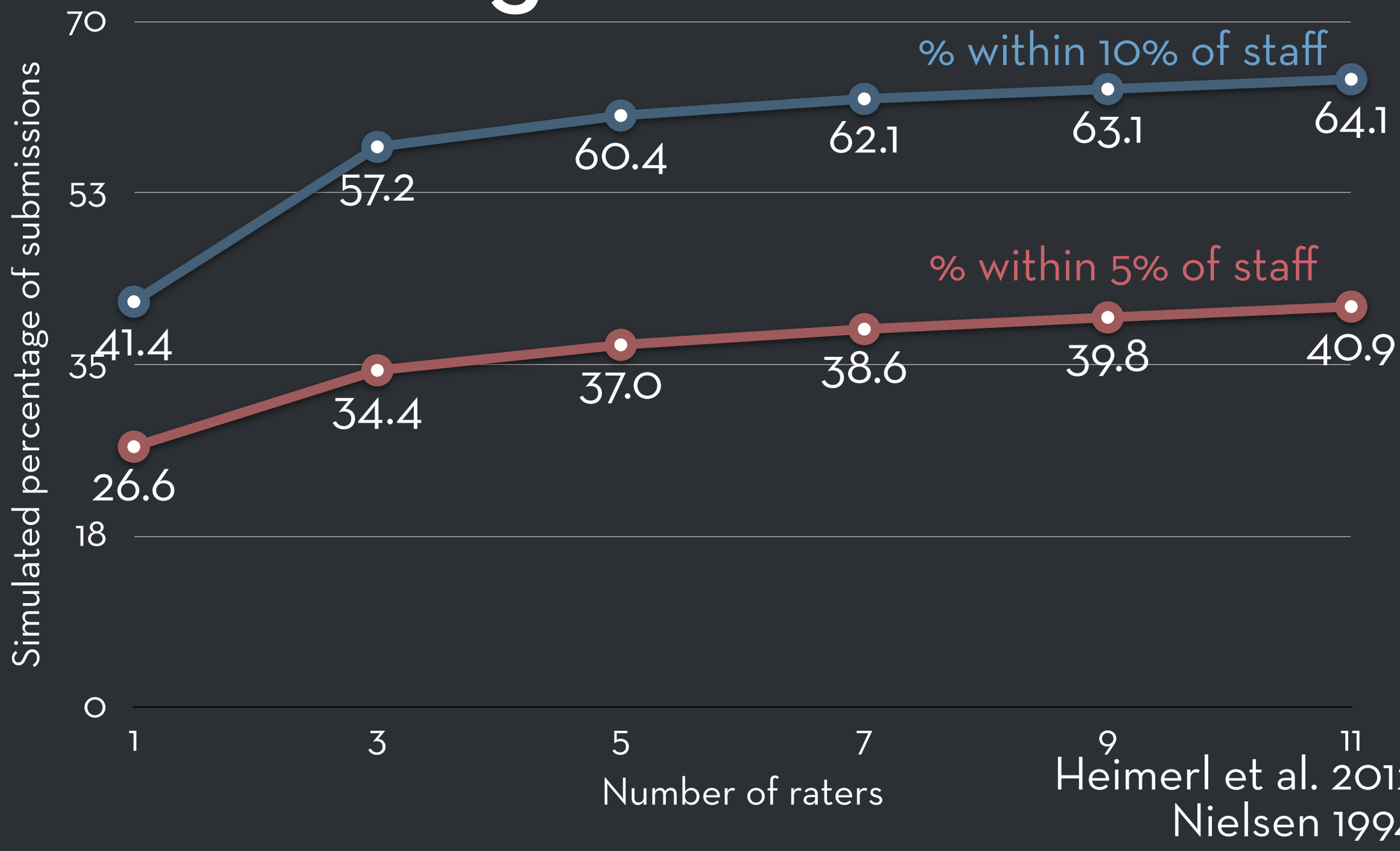
*"... helped me see problems that **i didn't see** before the peer evaluation..."*

50 similar responses

Consistent with in-person effects

Boud (1994)

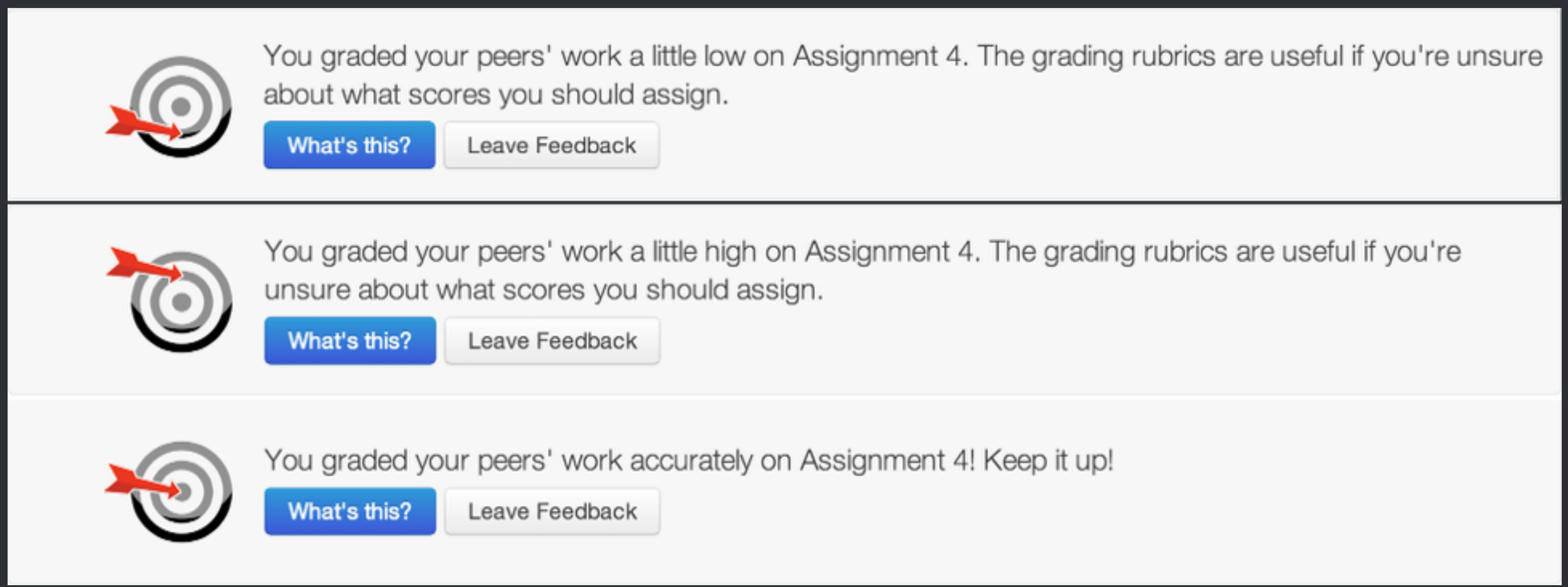
More peers yield quickly diminishing returns



Heimerl et al. 2012
Nielsen 1994

Feedback on grading improves accuracy

- n = 756
- Between subjects feedback v. no feedback



Grading Feedback reduced error

- 7.74% for control to 6.77% with feedback [t(4998)=3.38, p<0.01]
- Students appreciated receiving feedback
- Learning benefit: knowing how to appraise work right builds confidence
- Helps build shared norms

Qualitative feedback

Improvement-oriented feedback beyond the rubric

Some feedback minimal/ superficial

Minimal “Great idea!”

Superficial “I can't read the words in the pics clearly”

Better “Solution requirement is vague here but I'm excited to see where you take this in the storyboards!”

The return of the novices-as-experts paradox

Experts:
capture the structure of rubric

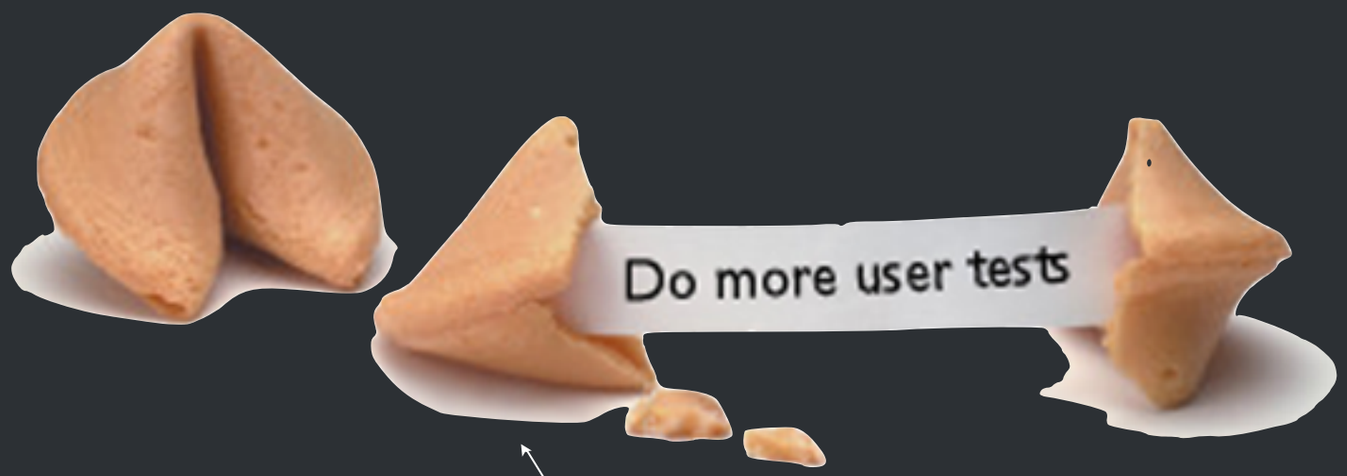
"fully interactive, page flow is complete... make it clearer what people should do next"

Peers:
Focus on superficial features, even when asked not to

"unpolished...Try to make UI less coloured."

Fortune cookies for qualitative, personalized feedback

- Peers can recognize errors from a list of patterns, even if they can't articulate them
- Most errors are variations on a theme



+ "...because _____"

Cue

Variation

Overall evaluation/feedback

Note: this section can only be filled out during the evaluation phase.

Overall feedback:

How could this student best improve his/her submission? From among the following, copy one or more pieces of advice that would help the student. Paste your advice in the feedback box below.

- Clarify the concerns, goals, and expectations of the user tests.
- Make the user tests more structured.
- Make the user tests more consistent across participants.
- Make the prototype more interactive so the user test represents a more real-life interaction.
- Determine the implications of the user succeeding (or not) on each task on the prototype.
- Make fewer assumptions about users/Reduce bias in user test.
- Other

Copy, then paste

Make the prototype more interactive so the user test represents a more real-life interaction: The prototype does everything you're testing, but it couldn't hurt to make it more interactive. If the user can't possibly stray from the things you want to test, how do you know that the user can actually use the full application without making mistakes?

Fortune cookies feedback actionable and detailed

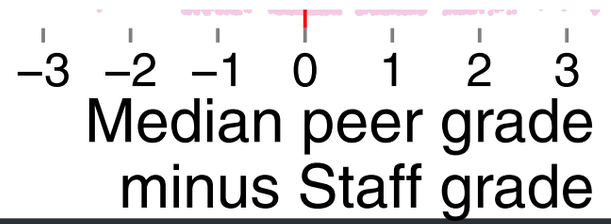
- “Clarify the concerns, goals, and expectations of the user tests: try to expand the information in the implementation plan”

- “Prototype should relate to the user needs in the storyboard more. I dont see the proposed functionallity from the storyboards here in the prototypes.”

Improving assessment

Using data on peer-staff disagreement

Low-fi prototyping	Point of view
	Prototype 1
	Prototype 2
	Storyboard 1
	Storyboard 2
Implementation Plan	Deadlines
	Heuristic evaluation
	Weekly plan
	Navigation skeleton
Functionality & Test Plan	Functionality
	Goals
	Test appropriateness
	Test completeness
	Alternate redesign
Test results & Iteration	Redesign complete
	Test changes
	Test photos
	Test process
	Test results



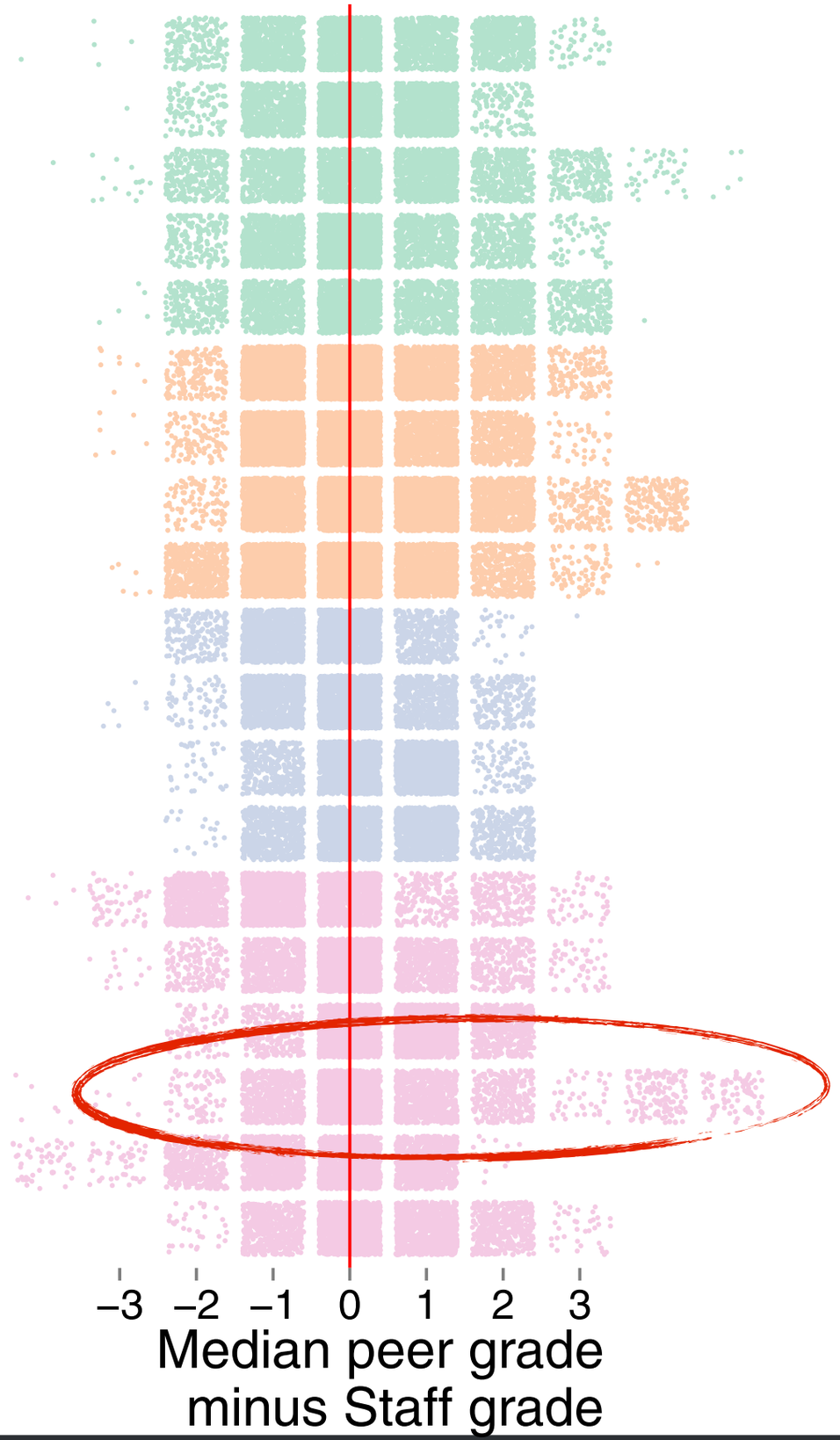
Low-fi prototyping

Implementation Plan

Functionality & Test Plan

Test results & Iteration

Point of view
Prototype 1
Prototype 2
Storyboard 1
Storyboard 2
Deadlines
Heuristic evaluation
Weekly plan
Navigation skeleton
Functionality
Goals
Test appropriateness
Test completeness
Alternate redesign
Redesign complete
Test changes
Test photos
Test process
Test results



Separating orthogonal questions improves agreement

Orthogonal attributes combined

Did the student upload interesting photos?

Orthogonal attributes separated

4% better agreement

1. Did the student upload photos?
2. Were photos interesting?

Parallellizing rubric cells improves agreement

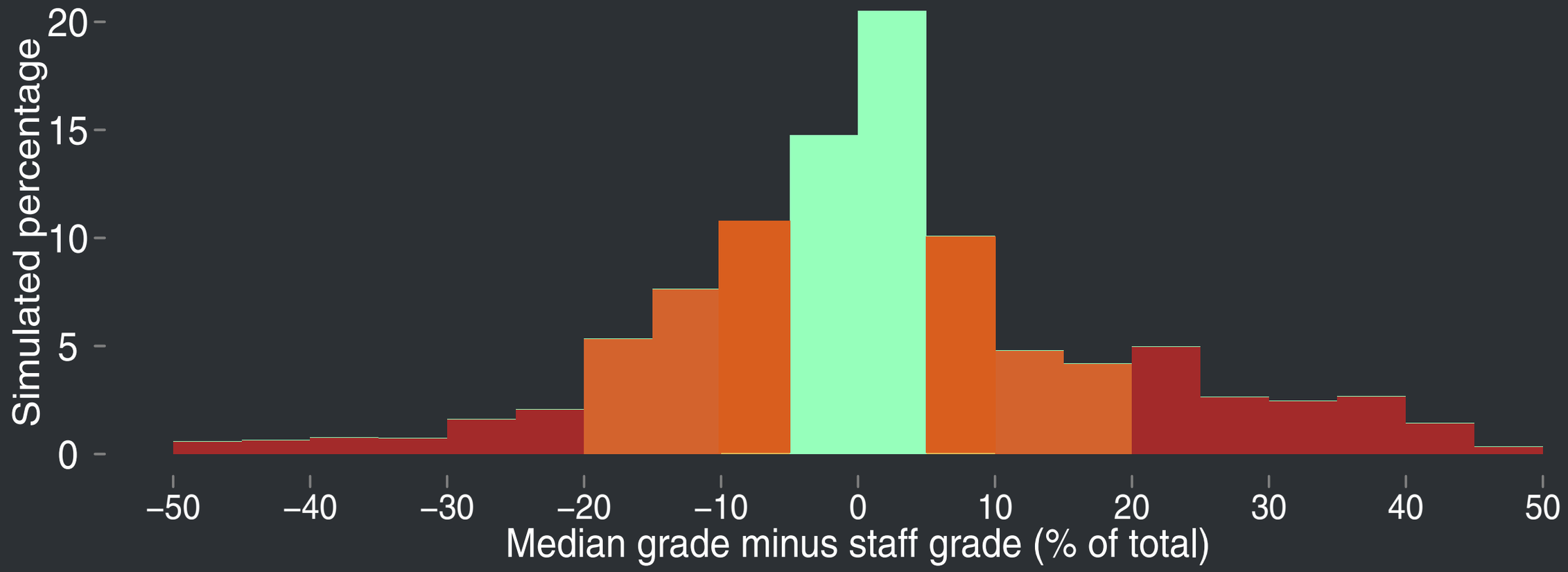
	1 point	3 points
Non-parallel structure	The storyboards are hard to follow or do not address the point of view.	The storyboards reasonably address the point of view, and are reasonably easy to understand

Parallellizing rubric cells improves agreement

	1 point	3 points
Parallel structure	The storyboards are hard to follow or do not address the point of view.	The storyboards are easy to follow and reasonably address the point of view

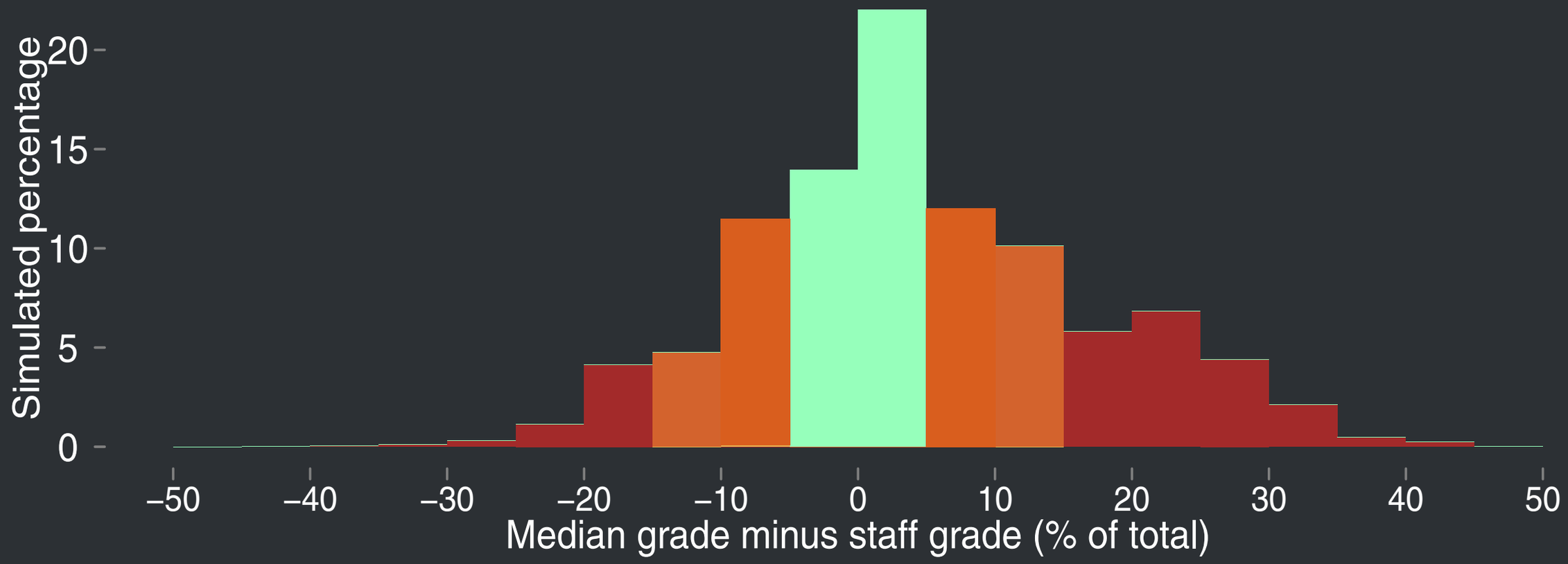
8% better agreement

Revising rubrics improves agreement



Peer-staff agreement increases for the median submission
within 5% improves 34% -> 42%
Peer-staff agreement outliers reduced
80% of submissions within 20% -> 15% of staff

Revising rubrics improves agreement



Peer-staff agreement increases for the median submission
within 5% improves 34% -> 42%
Peer-staff agreement outliers reduced
80% of submissions within 20% -> 15% of staff

Our vision

Peer processes can provide students deeper feedback, improve motivation and learning

(And do so at scale)

The 7 habits of highly successful peer assessment

- Assignment-specific rubrics
- Iterate before release (pre and during)
- Assignment-specific training
- Self assessment at the end
- Staff grades as ground truth
- Adaptive grade aggregation
- Provide feedback to graders

http://hci.st/assess

Machine grading?

Can algorithms reduce busywork and amplify peer processes?

Piech et al (2013)
Kulkarni et al (2014)

Coming up

L@S: Scaling Short-answer Grading by Combining Peer Assessment with Algorithmic Scoring

CHI workshop

Peer and Self Assessment in Massive Online Classes

Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, Scott R. Klemmer

http://hci.st/assess

Back up, back up!

References

- Guzdial, M., and Turns, J. Effective discussion through a computer-mediated anchored forum. The Journal of the Learning Sciences 9, 4 (2000), 437–469.
- Blitzer, D. The wide world of computer-based education. Advances in computers 15 (1976), 239–283.
- CommentSpace: Structured Support for Collaborative Visual Analysis Wesley Willett, Jeffrey Heer, Joseph Hellerstein, Maneesh Agrawala, ACM Human Factors in Computing Systems (CHI), 2011
- Motivating Participation by Displaying the Value of Contribution. Al Mamunur Rashid, Kimberly Ling, Regina D Tassone, Paul Resnick, Robert Kraut, John Riedl, CHI 2006
- Group Processes in the Classroom, Second Edition.Schmuck, Richard A.; Schmuck, Patricia A.
- Apprenticeship in thinking: Cognitive development in social context. Rogoff, Barbara, Oxford University Press, 1990
- Cooperation in the Classroom., David W. Johnson, Roger T. Johnson, Edythe Johnson Holubec, Interaction Book Company, 1991
- Managing Cultural Diversity: Implications for Organizational Competitiveness, Taylor H. Cox and Stacy Blake, The Executive, Vol. 5, No. 3 (Aug., 1991),
- Faris, O. (2009) The Impact of Homogeneous vs. Heterogeneous Collaborative Learning Groups in Multicultural Classes on the Achievement and Attitudes of Nine Graders towardsLearning Science. Online Submission, Feb. 2009.
- Ost, Ben (2010) The role of peers and grades in determining major persistence in the sciences, Economics of Education Review
- Veloski et al (1999) Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence

Coming up

L@S: Scaling Short-answer Grading by Combining Peer Assessment with Algorithmic Scoring

CHI workshop

Peer and Self Assessment in Massive Online Classes

Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, Scott R. Klemmer

Stanford University, Coursera Inc., and UC San Diego