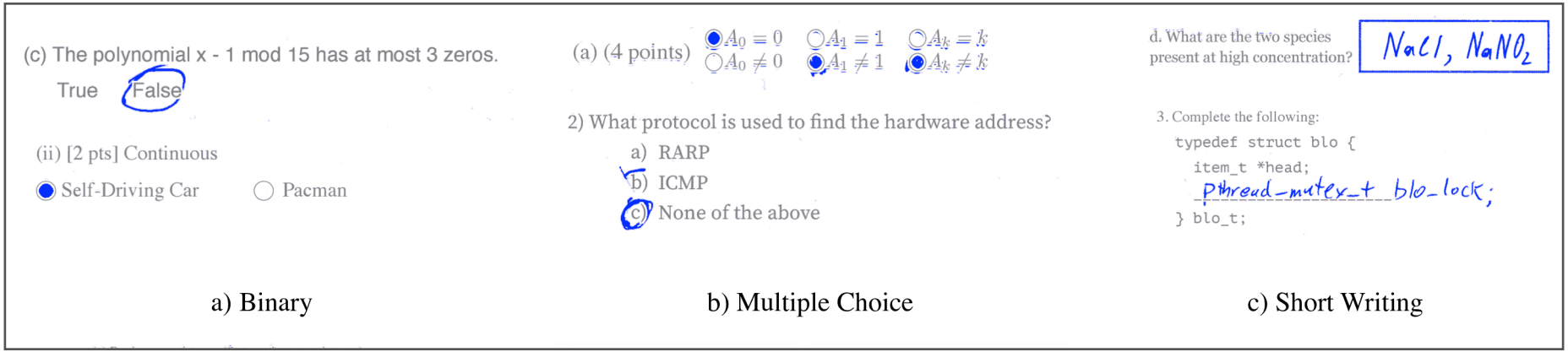
How Do Professors Format Exams?

An Analysis of Question Variety at Scale



Paul Laskowski, Sergey Karayev, and Marti Hearst

Learning at Scale 2018

Who

Paul Laskowski

UC Berkeley

Professor at School of Information

Sergey Karayev

Gradescope, Inc.

Co-founder, Director of Research

Marti Hearst

UC Berkeley

Professor at School of Information & CS

What

- We analyzed questions on nearly 1,800 paper exams graded in real STEM courses
- We annotated the type of each question (multiple choice, short/medium/long writing, drawing)
- We found interesting differences by type in position, use across subjects, student performance, and reliability
  - e.g. 5+ binary choice questions are required to match reliability of 1 long writing question

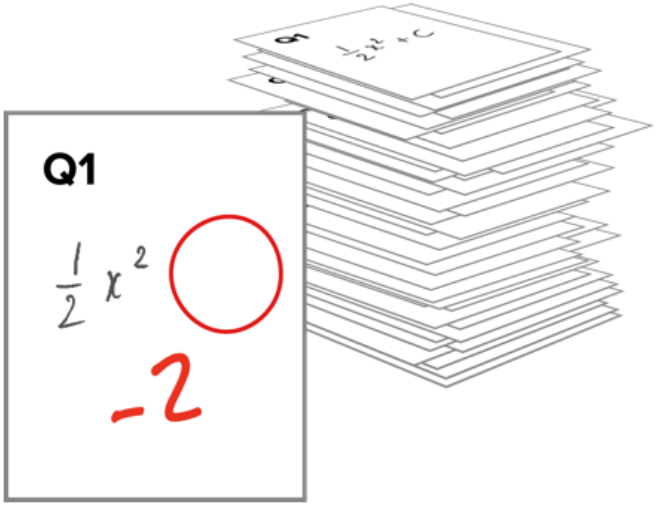
Why

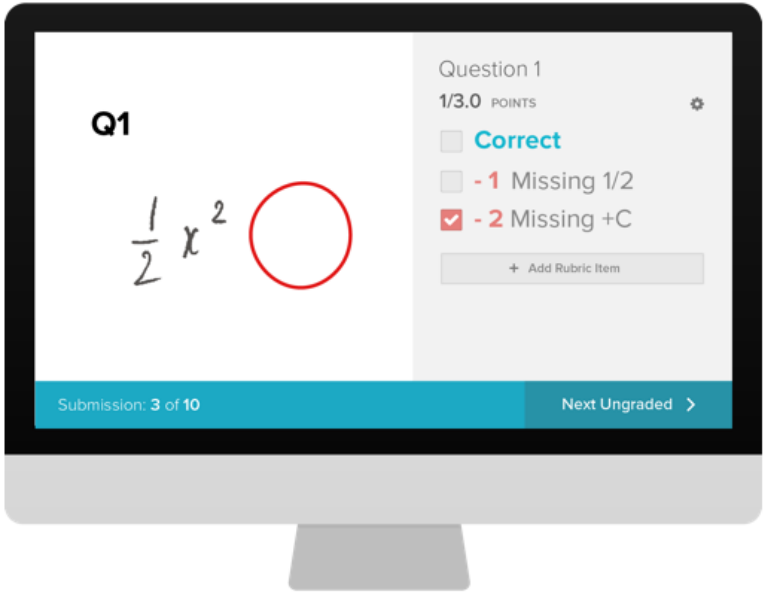
- Unique vantage point on exams “in the wild”
- Near term: get an **understanding** of what is happening today
- Longer term: **improve** exam writing for both for instructors and students

Related Work

- The trend: oral open-response -> written open-response -> multiple-choice
- Good assessment is valid and reliable.
  - Ideally, summative assessment is also discriminative and easy to grade without bias.
- Many studies on open-response vs. multiple-choice
  - Well-written MCs *can* be as good as essays
  - MCs are hard to write, easy to grade. Essays are the opposite.
- Ordering effects are unclear

Data Source

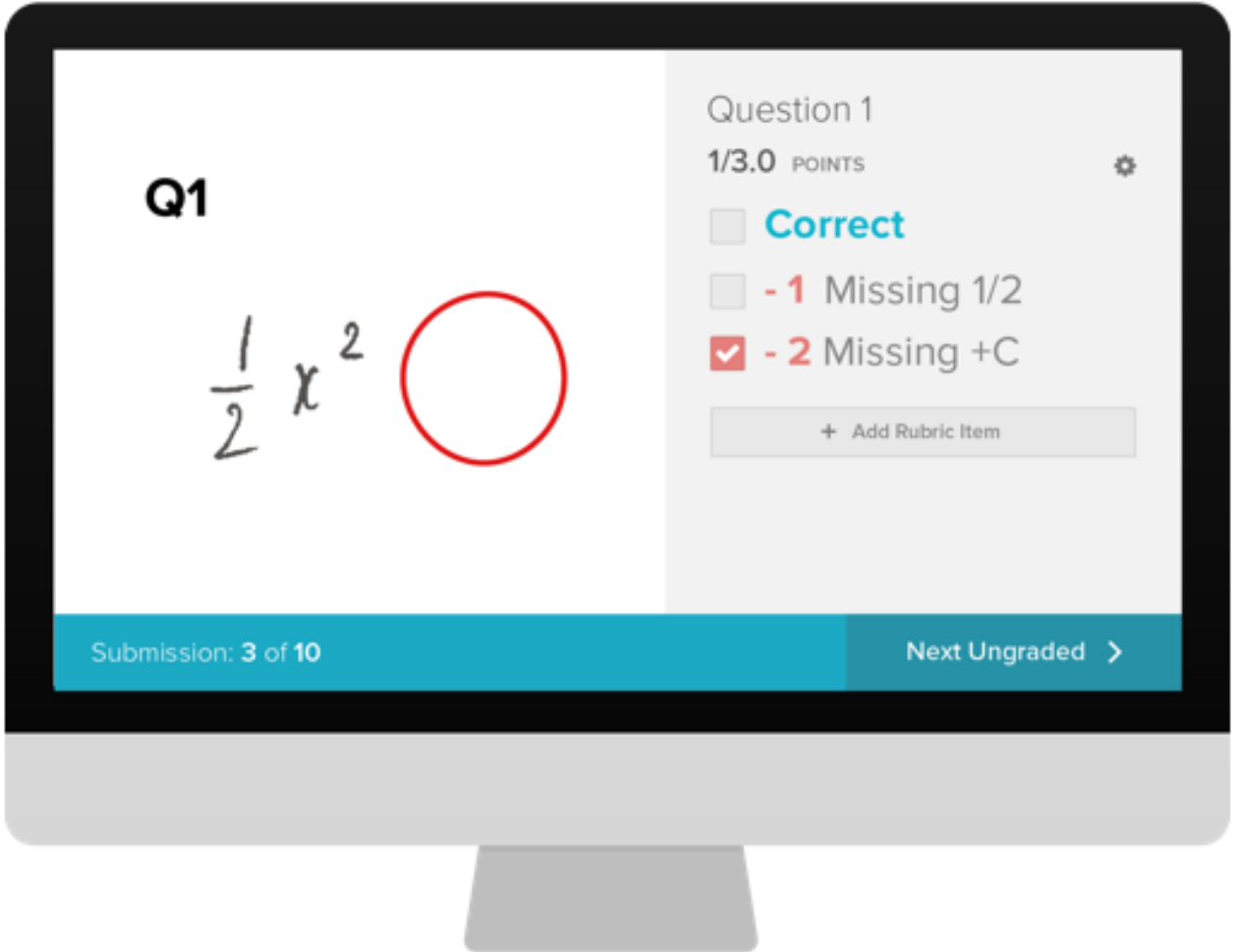




Gradescope: a Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work

Arjun Singh, **Sergey Karayev**, Kevin Gutowski, Pieter Abbeel

Learning at Scale 2017



- No constraints on type of questions
- Consistent scoring due to shared, modifiable rubric

Data Annotation

- 40M+ student answers
  - graded by 6,000+ instructors (mostly STEM)
  - at 600+ schools (mostly US higher-ed)
- Built special annotation interface, hired educators
- Annotated over 50% of 50,000+ questions that had at 100+ student answers at the time

REGION 1 | + ADD REGION | Unsuitable | Save

RESPONSE AREA

- ○ Other
- ○ Bubble or Square MC
- ○ Other kinds of MC
- ○ Small fill-in-the-blank
- ○ Line-sized
- ○ Paragraph-sized
- ○ Page-sized
- ○ Matrix or grid
- ○ Diagram, plot, or chem structure
- ◉ Box for drawing

TYPICAL RESPONSE LENGTH

- ◉ N/A
- ○ Character, word, or words
- ○ Sentence or a couple of lines
- ○ Paragraph or more

RESPONSE TYPE

Binary MC | Multiple MC | Text | Code | Math | Science Symbols | Diagram | Diagram Annotation | **Chemical Structure** | Self-boxed | Justification/Extra Work | Final Answer | Not English | Other

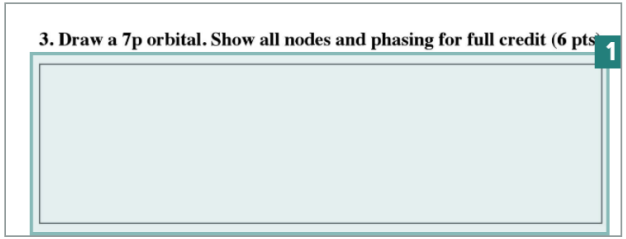
Remove | ☑ Set Bounding Box

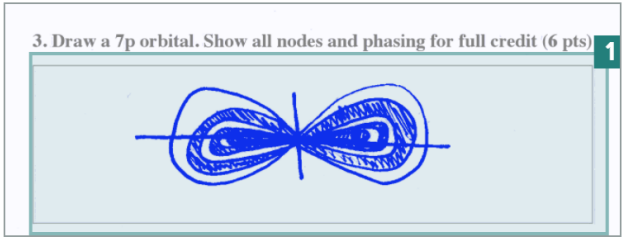
Rubric 6.0 points, floor, ceiling

0	Correct
-3	Incorrect phasing
-6	Fully incorrect

Sample Answers 2 answers

Mask pages





Review Groups | Grade | Next Annotated | Next Random | **Save & Next Random ›**

REGION 1

+ ADD REGION

Unsuitable

Save

RESPONSE AREA

- Other
- Bubble or Square MC
- Other kinds of MC
- Small fill-in-the-blank
- Line-sized
- Paragraph-sized
- Page-sized
- Matrix or grid
- Diagram, plot, or chem structure
- Box for drawing

TYPICAL RESPONSE LENGTH

- N/A
- Character, word, or words
- Sentence or a couple of lines
- Paragraph or more

RESPONSE TYPE

Binary MC | Multiple MC | Text | Code | Math

Chemical Structure | Self-boxed | Justification/Ext

Remove

Set Bounding Box

Final Dataset

- Almost 23,000 questions
- From almost 1,800 exams
- Corresponding to graded answers of 120,000+ students

Question Type Samples

(c) The polynomial x - 1 mod 15 has at most 3 zeros.

True False

(ii) [2 pts] Continuous

Self-Driving Car Pacman

a) Binary

(a) (4 points) $A_0 = 0$ $A_1 = 1$ $A_k = k$

$A_0 \neq 0$ $A_1 \neq 1$ $A_k \neq k$

2) What protocol is used to find the hardware address?

a) RARP

b) ICMP

c) None of the above

b) Multiple Choice

d. What are the two species present at high concentration? NaCl, $NaNO_2$

3. Complete the following:

```
typedef struct blo {
  item_t *head;
  pthread_mutex_t blo_lock;
} blo_t;
```

c) Short Writing

(e) Pre-image resistance (first pre-image resistance)

Given a hash value h, it should be computationally infeasible to find x such that h = h(x)

(c) (2 pts) What is the distance to default?

Distance to default is defined as $\ln(\frac{V}{B})$

Since the debt level is 90% of the value

from value $\frac{1}{0.9} = \frac{V}{B} = 1.1111\ldots$

$\ln(\frac{1}{0.9}) = 0.105$

d) Medium Writing

1. (10 points) An unknown compound contains carbon, hydrogen, and oxygen. A 12g sample was combusted to produce 36.62g of carbon dioxide and 14.99 of water. What is the formula of the unknown compound?

$36.62\,g\ CO_2 \times \frac{12.01\,g}{43.99} = 9.997\,g\ C$ > $\frac{11.67939}{5.321970}$

$14.99\,g\ H_2O \times \frac{2.02\,g\ H}{18.015} = 1.6812\,g\ H$

$9.997\,g\ C \times \frac{1\,mol\ C}{12.01\,g\ C} = 0.8323\,mol\ C$

$1.6812\,g\ H \times \frac{1\,mol}{1.01\,g\ H} = 1.6655\,mol\ H$

$3.3217\,g\ O \times \frac{3\,mol\ O}{15.999\,g} = 0.20762\,mol\ O$

$\frac{0.8323}{0.20762} = 4\ C$ $\frac{1.6655}{0.20762} = 8\ H$ $\frac{0.20762}{0.20762} = 1$

C_4H_8O

4. (20 pts) Validation is a key concept in Minsky's financial instability hypothesis. Write an essay to explain this concept. Pay attention to how validation influences lending standards. Include at least one example of validation from the subprime lending crisis. What are the implications of validation leading up to the Great Recession for future business cycles?

The hypothesis centers around the concept that 3 cycles exist in the marketplace, because of the inherent fluctuation in the expectations and actions of people. The degree in the validation for lending standards, changes as which prospects change.

When the economy is doing well, lending standards are low and banks and lenders validate easily. The economy continues to do well & grow, and and lending increases day of the expansion of credit (because validation is easy). What should be happening is finance companies into speculative finance, where too much money is then lent out to cover of the quickly issued investment.

Lastly, this happens into a poor finance environment where debts are paid through increasing debt. This is like the subprime lending crisis where validation and lending went over the edge because of the exponential credit.

e) Long Writing

3. Draw a 7p orbital. Show all nodes and phasing for full credit (6 pts)

A). (4 pts) Use the HA-tag to purify your favorite protein. Draw a diagram showing the purification setup.

f) Drawing

(c) The polynomial x - 1 mod 15 has at most 3 zeros.

True False

(ii) [2 pts] Continuous

Self-Driving Car Pacman

a) Binary

(e) Pre-image resistance (first pre-image resistance)

(a) (4 p

2) What

1. (10 points) An unknown co
combusted to produce 36.6

zeros.

(a) (4 points) ● $A_0 \equiv 0$ ○ $A_1 \equiv 1$ ○ $A_k \equiv k$
○ $A_0 \not\equiv 0$ ● $A_1 \not\equiv 1$ ● $A_k \not\equiv k$

2) What protocol is used to find the hardware address?

a) RARP

b) ICMP

c) None of the above

b) Multiple Choice

1. (10 points) An unknown compound contains carbon, hydrogen, and oxygen. A 15g sample was

4. (20 pts) Validation is a key concept in Minsky's financial instability hypothesis. Write an essay to explain this

$= k$
$\neq k$

re address?

d. What are the two species present at high concentration?

$\boxed{NaCl,\ NaNO_2}$

3. Complete the following:

```
typedef struct blo {
  item_t *head;
  pthread_mutex_t blo_lock;
} blo_t;
```

c) Short Writing

ty hypothesis. Write an essay to explain this
eds. Include at least one example of validation from

(e) Pre-image resistance (first pre-image resistance)

Given a hash value h, it should be computationally infeasible to find x such that $h = h(x)$

(c) (2 pts) What is the distance to default?

Distance to default is defined as $\ln\left(\frac{V}{B}\right)$

Since the debt level is 90% of the value

from value $\frac{1}{0.9} = \frac{V}{B} = 1.1111\ldots$

$\ln\left(\frac{1}{0.9}\right) = 0.105$

d) Medium Writing

1. (10 points) An unknown compound contains carbon, hydrogen, and oxygen. A 15g sample was combusted to produce 36.62g of carbon dioxide and 14.99 of water. What is the formula of the unknown compound?

$36.62\,g\ CO_2 \times \frac{12.01\,g}{43.99} = 9.997\,g\ C$

$14.99\,g\ H_2O \times \frac{2.02\,g\ H_2}{18.019} = 1.6812\,g\ H$

$9.997\,g\ C \times \frac{1\ mol\ C}{12.01\,g\ C} = 0.8323\ mol\ C$

$1.6812\,g\ H \times \frac{1\ mol}{1.01\,g\ H} = 1.6645\ mol\ H$

$3.3217\,g\ O \times \frac{1\ mol\ O}{15.999\,g} = 0.20762\ mol\ O$

$\frac{0.8323}{0.20762} = 4\ C \quad \frac{1.6645}{0.20762} = 8\ H \quad \frac{0.20762}{0.20762} = 1$

$\boxed{C_4 H_8 O}$

4. (20 pts) Validation is a key concept in Minsky's financial instability hypothesis. Write an essay to explain this concept. Pay attention to how validation influences lending standards. Include at least one example of validation from the subprime lending crisis. What are the implications of validation leading up to the Great Recession for future business cycles?

The hypothesis centers around the concept that b. cycles exist in the marketplace, because of the inherent fluctuation in the expectations and actions of people. The changes in the validation for lending standards change as market prospects change.

When the economy is doing well, lending standards are loose and banks and lenders validate easily. The economy continues to do well & grow, and lending increases along w/ the expansion of credit. (because validation is easy). What started as hedge finance transforms into speculative finance, where too much money has been lent out to cover all the spending and investment.

Lastly, this transform into a ponzi finance environment where debts are paid through increasing debt. This is like the subprime lending crisis where validation and lending went over the edge because of the expanded credit.

e) Long Writing

esis. Write an essay to explain this
de at least one example of validation from
p to the Great Recession for future

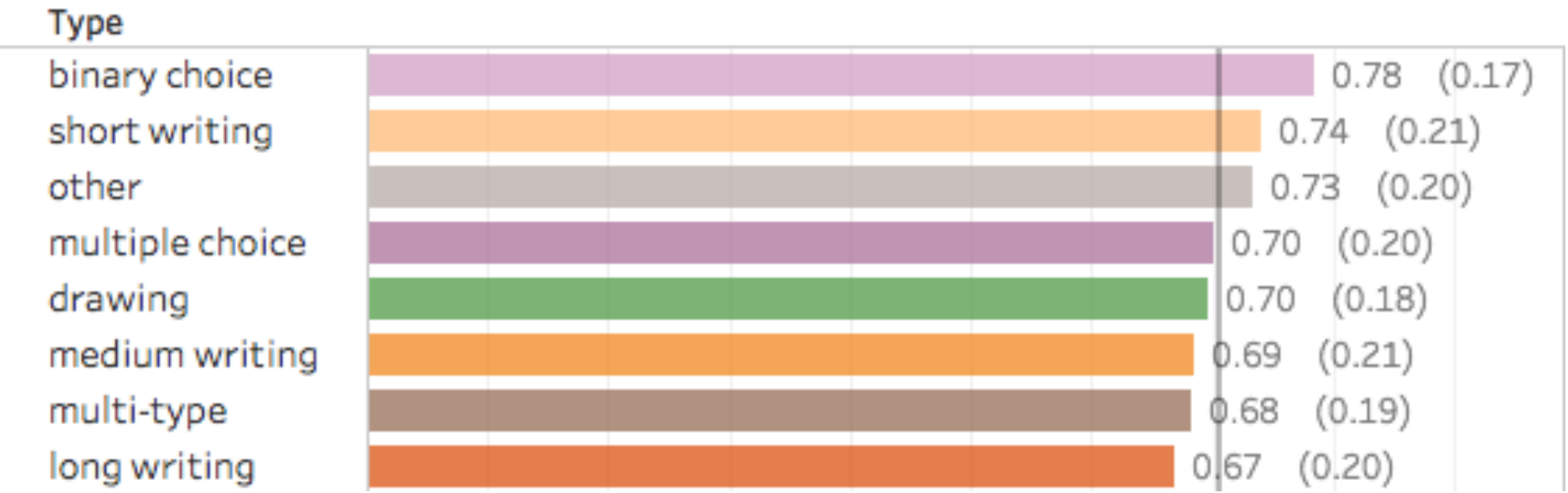
3. Draw a 7p orbital. Show all nodes and phasing for full credit (6 pts)

A). (4 pts) Use the HA-tag to purify your favorite protein. **Draw a diagram** showing the purification setup.

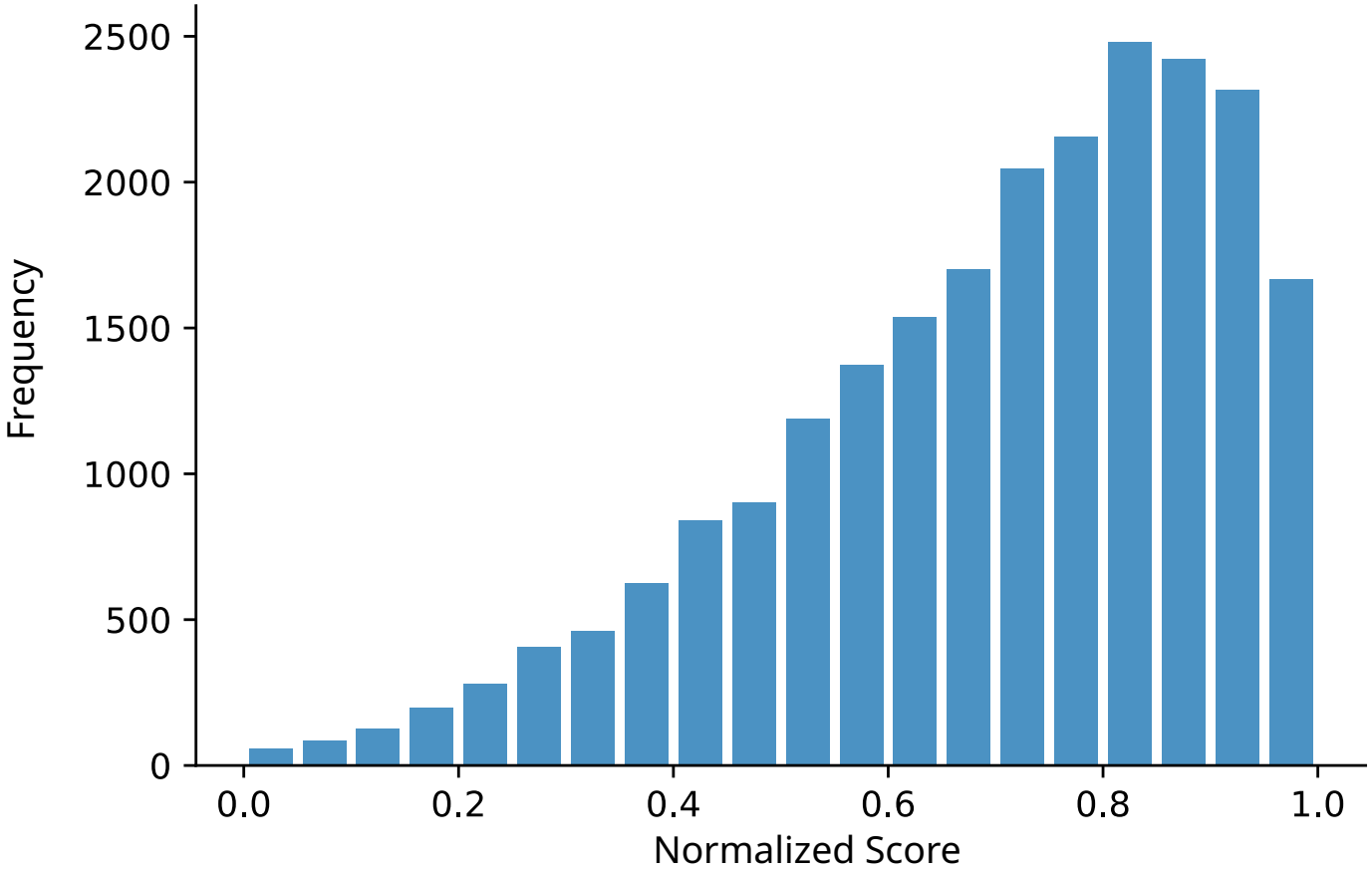
f) Drawing

How do scores differ by question type?

Mean Scores by Question Type



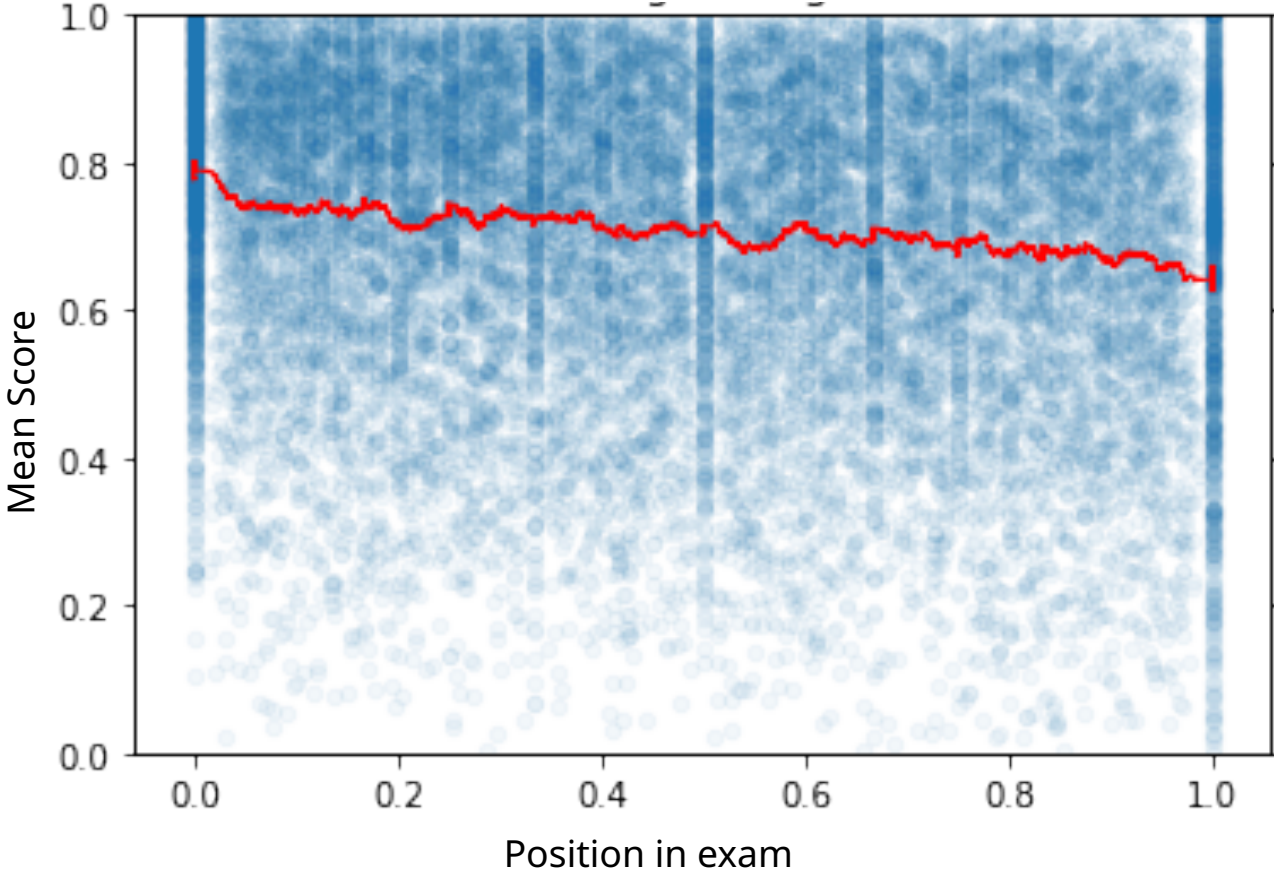
Overall Distribution of Question Mean Scores



N=22,854 questions

How do instructors sequence questions?

Question Score by Position in Exam



- Mean scores drop with position in exam.

First Half vs Second Half

- More frequently occur in **first** half of exam:
  - Binary choice (70%)
  - Multiple choice (56%)
  - Short writing (53%)
- More frequently occur in **second** half of exam:
  - Long writing (60%)
  - Medium writing (53%)

Question Scores by Position in Exam

- Mean scores drops with position in exam for all question types.

How reliable are different question types?

Question Reliability

$$\text{reliability} = \frac{\text{signal}}{\text{signal} + \text{noise}}$$

Question Reliability

Ideal: give each student an infinite number of questions

$$\text{reliability} = \frac{\text{signal}}{\text{signal} + \text{noise}} = \frac{\text{variance in student ability}}{\text{total score variance}}$$

Observed data

Question Reliability

~~Ideal: give each student an infinite number of questions~~

Estimate with a linear mixed model (Hoyt 1941, Cronbach 1951)

$$\text{reliability} = \frac{\text{signal}}{\text{signal} + \text{noise}} = \frac{\text{variance in student ability}}{\text{total score variance}}$$

Observed data

Our Model

$$score_{s,q,a,t} = G_t + K_{s,t} + A_{s,a} + u_{s,q,a,t}$$

Observed score of student s on question q in assignment a of type t

Global question type average

Student aptitude for the type

Student-assignment affinity

Student-question affinity

Our Model

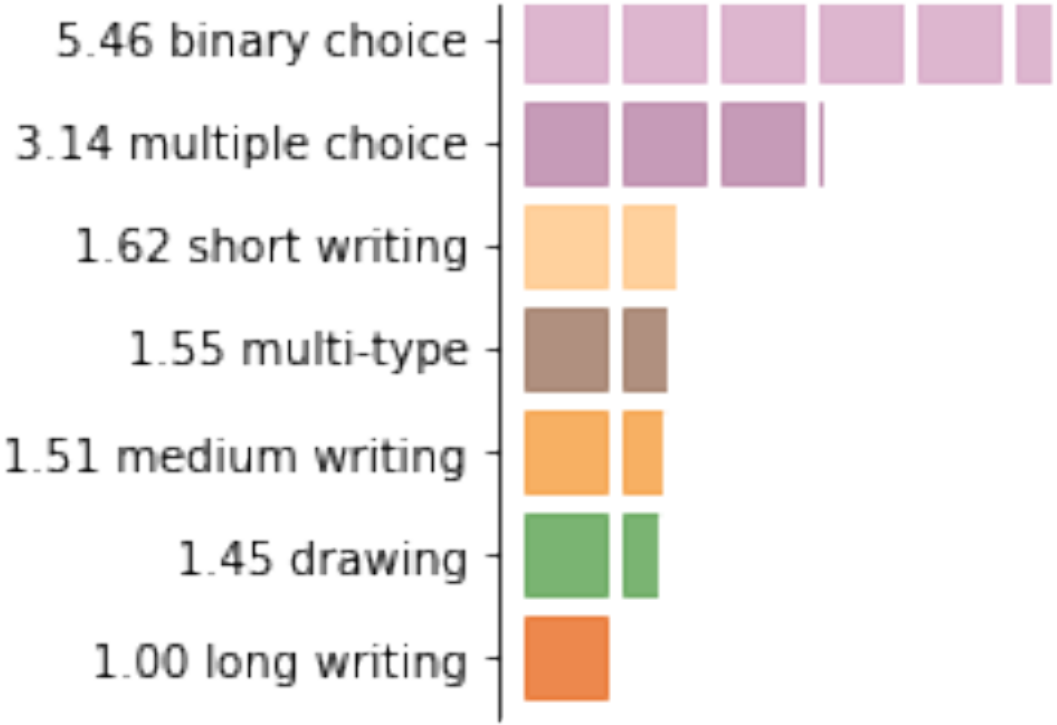
$$score_{s,q,a,t} = G_t + K_{s,t} + A_{s,a} + u_{s,q,a,t}$$

- Key assumption is that each effect is independent.
- Many factors not accounted for:
  - Students that work hard inspire others to work hard.
  - Instructors make up for hard questions/assignments with easy ones.
  - Consecutive questions cover similar material

Reliability Estimates

	Reliability	Error
Binary	0.036	0.002
Multiple Choice	0.061	0.001
Short Writing	0.112	0.001
Medium Writing	0.120	0.001
Long Writing	0.170	0.001
Drawing	0.144	0.002
Multi-type	0.117	0.001

Number of Questions Giving Equivalent Reliability



EBEL, R. L. Can teachers write good true-false test items? *Journal of Educational Measurement 12*, 1 (1975), 31–35.

To what degree are different students suited for different question types?

$$score_{s,q,a,t} = G_t + K_{s,t} + A_{s,a} + u_{s,q,a,t}$$

Observed score of student s on question q in assignment a of type t

Global question type average

Student aptitude for the type

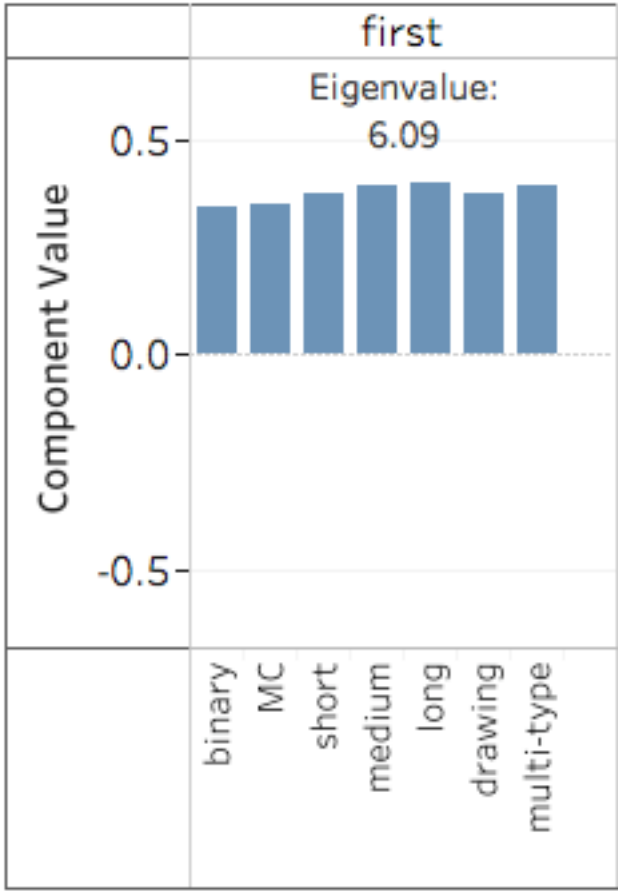
Student-assignment affinity

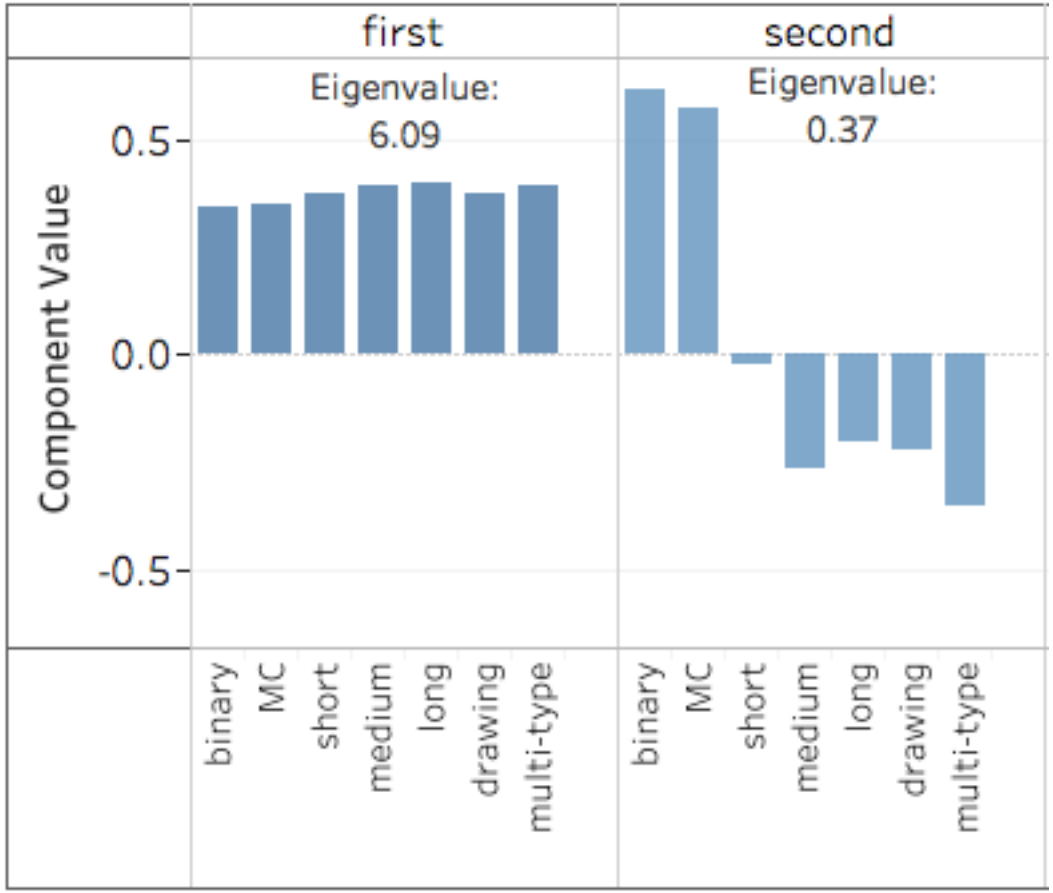
Student-question affinity

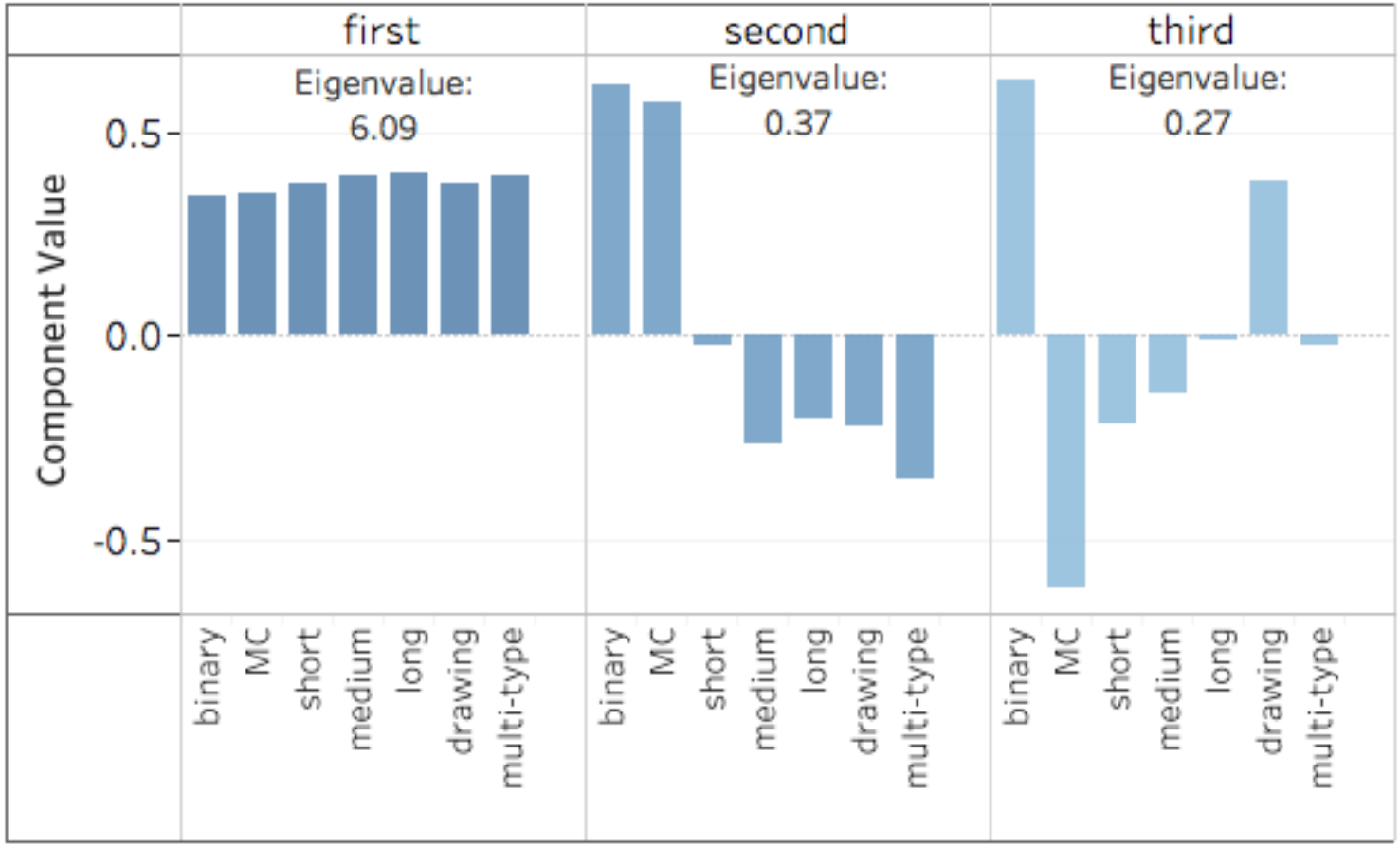
	Binary	MC	Short	t = Medium	Long	Drawing	Multi
Student A							
s = Student B				$K_{s,t}$			
Student C							
Student D							
Student E							
...							

Question Type Aptitude Correlations

	binary choice	multiple choice	short writing	medium writing	long writing	drawing	multi-type
multiple choice	0.753						
short writing	0.78	0.791					
medium writing	0.762	0.794	0.913				
long writing	0.806	0.805	0.915	0.995			
drawing	0.796	0.733	0.81	0.884	0.921		
multi-type	0.745	0.777	0.897	0.97	0.98	0.95	


first
Eigenvalue:
6.09
Component Value
0.5
0.0
-0.5
binary
MC
short
medium
long
drawing
multi-type


first
second
Eigenvalue:
6.09
Eigenvalue:
0.37
Component Value
0.5
0.0
-0.5
binary
MC
short
medium
long
drawing
multi-type
binary
MC
short
medium
long
drawing
multi-type


first
second
third
Eigenvalue:
6.09
Eigenvalue:
0.37
Eigenvalue:
0.27
Component Value
0.5
0.0
-0.5
binary
MC
short
medium
long
drawing
multi-type

Summary

- Revealed some common patterns in university STEM exams today
  - Start has binary and multi-MC, ending has open response
  - Mean score drops with position in exam
- Found significant differences in question type reliability
- Found that student aptitude for question types is largely correlated
  - But: binary/multiple choice questions are less correlated with others
- More interesting findings in the paper!

Thank You!

Let's Collaborate:

sergey@gradescope.com