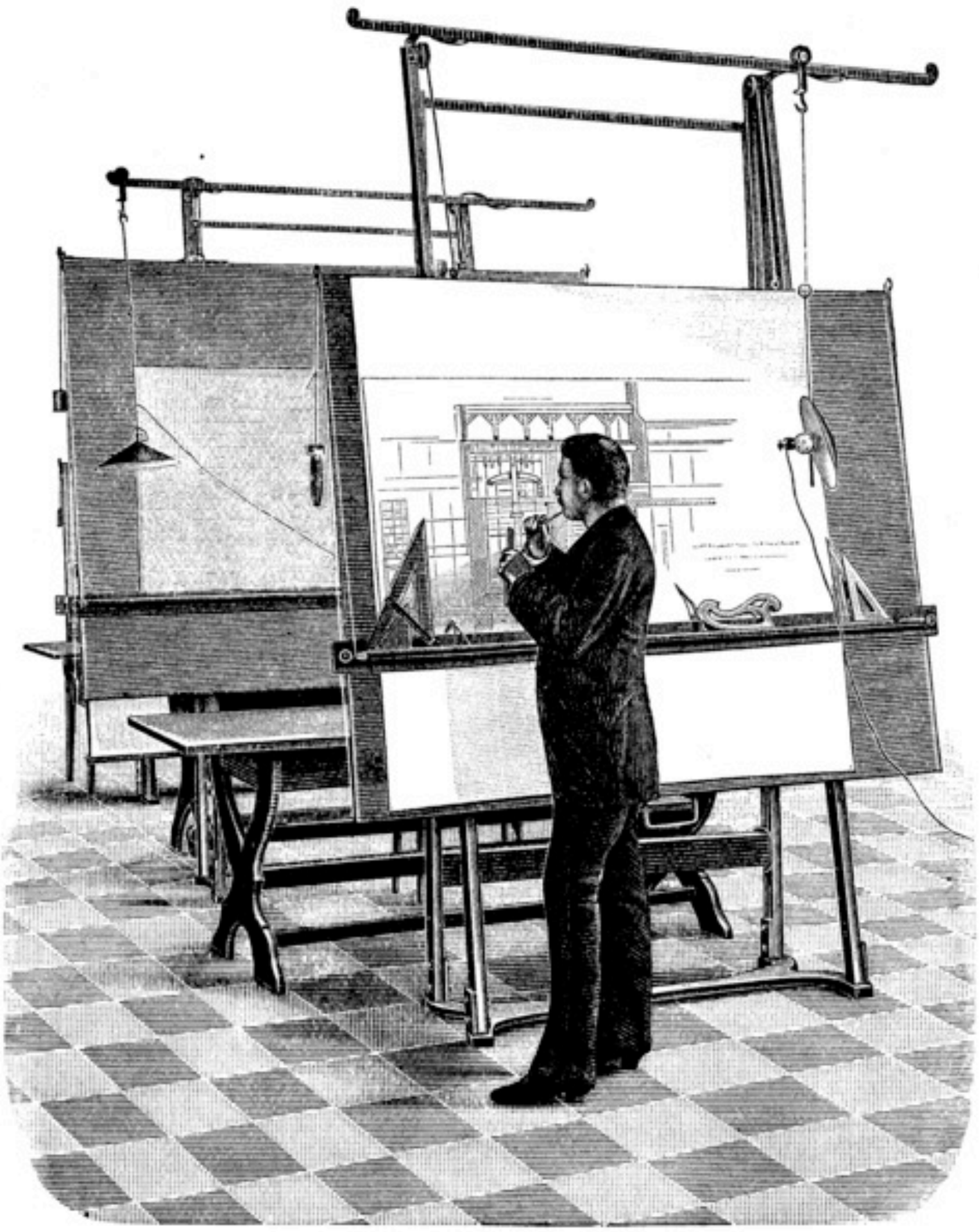
SCALING STUDIO CRITIQUE

Scott Klemmer & Chinmay Kulkarni

with Koh Pang Wei, Huy Le, Daniel Chia, Nisha Masharani, Kathryn Papadopoulos, Daphne Koller

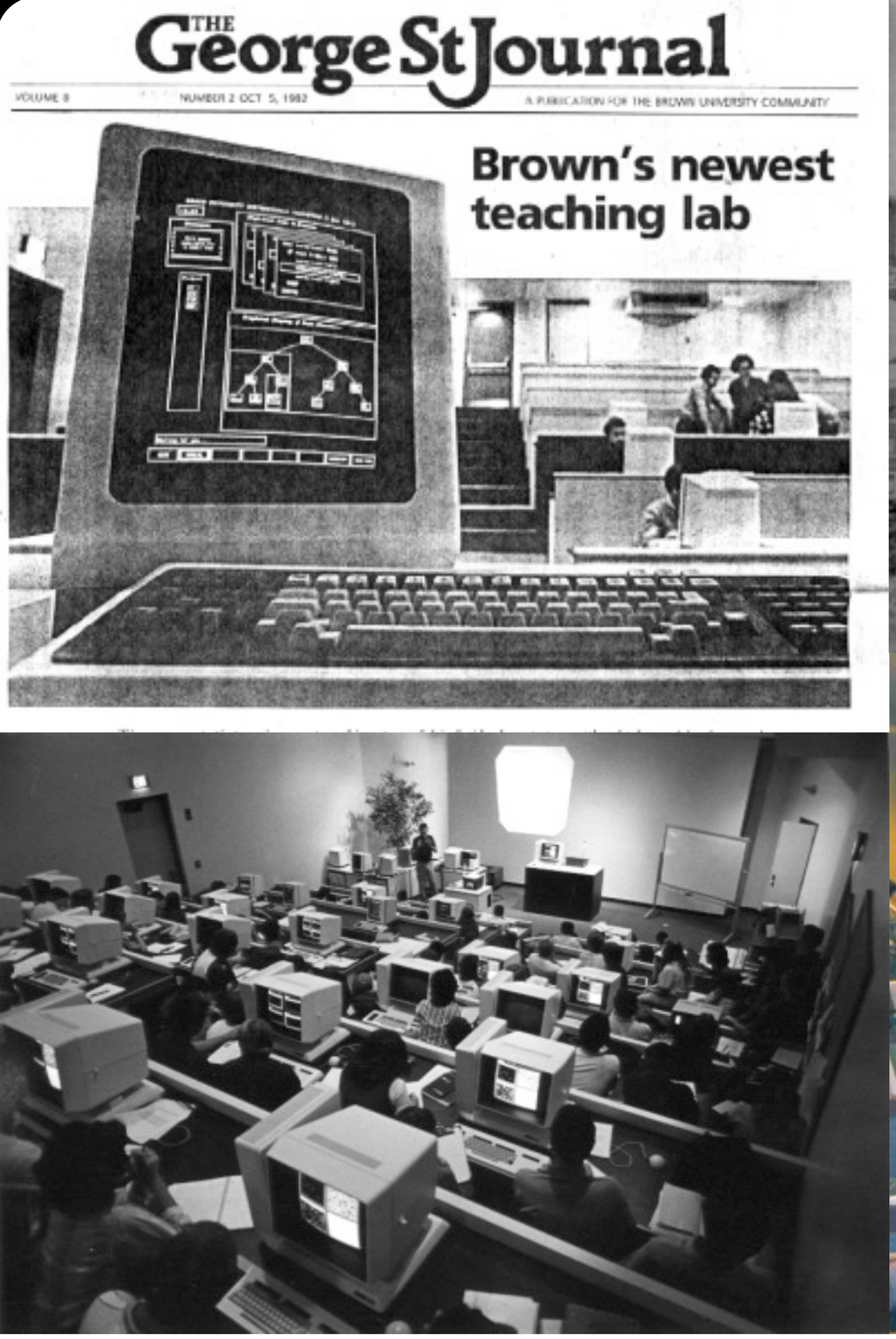
As many of you know, I have a background in design, and I'm a huge fan of it's teaching methods.

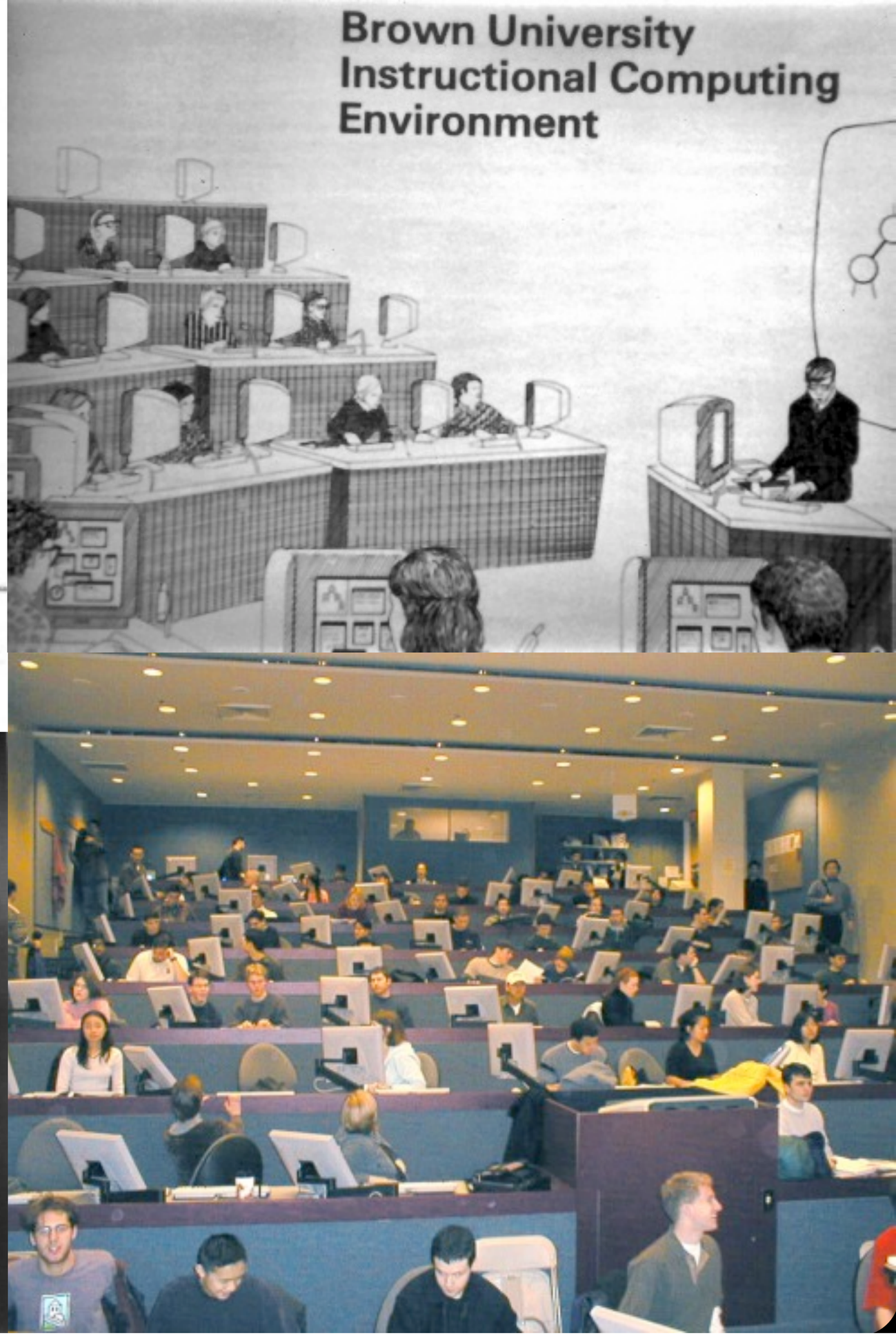


This “technology” was introduced with the founding of the École des beaux-arts in Paris in 1819, and has endured for nearly 200 years.

Got refined at the Bauhaus in the early 20th century, and with the emigration of many of the faculty and alumni to the US, became the foundation for design education in the US.

I think that art and design education can teach us a lot. And the first lesson has to do with physical space. Barbie dolls, umbrellas, new ideas, old ideas, good ideas, and bad. Collocated, cluttered studios are hallmarks of art and design education.





When took computer science courses as an undergraduate at Brown, we actually had very much of a studio work environment. All of the programming assignments were done on Unix workstations. And all the Unix workstations -- for undergraduates -- were in one large lab, shown here over several different generations of technology. Last week I was at UIUC. I spoke with this about Karrie Karahalios, an MIT alum, and she had experienced both the before and after, saying that “When they got rid of the athena cluster, it really affected the social interaction, sense of cohort, and learning.”

PROBLEM-FINDING

as well as

PROBLEM SOLVING

My experiments with self-assessment came about for several reasons. One of them is that, the first time I taught a design course of Stanford, in the fall of 2005. 13%.

RoutledgeFalmer
Taylor & Francis Group

ENHANCING LEARNING through Self Assessment

David Boud

Category	Unsatisfactory	Bare minimum	Satisfactory effort & performance	Above & Beyond
Point of view	0: No point of view or irrelevant point of view.	1: The student's point of view does not express a problem or opportunity, does not create requirements for a solution, or does not relate to the design brief.	3: The point of view relates to the brief and the problem and solution are clearly stated, but the solution requirement is too general (any solution meets the requirement) or is too specific (allows for only one solution).	5: The problem and solution requirement are clearly stated in the point of view; the requirement allows for focus without being too constraining.
Storyboard #1	0: The student did not submit a storyboard.	3: The storyboard is hard to follow or does not address the problems and solution requirements identified in the point of view. .	5: The storyboard reasonably addresses the problems identified in the point of view, but a reader may have lingering questions about the situation depicted.	7: The storyboard is easy to follow. Someone else could come up with a distinct prototype that would correspond with the point of view just from looking at the storyboard.
Storyboard #2	0: The student did not submit a second storyboard.	3: The storyboard is hard to follow and does not address the problems and solution requirements identified in the point of view.	5: The storyboard reasonably addresses the problems identified in the point of view, but a reader may have lingering questions about the situation depicted and/or it does not diverge and represent a different solution from the first storyboard.	7: The storyboard is easy to follow. Someone else could come up with a distinct prototype that would correspond with the point of view just from looking at the storyboard. It clearly diverges and represents a different solution from the first storyboard.
"Out of the box" Storyboards	0: The student submitted average, good, or reasonable storyboards.			1: Either one or both storyboards submitted were incredibly creative and innovative. Only the top 5% of the class will have storyboards that fit into this category.
Prototype #1	0: No prototype or completely irrelevant prototype.	3: The prototype is not interactive or interactions are broken, and there is no defined purpose for many elements in the prototype.	5: The prototype is mostly complete, although there are some functions that are not yet interactive, some elements have no defined purpose, and it can be difficult to know how to use certain parts of the prototype.	7: The prototype has enough detail that the user can see all of the interactions, understand how they work, and a programmer could use the prototype to create a functional application with a defined flow.
Prototype #2	0: No second prototype or completely irrelevant prototype.	3: The prototype is not interactive or interactions are broken, and there is no defined purpose for many elements in the prototype.	5: The prototype is mostly complete, although there are some functions that are not yet interactive, some elements have no defined purpose, and it can be difficult to know how to use certain parts of the prototype, and/or it does not diverges from the first prototype and explore different interfaces implementing the same idea.	7: The prototype has enough detail that the user can see all of the interactions, understand how they work, and a programmer could use the prototype to create a functional application with a defined flow. The prototype clearly diverges from the first prototype and explores different interfaces implementing the same idea.

Rubric, http://cs147.stanford.edu

Use a rubric to grade. When we grade in our teaching staff meeting, we often also link submissions that excel in a particular area to that cell of the rubric. It would probably be better to also link mediocre examples. We've done a little bit of that in some years, but it's a delicate matter.

cs147: Weekly Ritual

- Assignments due each Friday 8am
- Participate in studio
- Reflect on & self-assess your work

The in-person piece is that every Friday students gather for 12-15 person studios, led by a TA. **In studio**, students provide critique to each other. **A key piece** of this is that you do the self-assessment immediately after seeing each other's work. **So while there's not** an explicit curve to the class, there is a norming effect by seeing what others did. **In this class**, there's lots of peer critique, but no peer assessment. Peer assessment can work well, and it's something you may choose to explore.

Submit, Share

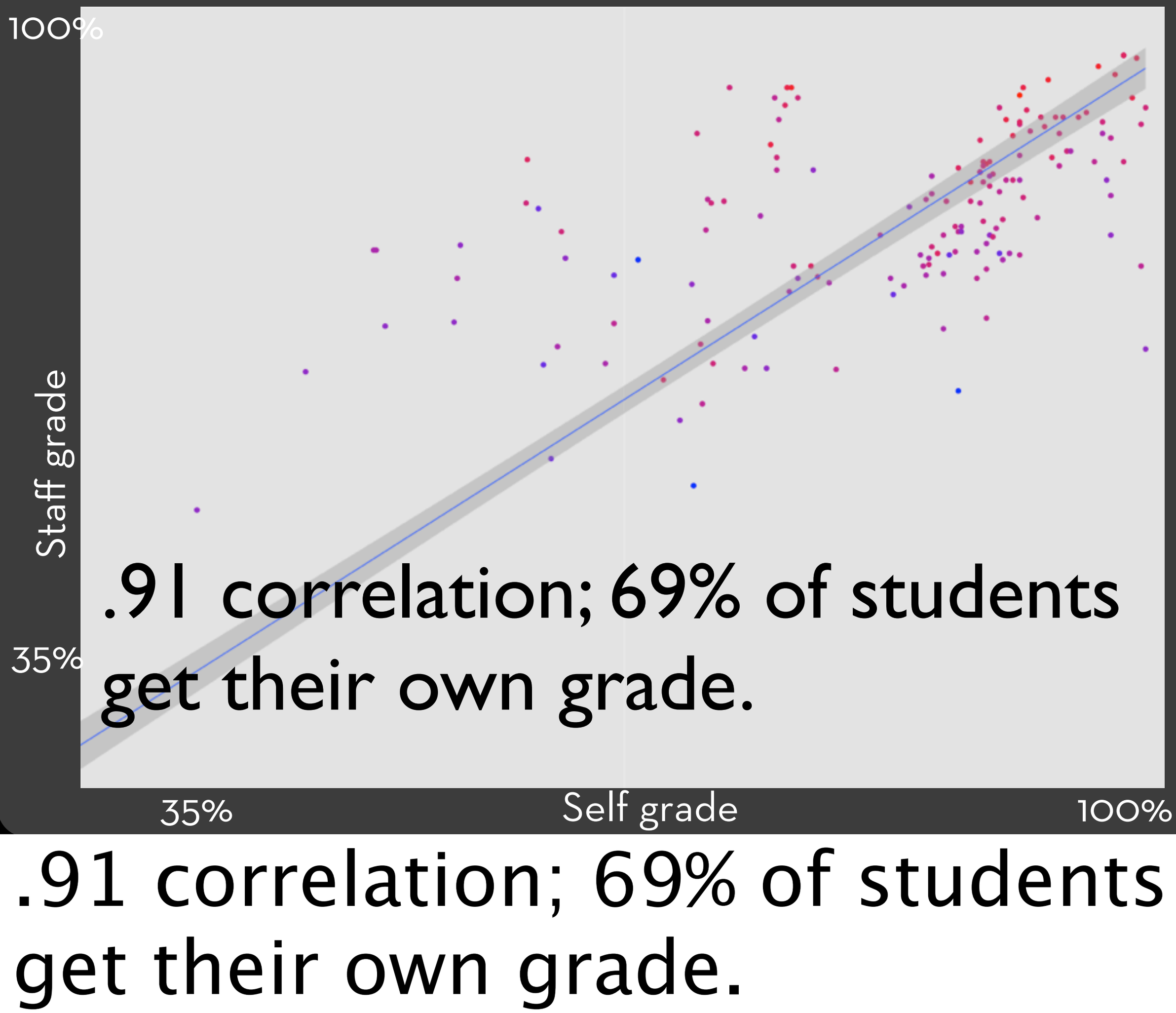
1. Subject #1 is currently in a meeting with two of her residents, discussing ways they can get involved in service. As the meeting goes on, the subject's computer is open to GMail the entire time - two instant message windows pop up within a few minutes of each other. The subject can't resist a brief glance every few minutes or so, but can't really respond in the moment, as she must listen to her residents.

Write only until the deadline; read only after. We link good examples from last year.

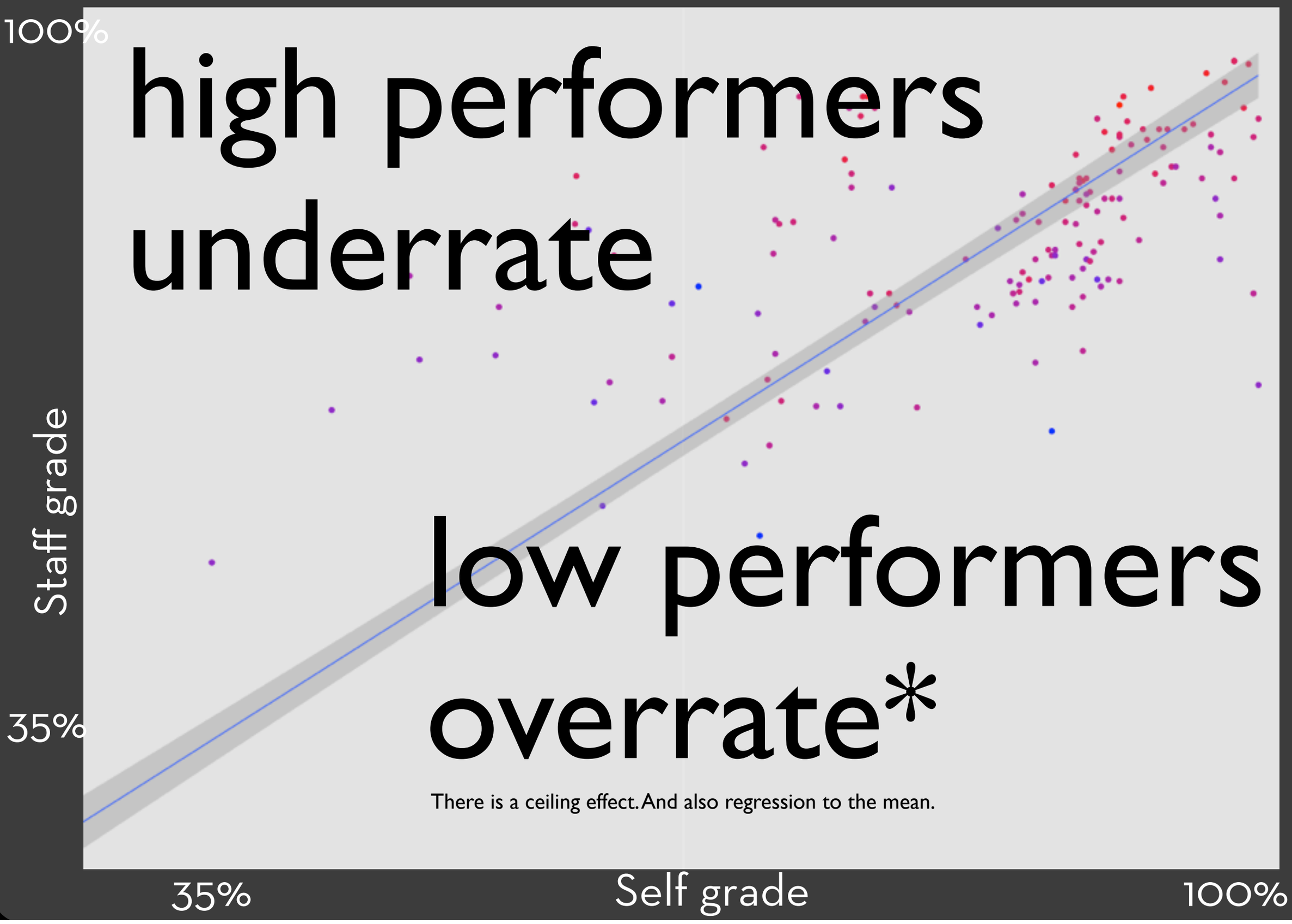
Providing "Bumpers"

If your grade is close to the TA grade, you get your own grade.

a game-theoretic analysis



.91 correlation; 69% of students get their own grade.



.91 correlation; 69% of students get their own grade.

Evaluators-> Coaches

Other universities leveraging these materials

- Berkeley, CMU, U Mass, Harvard, Maryland, Michigan, MIT, UNC, Olin, Pitt, Utah, Virginia Tech, Yale, UMD, Hawaii
- American university of Beirut, FAST National University, Islamabad, Sydney, Tehran Polytechnic, NTU Singapore, National University of Singapore, American University in Cairo, Open University in the Netherlands

(How) can we bring projects and peer critique online?

Three research questions

- How can student learn to assess others accurately?
- How can the system be robust to grading errors?
- How can students receive qualitative, personalized feedback?

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

Video lectures+ embedded questions, Weekly quizzes, open ended assignments

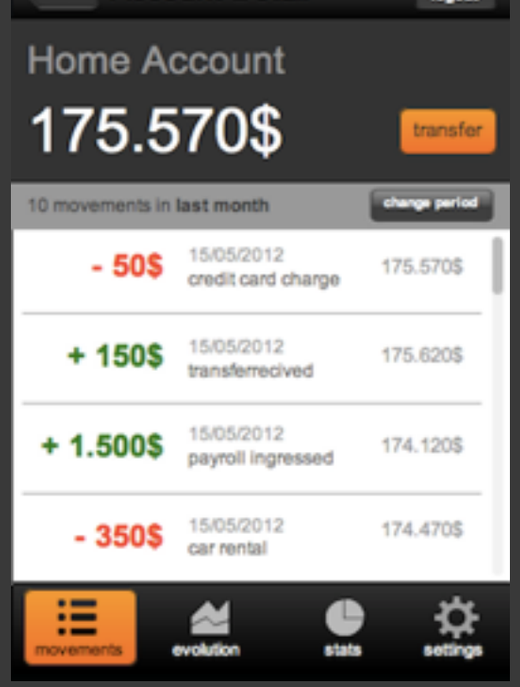
HCI Online	CS 147
7 weeks.	10 weeks.
1.5 hours of video lectures per week, broken up into 8-12 minute segments.	2 hours of in-person lectures per week, broken up into 1-hour class periods.
5 individual assignments.	Weekly, group assignments.
1 quiz per week with mini-quizzes embedded in lecture videos.	3 quizzes in total, taken in class.
Final project: complete design cycle to create a website.	Final project: complete design cycle to create mobile web app.
Peer- and self- assessments.	Staff- and self- assessments.
Statement of Accomplishment ("Apprentice" or "Studio").	Stanford University Credit.
No prerequisites.	Enrollment at Stanford; 2 quarters of programming

Cheating?

Per Jennifer Widom's advice, We designed for earnest students.

Student work

Task	Deadline	Status	Comment	Estimated hours	Actual hours spent
Create the project in Axure aplication	6/19/0212			30'	10'
Set up the project online (using dropbox) to access everyone	6/19/0212			10'	10'
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 me more time than expected to upload the first page	10'	30'
Test that the web can run in a mobile	6/20/2012		finally i'm not able to implement every functionality for mobile nav. I decide to implement a web simulation	30'	3
Create the navigational flow and main structure of the app in axure	6/21/2012			2	1
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			5	3
Upload and test	6/22/2012			1	2
Developing static main screens of the app I: List of accounts and general evolution	6/22/2012			3	4
			change name in settings and some interaction with the		



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)

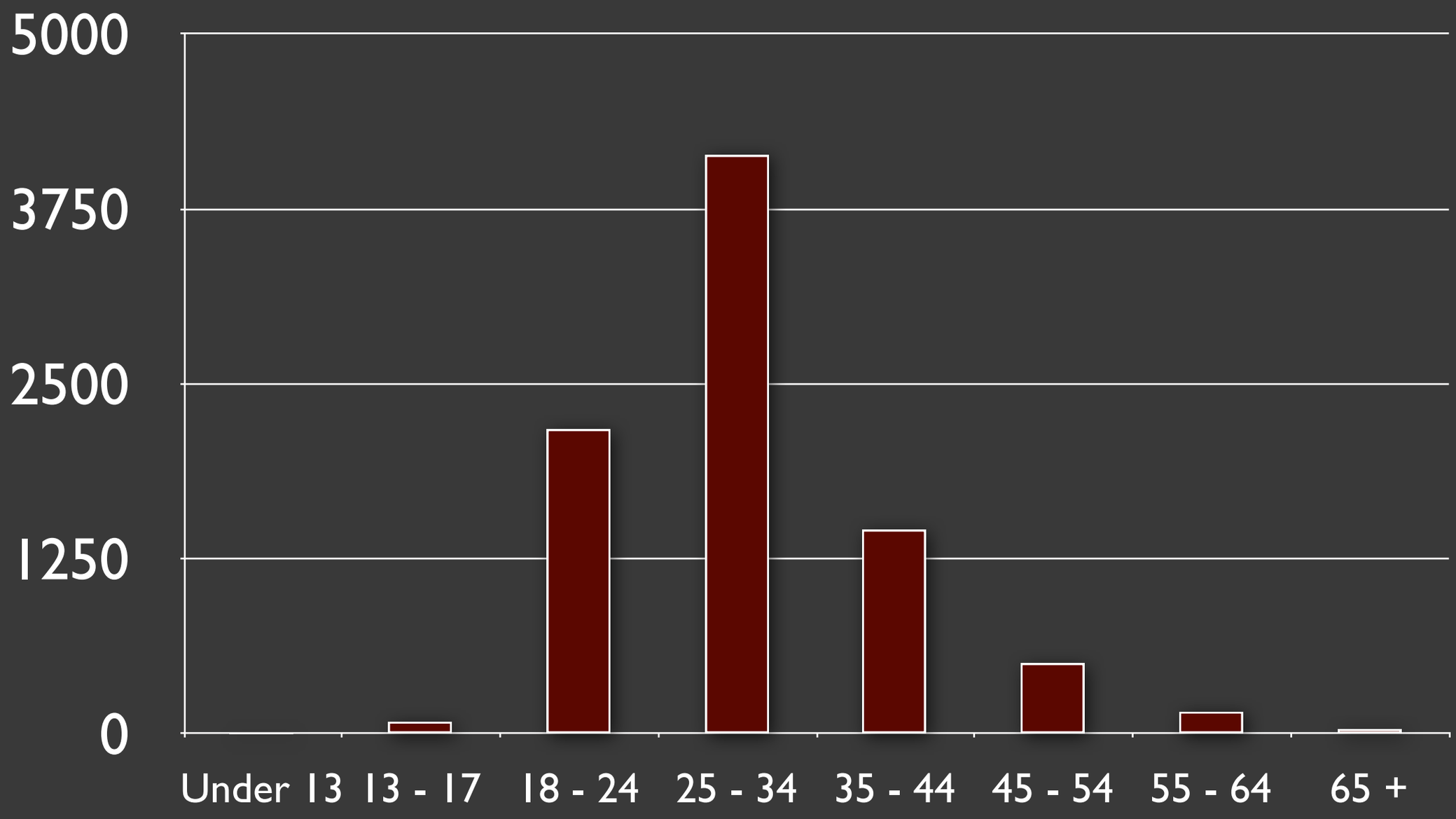
- Users don't guess what the stats option contains before visit the screen (3/3 users don't know it)

Summer 2012 numbers

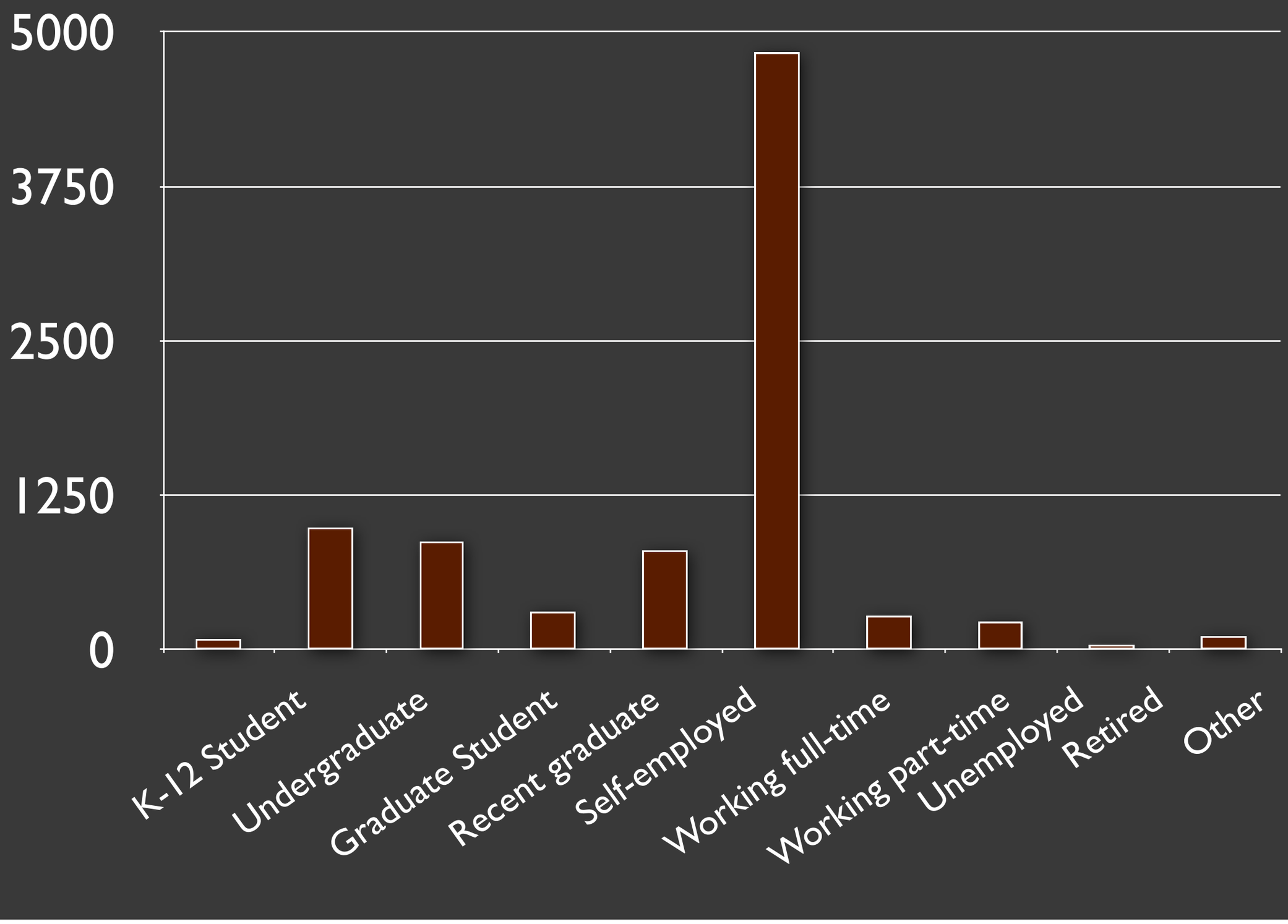
- 29,105 students watched video(s)
- 6,853 submitted quiz(zes)
- 2,470 completed an assignment
  - and 791 completed all 5
- From the 10,114 who filled out the initial survey...
  - 124 countries

We're really grateful that so many people wanted to participate in this crazy experiment. (Thankfully) unbeknownst to me ahead of time, a lot of my colleagues also watched the videos.

All ages; strong peak around 30



Many self-employed students



Anecdotally, a number of parents with young kids

10 self-organized meetup locations

- London, England
- Palo Alto, CA
- San Francisco, CA
- New York City, NY
- Buenos Aires, Argentina
- Bangladesh
- Portland, OR
- Mountain View, CA
- San Jose, CA
- Aachen, Germany

Students transcribed (some) lectures in 13 languages

- English, Spanish, Brazilian Portuguese, Russian, Bulgarian, Japanese, Korean, Slovak, Vietnamese, Chinese Simplified, Chinese Traditional, Persian, and Catalan

Through Amara

Chinese/China Study Group *Yucca (Student)* chinese china study qq group hci	jayant.jia 4 days ago	42	71	1.1k
Hungarian Study Group *Pal Hodosi (Student)* hungary hungarian magyar budapest szekesfehervar	Judit Ponya 20 hours ago	1	30	691
Pakistan Study Group *kanza batool haider (Student)* group pakistan	Farhan Ahmad 2 days ago	1	42	516
Ireland / Northern Ireland study group *Heather James (Student)* ireland	Dimitry Kireyenkov 2 days ago	0	17	172
Mexico | Grupo de Estudio y Encuentro *Laura Rodríguez-Loeches Zarragoitia (Student)* Mexico	Manuel Vidaurre 5 days ago	5	46	832
Ukrainian study group *Kirill Bigay (Student)* Ukraine Kiev Zaporizhzhya Odessa Lviv Kharkiv Dnipropetrovsk Berdyansk Donetsk Sevastopol Ivano-Frankivsk	Maksym Lushpenko 1 week ago	11	57	1.1k
Philippines Study Group *Christine balatbat (Student)* philippines pinoy	Claire Christine Arguelles 6 days ago	9	46	554
I.R. Iran study group *ali (Student)* Tehran Iran Persia Persian Iranian	Shaahin Mohammadi 4 days ago	0	16	303
Lithuanian Study Group *Audrius Prelgauskas (Student)* Lithuania Lietuva	Henrikas Kuryla 5 days ago	0	17	198
German-speaking study group *Marc Krenn (Student)* Europe german germany austria switzerland swiss	Katrin Siebler 1 week ago	4	24	393
Catalan study group *Marc Recasens (Student)* study catalan catalonia	Marc Recasens 1 week ago	5	39	604
South African/African Study Group *Siyabonga Africa (Student)* study group south africa	Edeh Victor Obinna 1 week ago	2	29	201
Spanish Study Group *Arnau Manyosa (Student)*	Jairo Felipe Pinedo Vergara 2 weeks ago	7	126	2.7k

From Coimbatore

Hello Scott,

This is Arun Martin, a usability analyst from Coimbatore, India. I have enrolled for your online HCI course last week and require your suggestions for the Design Brief on "Change."

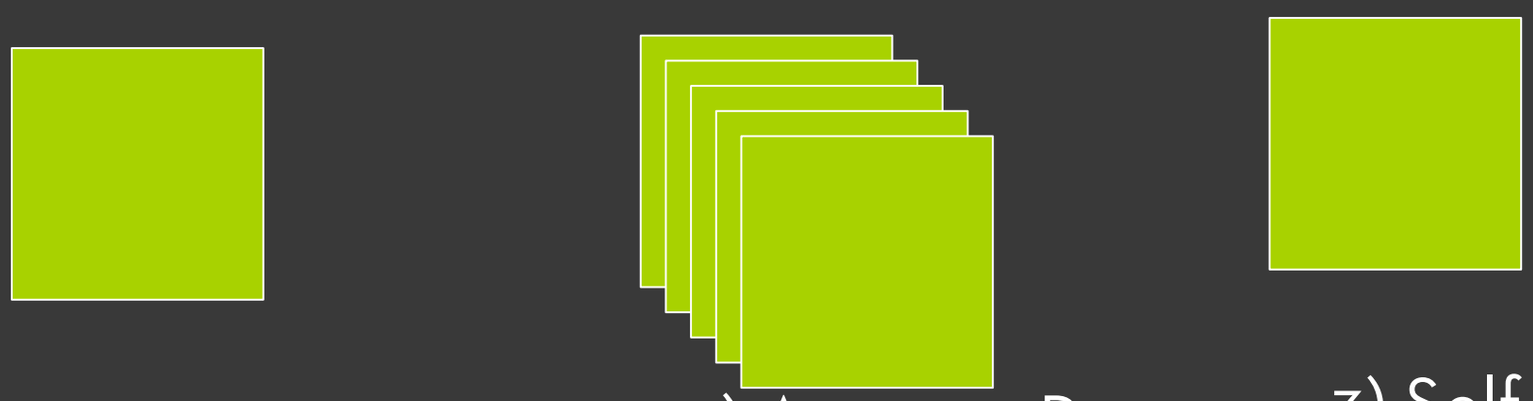
I plan to share my thoughts for commuters who start early from their offices, park their two-wheelers at a parking stand, and then commute to their offices through cabs or other modes of transport. I'm looking at the following challenge: How can technology help commuters to ascertain whether there are available parking spaces during morning hours so they can spend few minutes at home/ parking area to monitor their fitness levels and improve their health? Is this something that I can work on as a project?

Thanks,

Arun

I got this email at dawn

3-Step Peer Assessment



1) Calibration

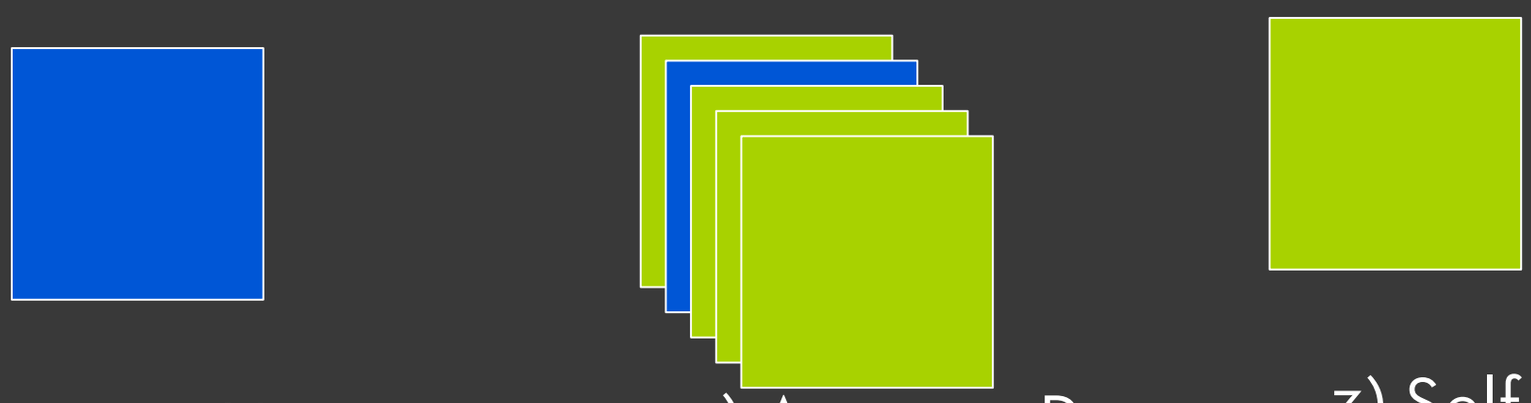
1) Assess 5 Peers

3) Self-Assess

For hci-online, students got the median of 5 grades

Creating Calibration & Ground Truth

- Assignments due most tuesdays at 11pm PDT
- Chinmay/Robi graded 8-10 submissions



1) Calibration

1) Assess 5 Peers

3) Self-Assess

Evaluation

How was the quality of the first prototype? Does it feel dynamic, like a working application? Was the student creative when implementing the interactions?

- **0 points:** No answer or completely irrelevant answer.
- **1 point:** The prototype is incomplete in significant ways. Many screens refer to screens that are not prototyped, and it's often unclear what a certain screen does.
- **3 points:** The prototype is mostly complete. The purpose of each screen is clear but maybe a user looking at the prototype may sometimes have a question about how to navigate between screens, how to use a form on a screen, or what some element on a screen is doing there.
- **5 points:** The prototype is detailed enough so that (1) a user can get a good feel for how the application works and flows and (2) a programmer can use the prototype to implement a simple skeleton application with a working interaction flow.

Intentionally Mostly numeric. Limited written. Language is an issue. see fortune cookie

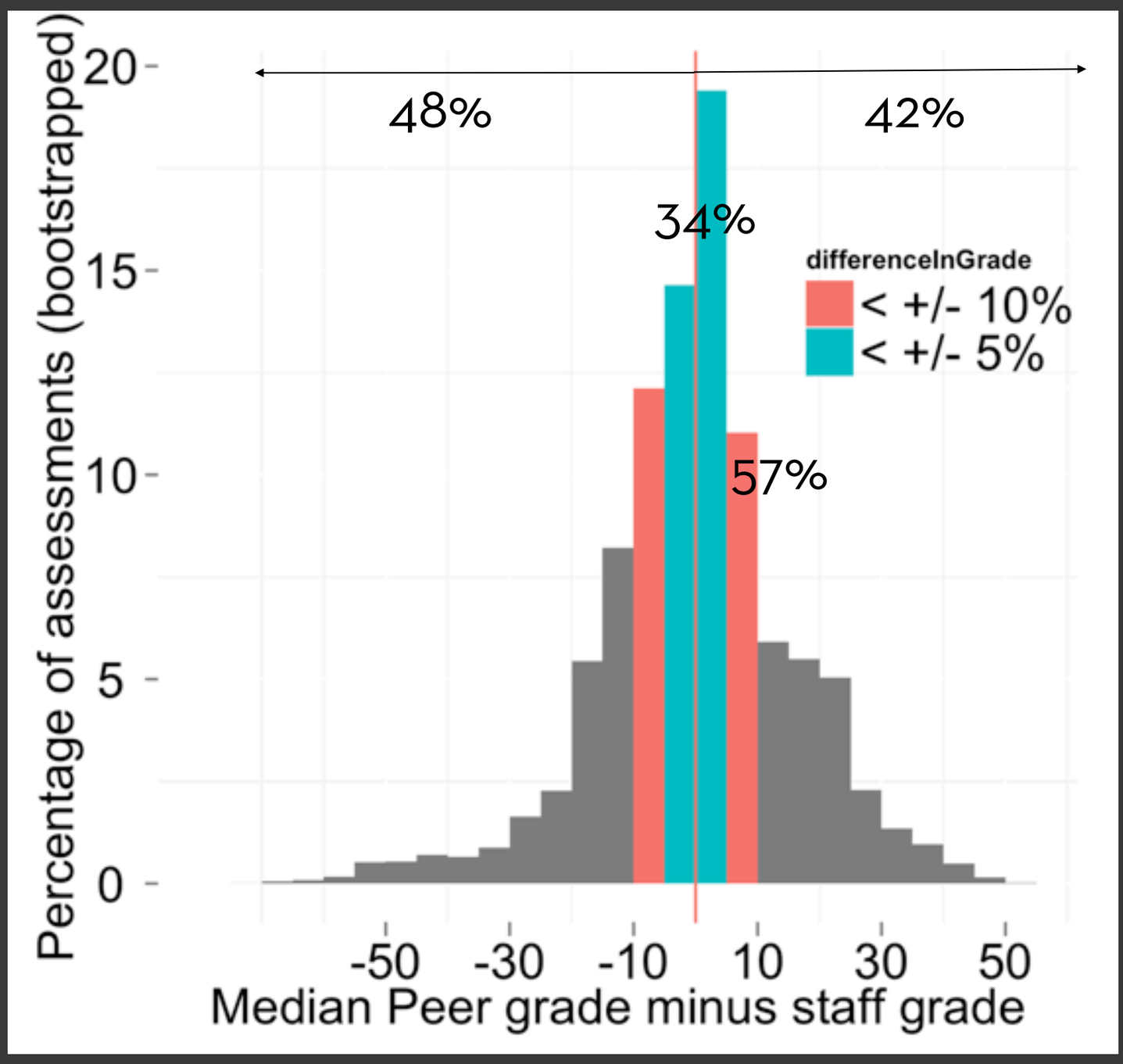
Summary: Assessments

	CS147	ONLINE
Correlation between self- and staff- assessments.	0.91	0.54
Correlation between self- and peer- assessments.	N/A	0.78
% of students who received their own grade.	69%	27%

While not high, these numbers are still kinda remarkable.

6,701 students submissions
across five assignments

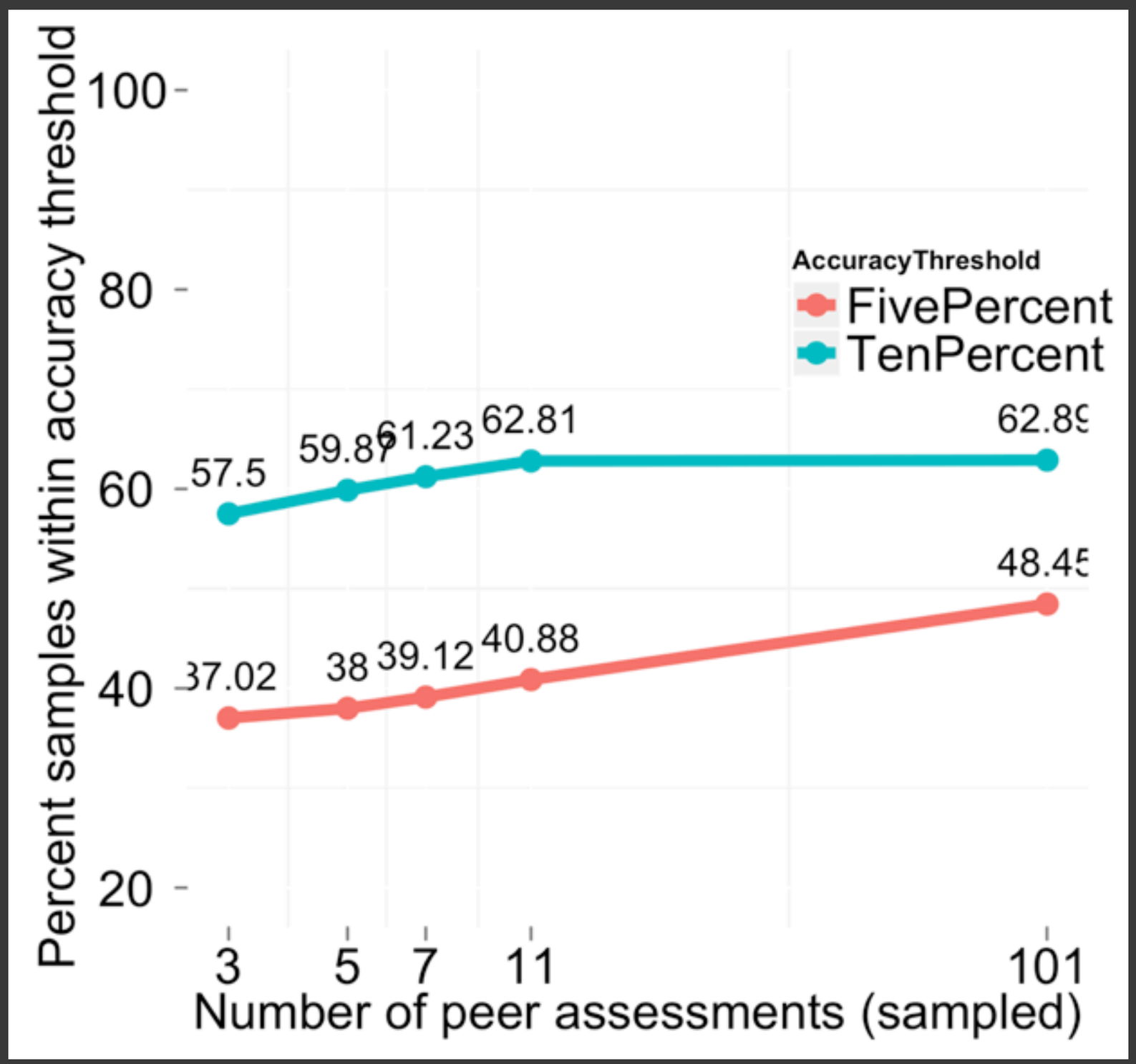
Errors were pretty balanced



This results in a relatively small number of ground truth submissions, and lots of grades for each. How do we use these to get an idea of how accurate grading is?

It is kinda remarkable that errors are distributed evenly. Errors may cancel out across the class. you get a higher and lower grade roughly evenly, so your final grade may have smaller error

More than five raters?



Curve essentially flat after 10 raters

The previous graph shows what happens when we have five raters. What happens when we have more?

Better than median?

- Best linear model:

X_1	X_2	X_3	X_4	X_5
0.035	0.170	0.208	0.329	0.329

  - 34.7% of samples within 5%, and 58.8% within 10%
  - Median model: 34% of such samples were within 5% of the staff grade, 57% within 10%.
  - Also tried: trimmed mean, etc.

Given a ranked list of assessments, is there a linear combination of grades that does better? marginally, but not by much.

This linear model gives 2% to the lowest, 32% to highest

Can we do better still?

- We hope so - and encourage your ideas

Budget-optimal Crowdsourcing using Low-rank Matrix Approximations

David R. Karger, Sewoong Oh, and Devavrat Shah
Department of EECS, Massachusetts Institute of Technology
Email: {karger, swoh, devavrat}@mit.edu

***Abstract*— Crowdsourcing systems, in which numerous tasks are electronically distributed to numerous "information piece-workers", have emerged as an effective paradigm for human-powered solving of large scale problems in domains such as image classification, data entry, optical character recognition, recommendation, and proofreading. Because these low-paid workers can be unreliable, nearly all crowdsourcers must devise schemes to increase confidence in their answers, typically by assigning each task multiple times and combining the answers in some way such as majority voting.**

In this paper, we consider a model of such crowdsourcing tasks and pose the problem of minimizing the total price (i.e., number of task assignments) that must be paid to achieve a target overall reliability. We give a new algorithm for deciding which tasks to assign to which workers and for inferring correct answers from the workers' answers. We show that our algorithm, based on low-rank matrix approximation, significantly outperforms majority voting and, in fact, is order-optimal through comparison to an oracle that knows the reliability of every worker.

I. Introduction

Background. Crowdsourcing systems have emerged as an effective paradigm for human-powered solving of large scale problems in domains such as image classification, data entry, optical character recognition, recommendation, and proofreading. Because these low-paid workers can be unreliable, nearly all crowdsourcers must devise schemes to increase confidence in their answers, typically by assigning each task multiple times and combining the answers in some way such as majority voting. In this paper, we consider a model of such crowdsourcing tasks and pose the problem of minimizing the total price (i.e., number of task assignments) that must be paid to achieve a target overall reliability. We give a new algorithm for deciding which tasks to assign to which workers and for inferring correct answers from the workers' answers. We show that our algorithm, based on low-rank matrix approximation, significantly outperforms majority voting and, in fact, is order-optimal through comparison to an oracle that knows the reliability of every worker.

I. Introduction

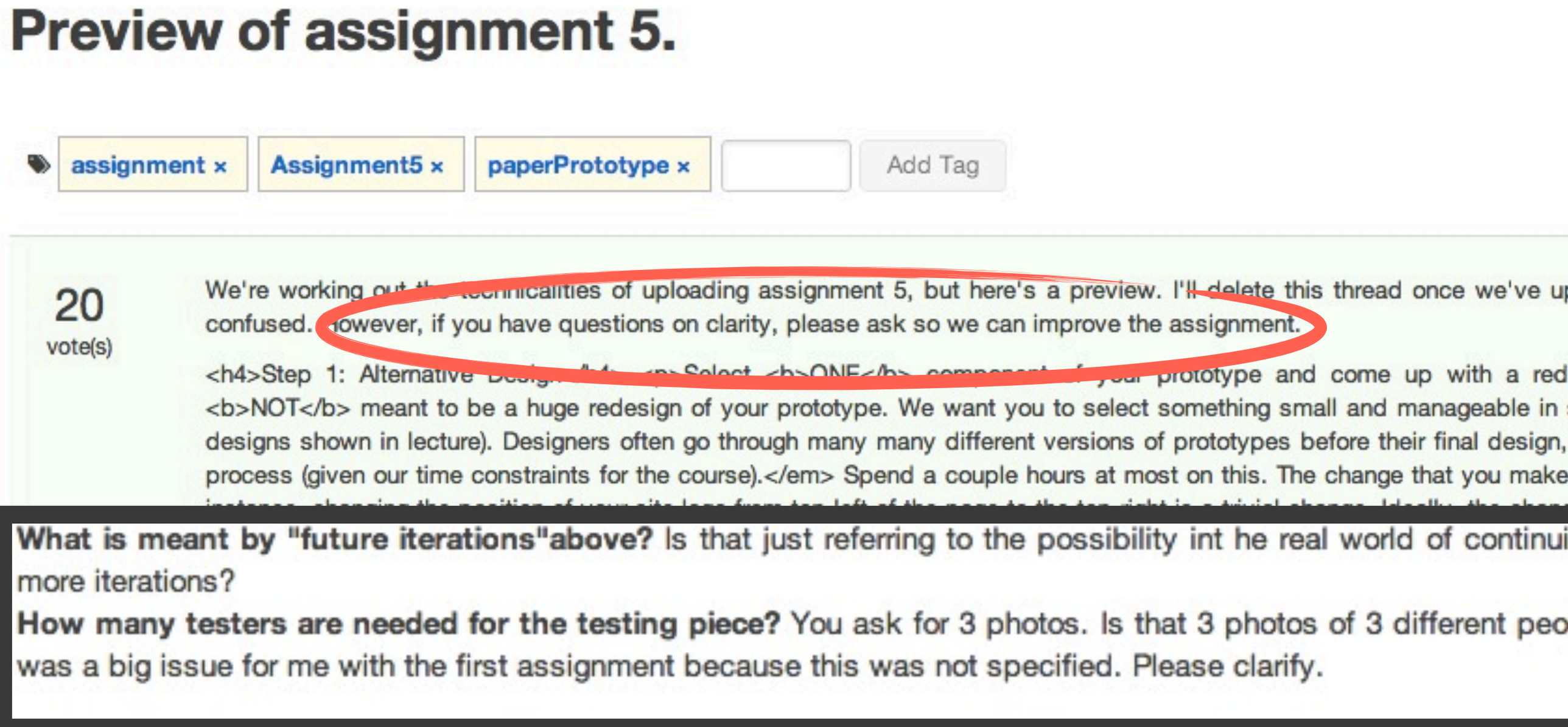
Background. Crowdsourcing systems have emerged as an effective paradigm for human-powered solving of large scale problems in domains such as image classification, data entry, optical character recognition, recommendation, and proofreading. Because these low-paid workers can be unreliable, nearly all crowdsourcers must devise schemes to increase confidence in their answers, typically by assigning each task multiple times and combining the answers in some way such as majority voting.

For such systems there is a natural core optimization problem to be solved. Assuming the taskmaster wishes to achieve a certain reliability in their answers, how can she do so at minimum cost (which is equivalent to asking how she can do so while asking the fewest possible questions)?

Several characteristics of crowdsourcing systems make this problem interesting. Workers are neither persistent nor identifiable; each batch of tasks will be solved by a worker who may be completely new and who you may never see again. Thus one cannot identify and reuse particularly reliable workers. Nonetheless, by comparing one worker's answer to others' on the same question, it is possible to draw conclusions about a worker's reliability, which can be used to weight their answers to other questions in their batch. However, batches must be of manageable size, obeying limits

This is MIT, and there's lots of smart people in this room. We can think of algorithms that get more signal from this data. As an example, [Student+Karger: budget optimal crowdsourcing].

Improving Assignments with student help

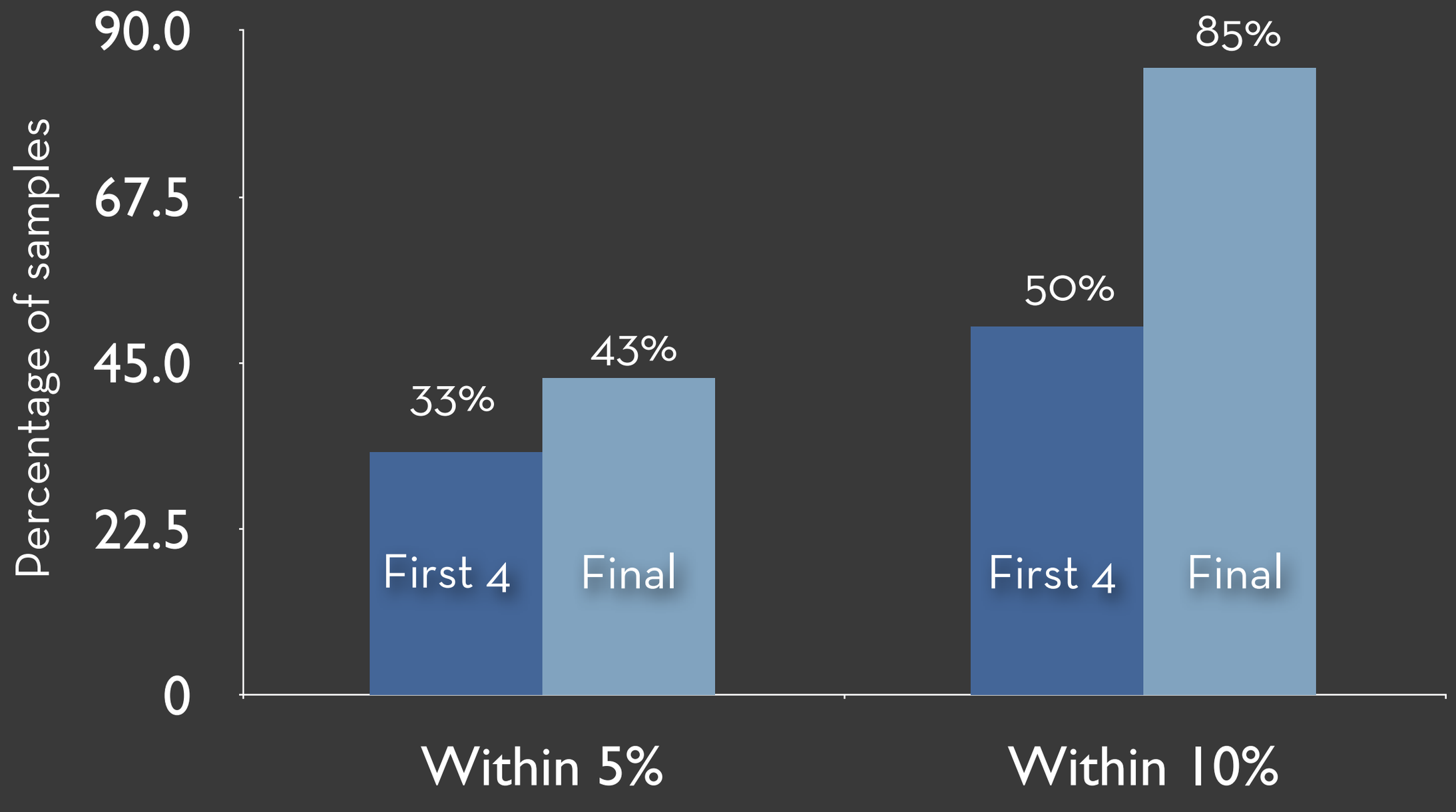


- Major Outcome: More Parallel Structure

One strategy is to try and extract more signal from the data. A complementary strategy is to get more signal in the data by improving assignments.

Logistics and clarity.

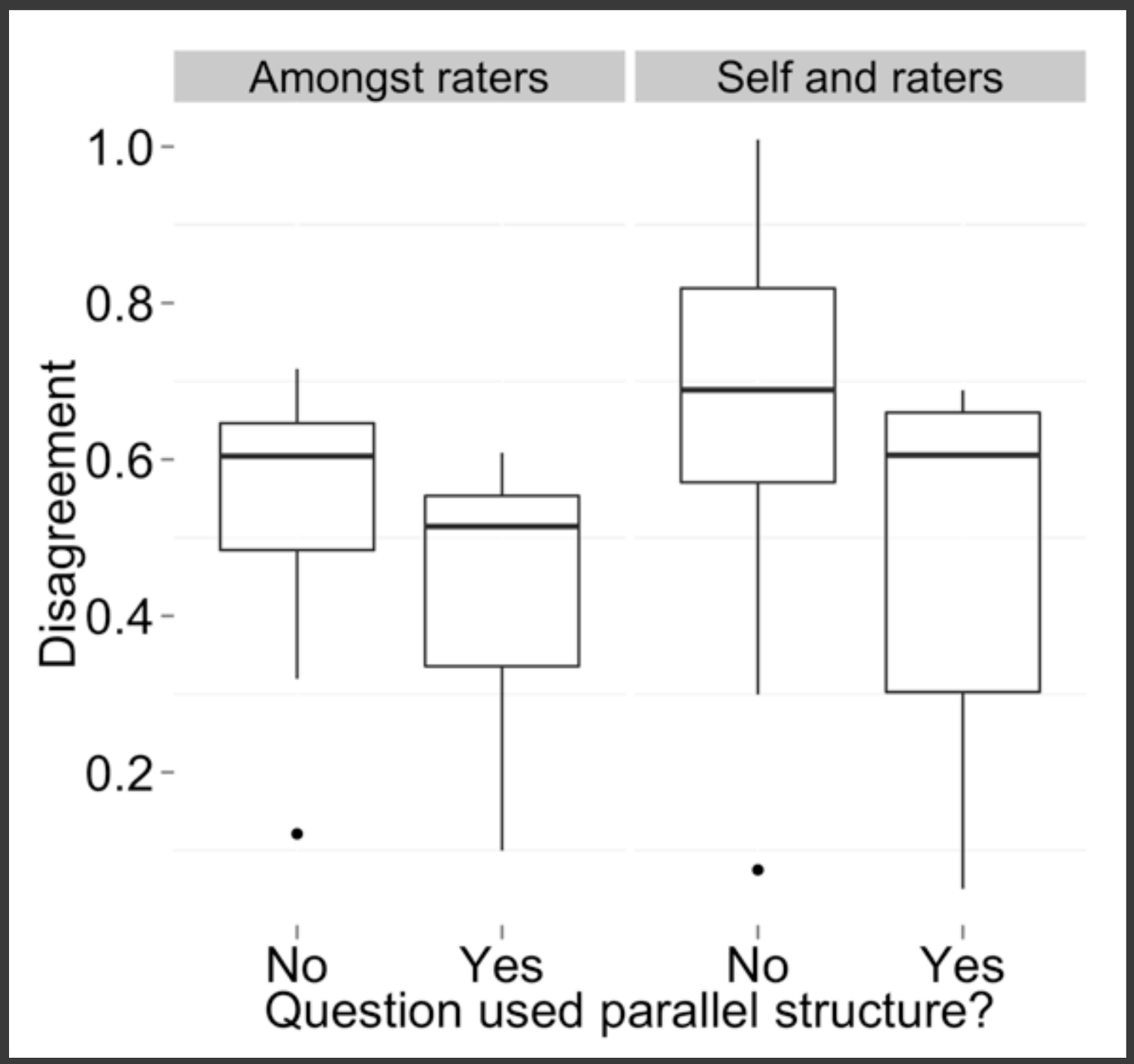
Higher agreement on final assignment



this analysis only uses raters who rated the last assignment. For these raters, accuracy is higher. 10% more samples were within 5% of staff grade, and 35% more within 10%.

The 85% is kinda remarkable.

Parallel structure helps



- F (1, 39) = 2.07, p < 0.05

example of parallel/non parallel str.

peer grading should allow us to leave more detailed comments for peers

I'm really excited about this peer grading, but I think we should be allowed to leave more detailed comments for our peers.

In particular some of the numeric scales seem a bit odd - the inspration diversity has like

- *2 points: Either the inspirations were obvious (that is, you could have come up with these without actually doing any observa they might confuse someone who had to read the inspirations and implement a solution based on them).*
- *3 points: Offered a diverse set of inspirations with insightful explanations. An HCI guru can spend an hour on the Internet ar*

Basically the options here are like "dude your inspirations were obvious" and "dude, you are like an HCI god!". At the very mir between.

I don't know that I would have had time to leave longer detailed comments, given that I had to complete reviews of 5 other stu the bumpy numeric system, and I would at least have liked the option of a text box to leave clarifying comments.

I get the sense the peer grading will be anonymous, but why couldn't it be the start of an interesting dialogue between peers - they continue through the course. Will I get another randomized set of five in the next assignment - that's interesting too, but I with my peers rather than getting sets of anonymous numbers from them ...

Anyhow, this is a great experiment and very much appreciate Prof. `Klemmer` and team's hard work. I think there's lots of roon likely have to wait till a second round of the course, but that's just fine - I'm along for the ride!

Many errors are variations on a few themes

Solution

+ “...because _____”

Provide Feedback as a packaged “fortune cookie” plus, optionally, a customized bit.
Explain the metaphor in terms of how implemented

Encouraging Richer Feedback

Overall feedback:

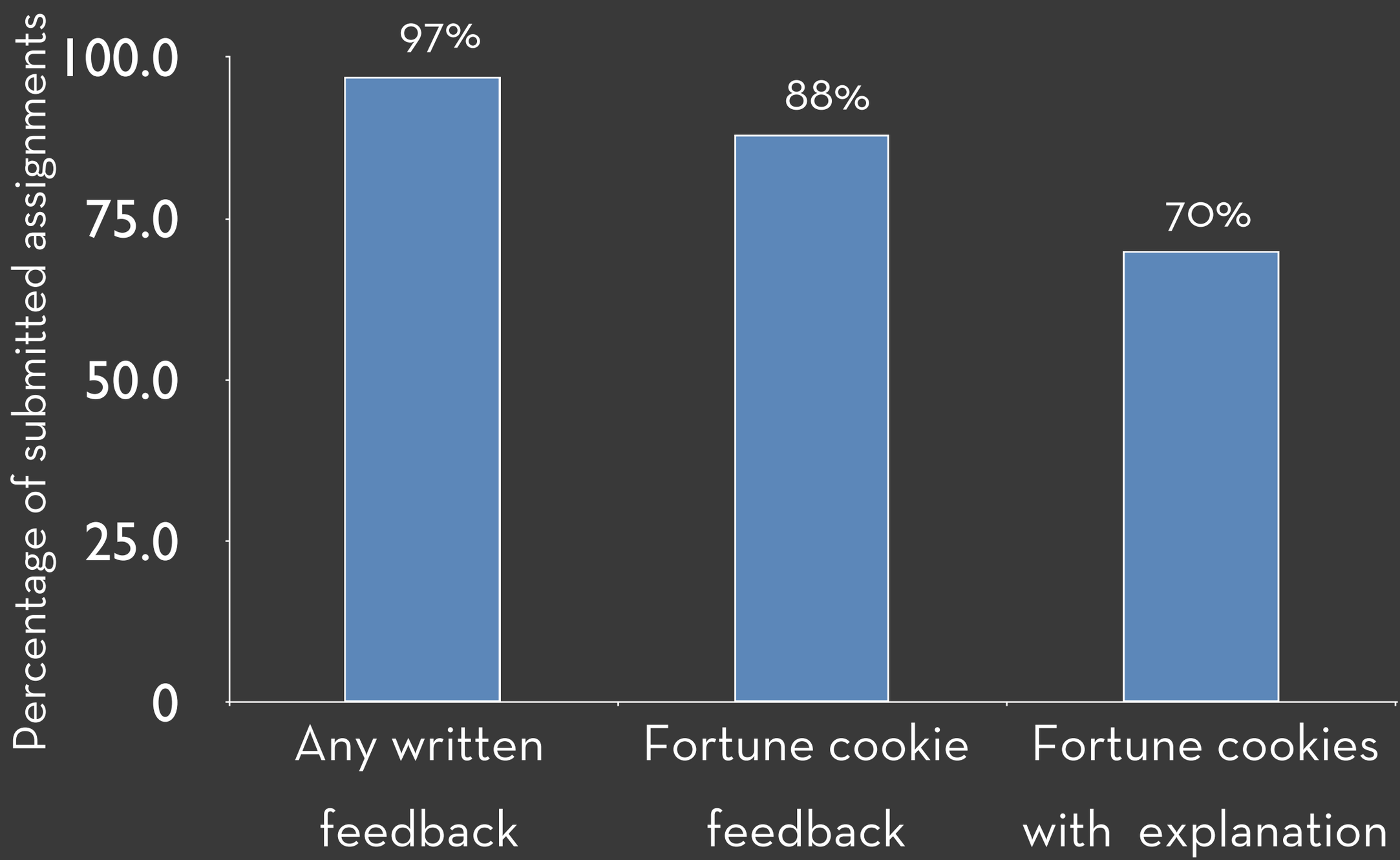
How could this student best improve his/her user evaluation? From among the following, copy one or more pieces of advice in the feedback box below.

- Test with more users
- Test in a different location/environment
- Test with a more functional prototype
- Create a more thorough script/plan
- Reflect on findings and find underlying reasons
- Test with a different alternate prototype
- Other

Test with more users: since a lot of the feedback about look and feel wasn't consistent across all users, testing wit
Test with a more functional prototype: especially for the part about looking inside an item, where functionality is not
Test with a different alternate prototype: maybe a different interaction flow would lead to a clearer interface, than jus

Use was optional

Fortune cookies provide personalized, qualitative feedback



Number with more than one feedback that was...

What do fortune cookies do?

- ***Recall*** to ***recognition***: a time honored HCI principle

- Most errors are variations on a theme: low-cost personalization

- Clean-slate critique hard: catalyze critical thinking

emphasize “implements a time honored principle of changing a recall task in to recognition” [cite ref]

Baking Fortune cookies
Nisha Masharani

Nisha was a summer intern in our group who had just completed her freshman year.

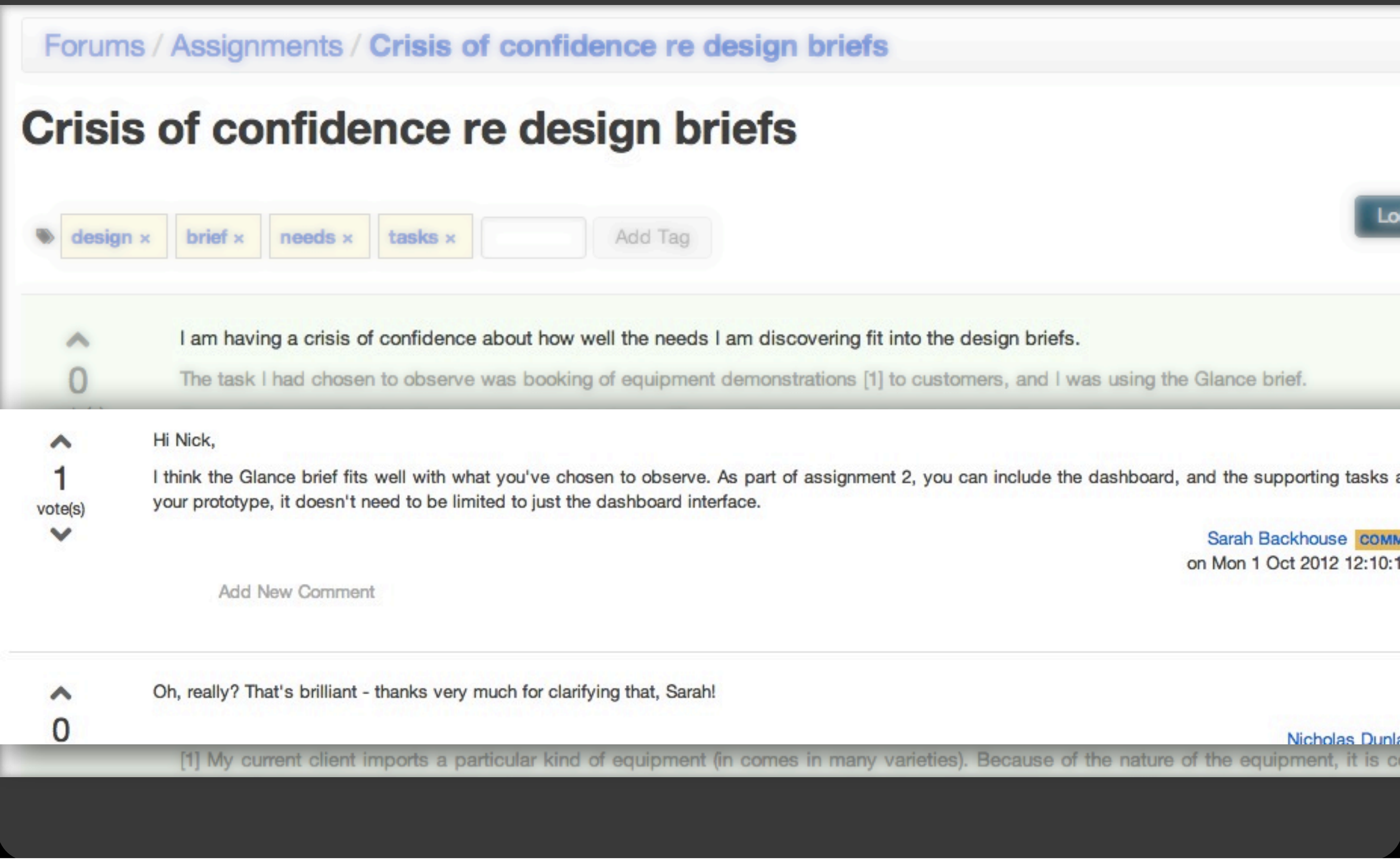
How to bake fortune cookies

- Look at three kinds of assignments

- Write feedback for assignment
- Aggregate
- 2/14 great, 11/14 needed minor changes, 1/14 not relevant
- Nisha took 3-4 hours per assignment

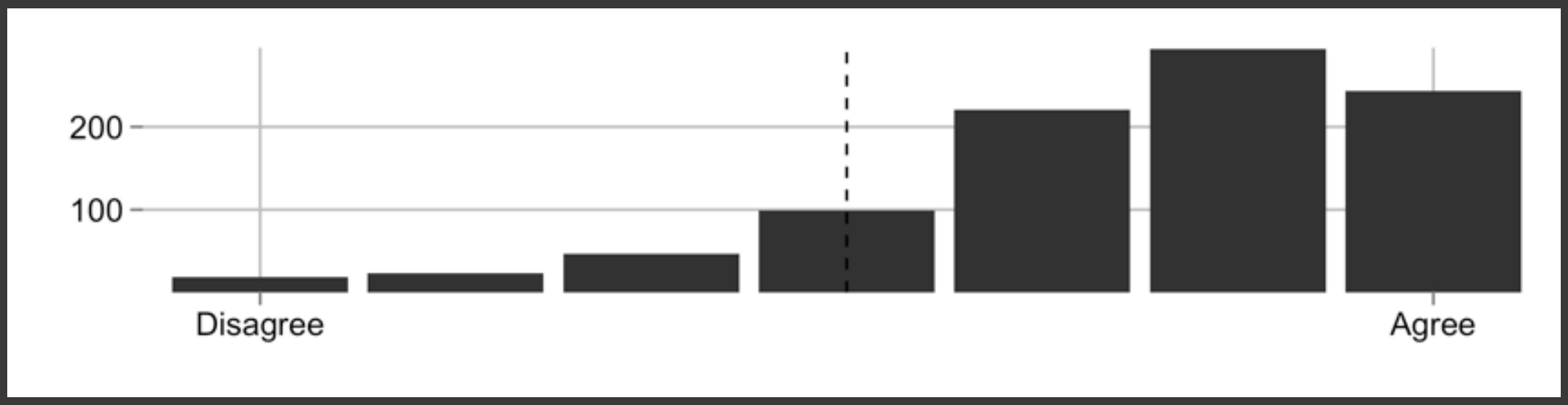
Nisha looked at assignments for 15 students: first and second assignment

Moving peer feedback earlier

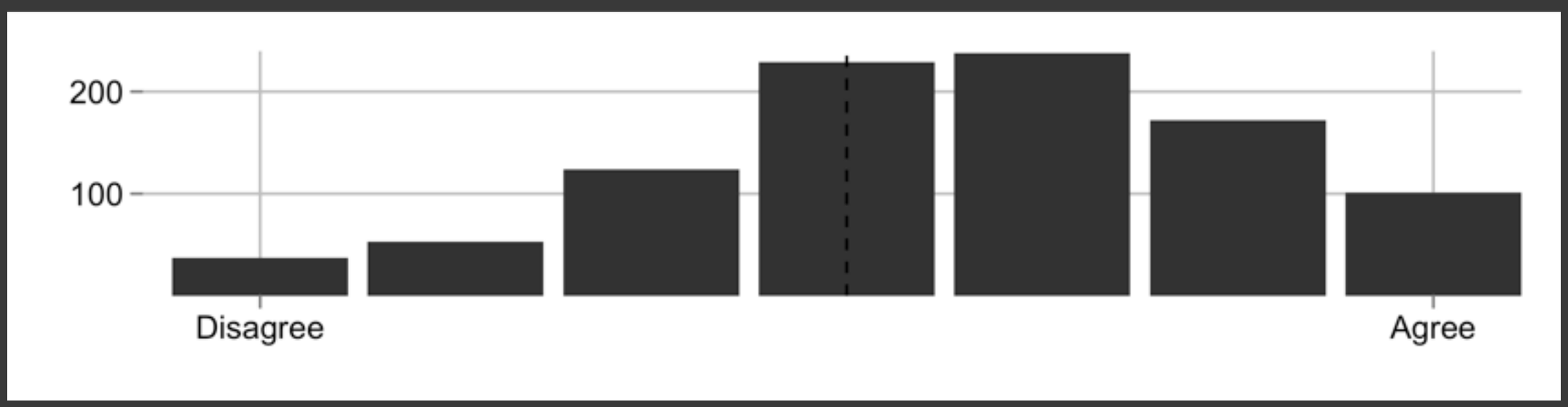


Student got the answer lasted 3 hours and 49 min

- I put sufficient effort into assessing peers



- Peers put sufficient effort into assessing me



Wilcoxon W = 610, 728, $p < 0.001$.

- How much did you feel like you learned from assessing **your own** work?
- How much did you feel like you learned from assessing **others'** work?

Students Took Ownership

- Sharing cool interfaces, resources, articles
- Collating reading lists, creating assignment aids
- Doing really creative work
- Helping other students
  - heuristic evaluation feedback
  - answering forum questions
  - extra peer assessment

Browse by Category

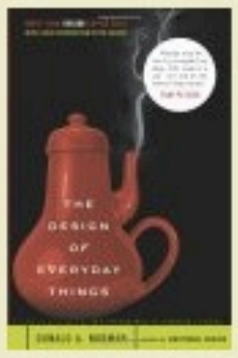
POWERED BY amazon.com

Lecture 1 - Introduction
Lecture 2 - Needfinding
Lecture 3 - Rapid Prototyping
Lecture 4 - Direct Manipulation
Lecture 6 - Representations
Lecture 7 - Visual Design
Lecture 8 - Information Design
Lecture 9 - Designing Experiments
Lecture 10 - Running Experiments

Recommended Books of Coursera Stanford HCI Course

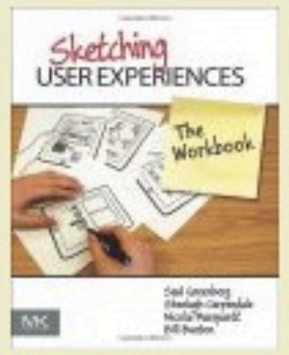
Lecture 1 - Introduction

Shopping Cart



The Design of Everyday Things
by Donald A. Norman
$11.53
Introduction

Sketching User Experiences: Getting the De...
by Bill Buxton
$29.48

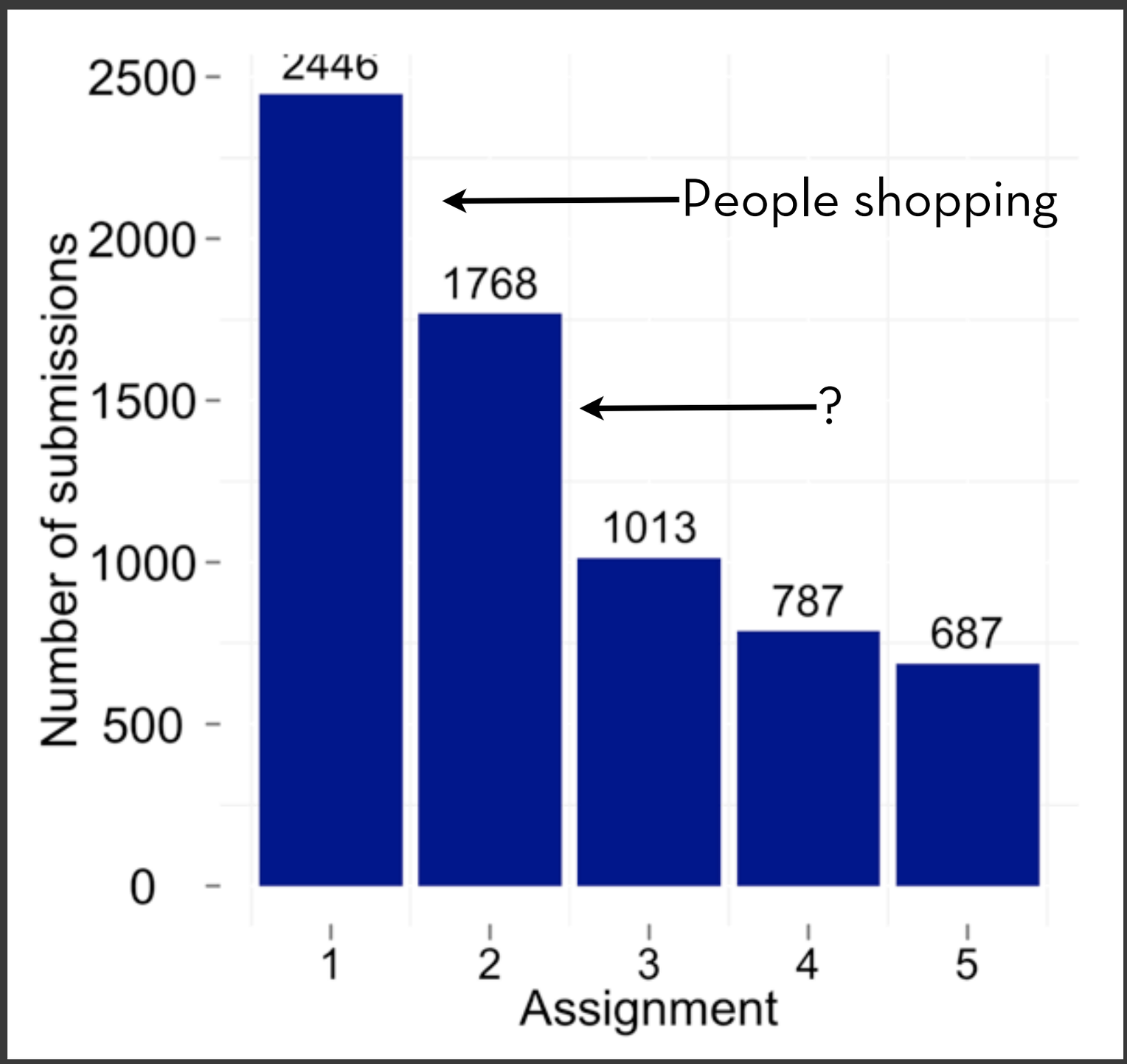


Sketching User Experiences: The Workbook
by Saul Greenberg
$17.38

From Counterculture to Cyberculture: Stewar...
by Fred Turner
$10.01

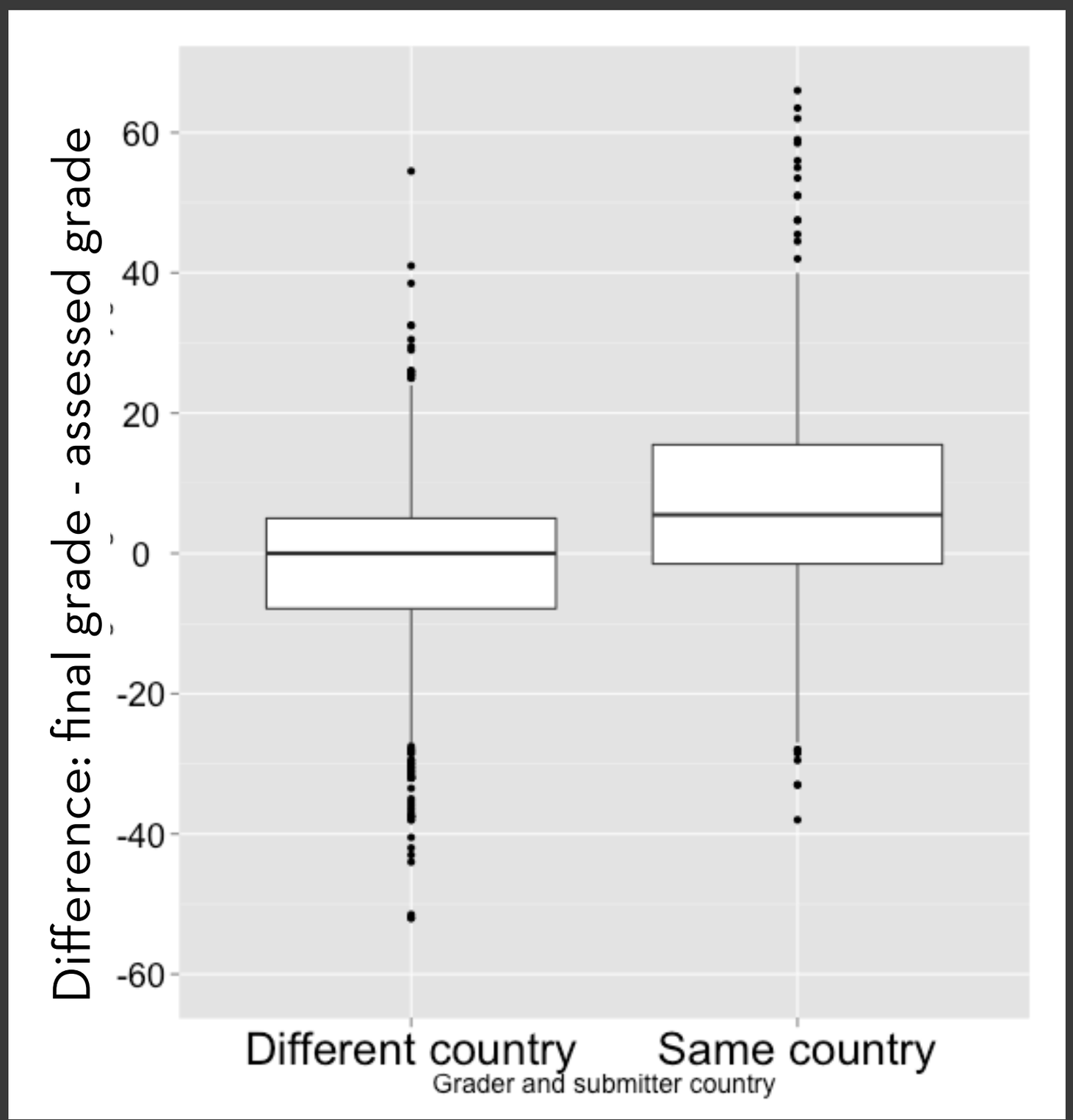
What the Dormouse Said: How the Sixties Cou...
by John Markoff
$9.10

Student Submissions Tapered Off



I told you before that we had 6000-odd submissions. Sadly, these submissions are not evenly distributed. While it is understandable that some students are shopping, those that complete two assignments have made a considerable investment.

Patriotic Grading



Thanks to...

- Stanford: Jane, Robi, Alex, Chinmay, Kathryn...
- Coursera: Huy, Daniel, Pang Wei, Ngiam, Daphne, Andrew...
- Colleagues who let me show their work
- Students going above and beyond: Amy...