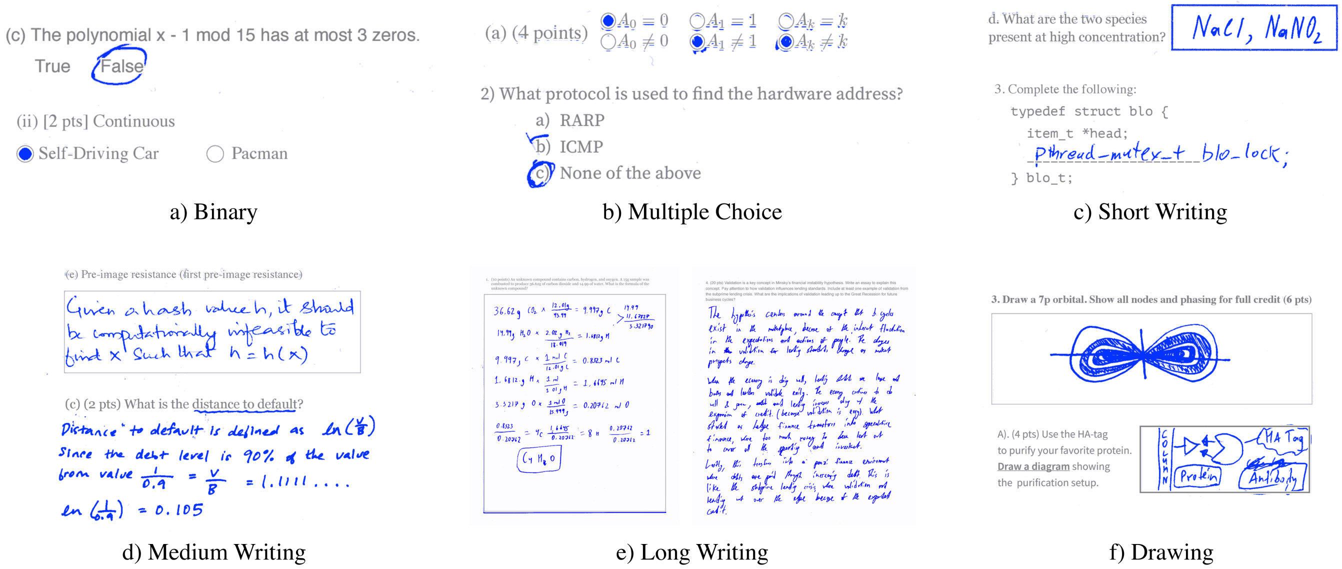
How Do Professors Format Exams?
An Analysis of Question Variety at Scale

Paul Laskowski
UC Berkeley
Berkeley, CA, USA
paul@ischool.berkeley.edu

Sergey Karayev
Gradescope
Berkeley, CA, USA
sergeyk@gradescope.com

Marti A. Hearst
UC Berkeley
Berkeley, CA, USA
hearst@berkeley.edu



(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

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

Figure 1. Examples of different question types in our dataset. (Content and handwriting have been anonymized.)

ABSTRACT

This study analyzes the use of paper exams in college-level STEM courses. It leverages a unique dataset of nearly 1,800 exams, which were scanned into a web application, then processed by a team of annotators to yield a detailed snapshot of the way instructors currently structure exams. The focus of the investigation is on the variety of question formats, and how they are applied across different course topics.

The analysis divides questions according to seven top-level categories, finding significant differences among these in terms of positioning, use across subjects, and student performance. The analysis also reveals a strong tendency within the collection for instructors to order questions from easier to harder.

A linear mixed effects model is used to estimate the reliability of different question types. Long writing questions stand out for their high reliability, while binary and multiple choice questions have low reliability. The model suggests that over three multiple choice questions, or over five binary questions, are required to attain the same reliability as a single long writing question.

A correlation analysis across seven response types finds that student abilities for different questions types exceed 70 percent for all pairs, although binary and multiple-choice questions stand out for having unusually low correlations with all other question types.

L@S'18 London, UK

ISBN 978-1-4503-5886-6/18/06.
DOI: http://dx.doi.org/10.1145/3231644.3231667

ACM Classification Keywords

K.3.1 Computers and Education: Computer Uses in Education

Author Keywords

learning at scale, essay questions, multiple choice questions, summative evaluations, examinations

INTRODUCTION

One of the goals of Learning at Scale is to provide new views of educational practices across a wide range of instructors, students, subjects, courses, and media. This paper approaches this goal by both describing and analyzing a unique educational dataset: the question format of nearly two thousand college-level summative examinations across a range of STEM topics, each administered to at least 100 students. This dataset is unique in that the exams are created "in the wild" by individual instructors, as opposed to standardized tests, and are presented to students on paper, as opposed to electronically, allowing a wide range of question formats.

The data was collected by an educational company, Gradescope, and includes observations of 22,854 different questions from 1780 exams, with fully anonymized score information for over 120,000 students linked across exams. Such a large sample size is unprecedented in studies of question response types.

Unlike large datasets that arise in the context of massive open online courses (MOOCs), this dataset does not require that exam questions conform to a set of computer-supported formats. Instead, instructors write paper exams in a traditional fashion and then scan and upload them to take advantage of a suite of grading tools. Features of each question were manually extracted by a team of annotators, yielding an unusually faithful snapshot of the way instructors write actual exams, free of bias introduced by the constraints of digital exam interfaces.

This dataset allows for the documentation of the format of paper exams used in universities with much more granularity than previously possible. We go beyond the standard dichotomy of multiple-choice versus essay question, distinguishing among seven top-level categories of response types. We compare the distribution of these question types across course topics, and look at how their position within an exam is related to scores.

The anonymized student scores data allows analysis of factors related to student performance. We construct a measure of reliability for each question type, based on an analysis of variance with random effects. Using this measure, we estimate the number of questions of each type needed to attain an equivalent level of reliability and find dramatic differences between binary and multiple choice questions on the one hand, and open-response questions on the other.

We extend our linear mixed model to explain student performance on different question types. This approach separately accounts for idiosyncratic noise associated with individual questions, and variation in students' underlying ability for each response type. We use this technique to estimate a correlation matrix of performance for each question type. We further break down this correlation using principal component analysis to identify situations in which student performance on some question types may diverge from performance on others. We also leverage the large sample size to estimate standard errors through bootstrapping.

RELATED WORK

According to Gronlund and Waugh's classic text [10], effective instruction includes specifying in performance terms what content students are expected to learn, creating instructional activities that support that intended learning, and planning for and using assessment procedures that are "in harmony with the intended learning outcomes and the instructional activities." Before creating an examination, an instructor has to decide what the purpose of the exam is, and what material is to be assessed. They must also decide what format the questions will appear in, and good practice dictates informing the students of the format in advance, to allow for proper preparation.

Stecklein [22] writes that a test is effective only in relation to its particular function; an exam used to determine mastery of subject matter is not necessarily the same as one used to discriminate among students. He outlines the standard qualities of good tests: tests that yield what the instructor intended, in terms of *validity* (measures the learning objectives accurately) and *reliability* (high agreement is found between multiple administrations of the test). Another common measure of a good test is one whose test items discriminate well; the *discriminative* power of a test item can be determined by comparing the proportions of high-achieving and low-achieving students who answer the question correctly, across the student population, or within an achievement subgroup (for instance, an easy question may be answered correctly by all the high-achievers but discriminate well among the lower quarter of the class). Finally, a good test can be scored without bias and is easy to administer and score. Nicol [15] relates additional, modern conceptions about the use of testing in formative evaluation for providing feedback to students as they learn.

Fixed Choice vs Open Response Types

What format the assessment should take has been debated for more than a century. Numerous studies through the years have compared fixed response (e.g., multiple choice, true-false questions) questions to open format (e.g., essays, proofs, coding) questions.

Stecklein [21] writes that fixed response-style questions were widely adopted in the 1930's to remedy what was perceived as bias in the construction and scoring of written essay questions, which in turn were introduced to replace the bias inherent in the oral examination. However, the mechanistic style of fixed response questions, paired with the perception that exams created using them are often poorly constructed, and that their answers can be guessed based on artifacts of the question style, resulted in questioning of their use starting in the 1980's [19], and in a continued debate as to the relative benefits of the different styles of question.

The arguments against fixed response item types are that they are difficult to write well, that more questions need to be constructed than for essay exams, are subject to guessing, do not allow students to exercise and express their knowledge, and only test shallow knowledge, low on the Bloom taxonomy. The arguments in favor of fixed response questions are that they are less biased than open response questions, more questions can be asked during an exam, so more material can be covered and students have more options to show their knowledge, are easy to grade, and if well-constructed, can measure the same knowledge and competencies that open response questions can, including deep concepts; and in fact, numerous studies show that well-constructed multiple choice exams test equivalent knowledge as essay exams [10, 3].

Conversely, the arguments in favor of open response questions, such as essay exams, computer coding, and mathematical proofs, are that they allow students to really show their ability, and if well constructed, force students to recall information, allow students to go deeper into their knowledge and discover new relationships as they answer the questions, and perform deeper reasoning. The arguments against are that they are often poorly written and hence too open-ended, difficult and time-consuming to grade fairly, can result in only testing shallow reasoning if not written well, and, as mentioned above, have not been proven to test different knowledge than multiple choice questions in many cases [10].

The literature comparing open response to fixed item responses is still mixed. For STEM topics, nearly every controlled study finds equivalence in knowledge tested, but some find small differences in the quality of what is being measured, [23]. For tests of writing, qualitative differences seem to be present [19]. Some studies show that matching the way students prepare to take an exam to that exam's structure can be a better predictor of outcome than the inherent benefits of the structure of the exam [20]. Rauch and Hartig [19] summarize several aspects of the recent findings and the ongoing debates.

Research on Item Ordering

The effects of the order of presentation of items within an exam have been studied on at least two dimensions: ordering in terms of difficulty, and in terms of how well the order of the content mirrors the order of presentation of the material in the course. A series of empirical studies, typically with 60-100 students as participants, have compared three orderings of test items: random, easy-to-hard, and hard-to-easy for the difficulty case, and random, start-to-end, and end-to-start in the content ordering case.

For example, Marso [14], noting that educational textbook writers instruct teachers to order their questions in particular ways (from easier to harder, to reflect the order presented in the course), conducted a controlled experiment to determine if question order affected assessment outcome. Results with educational psychology students showed that item arrangements based upon item difficulty (59 students), and order of class presentation (159 students) did not influence test score or required testing time. They did however find that students with high degrees of measured test anxiety performed less well on all arrangement orders, and so suggest that the first question or two be made easy to help relieve initial anxiety for high-stress students. Chidomere [4] found no effect of ordering in a study on marketing topics with 76 students.

Laffitte [13] reviews prior studies, noting that nearly all found no ordering effects, with a few exceptions, all in the math subject area. In a replication study with 82 introductory psychology students, he finds no effects of ordering. This study also asked students to subjectively rate the difficulty of the exams, and found no difference based on the item ordering. More recently, Pettijohn and Sacco [17] varied item ordering for both a non-comprehensive exam and a comprehensive exam for 66 psychology students. They too found no differences in outcomes based on item ordering, but they did find a difference in perceived difficulty among students.

However, Hambleton and Traub [11] did find an ordering effect on a general mathematics exam administered to 101 eleventh graders, finding that the mean number of correct questions was lower in the difficult-to-easy order, and that item order had an effect on stress generated during the test. A study by Plake et al. [18] with 170 students enrolled in educational psychology and similar majors taking placement exams for statistics courses also found ordering effects linked to anxiety. A trend in these studies seems to be that mathematics exams for which students have not been provided with course-based training in advance do show ordering effects linked to test anxiety.

THE DATA

The data was obtained from Gradescope, an educational company that provides tools to facilitate the grading of exams and homework. In a typical scenario, an instructor creates an assignment template in Word or LaTeX, has students write their answers on printed out copies, and then scans these for upload. Using a web interface, instructors specify the regions of the page that contain student responses. They are then able to grade the responses for each question in sequence, with a scoring rubric always visible and modifiable, as in Figure 2.

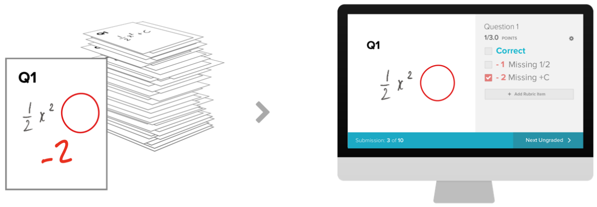


Figure 2. Instead of grading on paper, scans of the paper exams are uploaded to a web-based tool, where the grading interface focuses on the defined question area, and provides a scoring rubric.

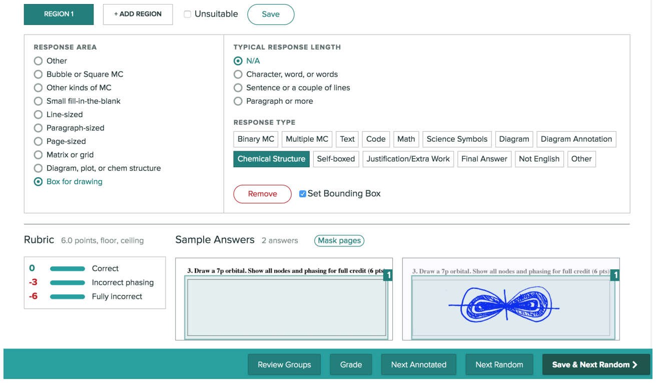
Since the application's inception in 2013, over 35M student answers have been graded by over 5,000 instructors at over 500 colleges and high schools. Notably, use is highest for computer science courses, suggesting that tech-savvy instructors are the most likely to use the system. Math, chemistry, engineering, and economics are next largest subjects. Furthermore, instructors who have adopted the system tend to be located at tier 1 universities in the United States.

Data Annotation

A subset of the raw data was annotated to identify several kinds of information about the format of the exam questions. Key variables include the following:

Response area: This refers to the type of region that holds student answers. Options include multiple choice, fill in the blank, line-sized, paragraph-sized, page-sized, matrix or grid, diagram, plot, or chemical structure, drawing, and other.

Typical Response Length: Annotators also indicated the amount of text, code, or math that the respondents filled into



□ Unsuitable

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

☑ Set Bounding Box

Rubric 6.0 points, floor, ceiling

0	Correct
-3	Incorrect phasing
-6	Fully incorrect

Sample Answers 2 answers

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

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

Figure 3. The annotation interface.

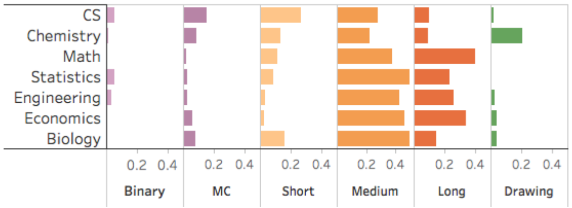


Figure 4. Frequency of question response type by topic.

these response items; for instance, an instructor might leave a paragraph size of space blank, but if the students in general wrote only a few words in response, this category was marked "Character, word, or words". Other choices were "Sentence or a couple of lines" and "Paragraph or more".

Response Type: This refines the prior two categories, including more details about the type of content written (math, code, or science symbols), and other details including if the student boxed their answer or wrote justifications and extra work. For multiple choice questions, the number of possible choices was also annotated, in order to separate questions with just two choices (*Binary*) for the main analysis. Additionally, we recorded whether exactly one choice (*1-of-N*) or multiple choices (*K-of-N*) could be selected by the student.

The annotation team went through 26,509 questions, randomly drawing from a database of exam and quiz assignments with at least 100 student answers graded between April 2013 and October 2017. Of these, 2,430 were labeled as unsuitable for annotation, either due to their not being exams (e.g., typed-out homework assignments), due to some in image processing, or because the annotator could not reasonably understand the question or answers.

Annotators went through candidate questions in the database in fully random order, with one exception: arbitrarily, they would annotate every question on the assignment. This was done to ensure that some exams were fully-annotated. The selection of such exams was left up to the annotators instead of being randomly generated, so there may be bias in which exams were fully annotated.

Refining the Dataset

Annotators were instructed to annotate each distinct answer region of a question separately. In some cases, the same question would be labeled with multiple regions. To simplify analysis of the results, such questions were grouped into a new Response Area type, called "Multi." Questions labeled as "Other" were omitted from subsequent analysis.

After initial analysis, we decided to reduce the number of question type categories by converting them to more descriptive groups as follows:

- If response area is multiple choice (MC) and has only two choices, call it "Binary MC"
- Else if response area is multiple choice, call it "MC"
- Else if response area is fill-in-the-blank, or line with a short response length, call it "Short Writing"
- Else if response area is line or paragraph and response length is short or medium, call it "Medium Writing"
- Else if response area is paragraph or page, call it "Long Writing"
- Else add plot to "Drawing" and add grid to "Other".

Figure 1 presents examples of each of these final categories.

Because formative assessment (frequently given out as homework) has fundamentally different characteristics from summative assessment (e.g., quizzes and exams), we limited the scope of the analysis to the latter by retaining only those submissions that we could verify were exams or quizzes. To compare courses according to topic, we manually categorized titles into high-level subject areas, including Computer Science, Chemistry, and Mathematics. Only seven topics were retained for which there were at least 200 annotated questions. The final dataset contains 22,854 questions from 1,780 exams, corresponding to graded answers of over 120,000 students.

Table 1 shows the resulting counts of question and exam types, and Figure 4 illustrates the frequency of question type by topic.

ANALYSIS

Question Scores

For every question in the dataset, we compute the mean normalized score, measured as the ratio of an answer's score to the max possible score on that question. The overall distribution of these mean normalized scores is shown in Figure 5. A clear ceiling effect is visible at 100%, which suggests follow-up work on replacing normal distribution assumptions in educational models with alternatives that can represent the observed score cut-off.

The mean score varies considerably among different question types, as seen in Figure 6. Binary choice questions score the highest on average, followed by short writing and multiple choice, all of which are above the overall mean. Among the question types that involve writing, mean score decreases with the space allotted, from short writing to medium to long. Drawing questions also stand out for particularly low scores.

Question scores by topic area show Engineering as the highest scoring, Statistics as the lowest, and Computer Science and Math in the middle at just above the mean score of 0.70.

		Questions									
Subject	Exams	Binary	MC	Short	Medium	Long	Drawing	Multi	Other	Total	
Computer Science	862	697	2,050	3,523	3,752	1,314	229	1,268	438	13,271	58%
Chemistry	288	44	319	469	797	331	745	813	33	3,551	16%
Mathematics	259	20	45	227	767	807	14	128	5	2,013	9%
Statistics	126	83	37	130	774	363	3	167	3	1,560	7%
Engineering	93	30	24	32	376	231	22	148	16	879	4%
Economics	65	7	42	22	344	253	29	37	5	739	3%
Biology	31	4	45	90	279	81	24	31	3	557	2%
Unknown	22	2	5	18	58	45	4	44	3	179	1%
Physics	34	0	2	6	23	45	2	27	0	105	0%
Total	1,780	887	2,569	4,517	7,170	3,470	1,072	2,663	506	22,854	100%
		4%	11%	20%	31%	15%	5%	12%	2%	100%	

Table 1. Summary of number of exams and question items by subject and question type.

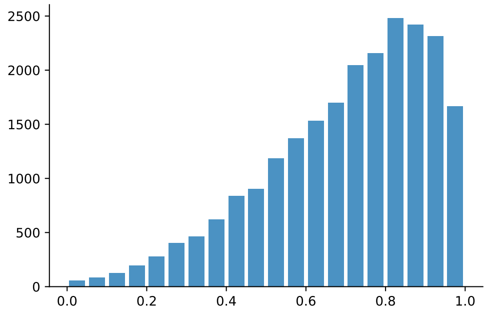


Figure 5. Distribution of mean scores for all questions.

	Count	Score Mean (std. dev.)
Binary	887	78.4% (17.4%)
1-of-N	1,791	71.2% (20.8%)
K-of-N	718	68.0% (19.0%)

Table 2. Counts and average scores for multiple choice questions.

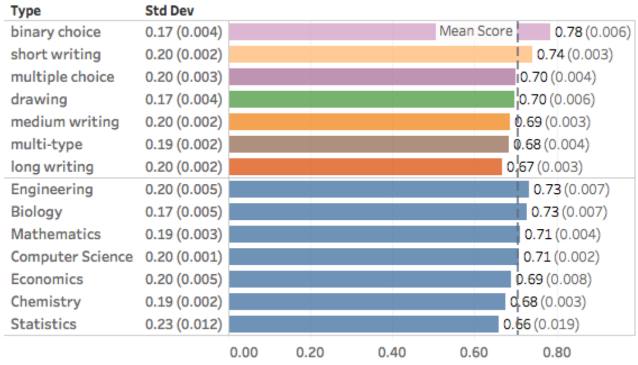


Figure 6. Mean student scores and standard deviations on exams by question type and course topic (standard errors in parentheses).

Multiple choice types

Researchers have reported on the differences in discriminability between True-False (TF), or binary choice questions and MC item types. For instance, as far back as 1941, Cronbach [5] showed experimentally the bias of students towards choosing True in a TF choice. Frisbie [9] found in one experiment that TF questions were significantly less reliable than MC items on the same subject material. But Ebel [7] found in a controlled experiment that teachers could get about the same degree of discriminability with TF questions as with multiple choice questions if they wrote five TF questions for every three MC questions.

For this dataset, Table 2 presents the detailed counts and average score mean and standard deviation for Binary, 1-of-N (exactly one choice must be selected), and K-of-N (several choices may be selected) multiple choice questions. The binary type has the highest mean score, followed by 1-of-N, and then K-of-N.

Over 95% of all multiple choice questions offered eight or fewer choices. The average number of choices for 1-of-N questions was 5.0; the average for K-of-N questions was 5.7. The number of choices was inversely correlated with the average score on the question (correlation strength of 0.18), ranging from 78.6% mean score for two choices to 66.0 mean score for eight choices.

K-of-N multiple choice questions are inherently more difficult due to the expanded answer space, reducing the likelihood that a guess will be correct, but they are also more amenable to assigning partial credit to student answers. We found that 33.0% of answers to K-of-N questions were given partial credit by instructors, in comparison to only 7.7% of answers to 1-of-N questions.

Exam Composition and Question Type Ordering

We observe that different question types tend to occur in different positions within an exam. As shown in Figure 7, binary, MC, multi-type, and short writing questions occur more frequently in the first half of the exam. Drawing and other

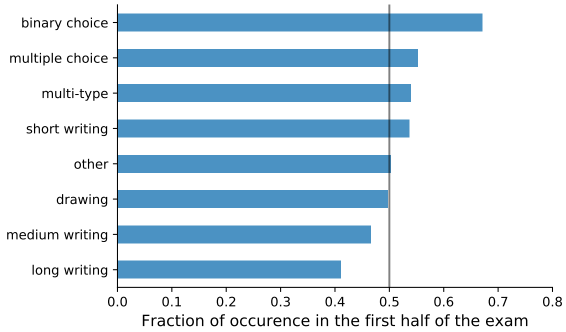


Figure 7. Frequency of position of question type varies in the first versus the second half of exams.

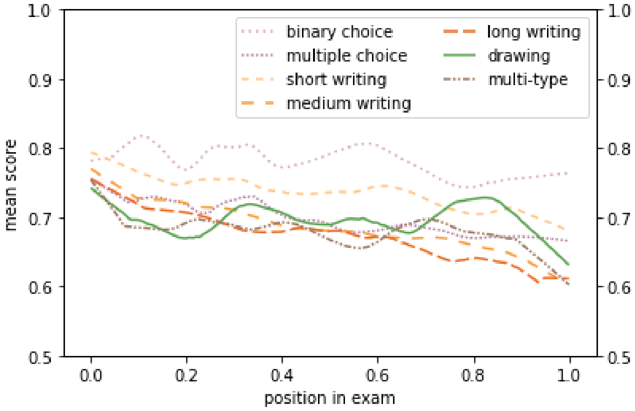


Figure 8. Lowess smoothing curves show the decline in mean score from the start of an exam to the end, broken out by the question type.

question types occur about equally frequently in the first and second halves. Medium and long writing questions occur more frequently in the second half of the exam. This suggests a common pattern for exams that use multiple types, in which fixed-response questions are followed by open-response questions.

We also observe a substantial difference in scores depending on where questions appear in an exam. Among questions that are the very first in an exam, the mean score is 0.79. Among final position questions, the corresponding mean is 0.65, 15 percentage points lower. Alternately, we can use linear regression to estimate the average decline in scores using all questions. With this method, the drop from the start to the end of an exam is 0.10 (std. err. = 0.005).

One might wonder if the decline in scores can be explained by the different mix of question types near the beginning and near the end of an exam. In fact, controlling for question type only reduces the average drop to 0.09 (std. err. = 0.005), suggesting that the decline is mainly caused by differences within each question type. In figure 8, Lowess smoothing curves are used to show the decline in scores for each question type, from the start to the end of an exam. The largest declines are found for long writing (0.12), medium writing (0.12), and short writing (0.08). The smallest declines are belong to drawing (0.03) and binary choice (0.04).

As discussed in the Related Work section, the preponderance of evidence shows that all other things being equal, test outcome is not affected by order of placement of items on the exam [14, 4, 13, 17], so these results suggest that on the whole, instructors are ordering their exams roughly from easier to more difficult questions. However, other factors, such as test taker fatigue, and differences in grading of early versus later questions, cannot be ruled out.

We find differences in exam composition by subject, as shown in Figure 9, which plots all exams that have more than half of all their questions annotated. Each exam is placed on a ternary plot corresponding to its subject, where triangle vertices correspond to exams composed entirely of one of (i) Binary/MC questions, (ii) Short, Medium, or Long Writing questions, or (iii) Other (Drawing and Multi-Type) questions. Points located inside the simplex correspond to mixtures of different question types, such that the center point is an exactly even mixture.

The figure shows that Computer Science exams in the dataset are often composed of a mix of all three types, Chemistry exams have a significant fraction of Other questions (quite often, chemical diagram drawings), and the remaining subjects are dominated by Writing questions. Figure 13 provides a more explicit view of the exams according to the sequence of question types, for Math and Chemistry exams of length 10 to 20 questions for which at least half the questions have been annotated.

QUESTION RELIABILITY

Within the literature on educational testing, two primary metrics for assessing exam questions are reliability and validity. [10]. Reliability refers to whether an exam question produces stable and consistent results when applied to the same students [8]. Validity refers to whether a question measures the knowledge or ability it is designed to measure and is not measurable with our dataset.

When measuring reliability, it is important to stress that student knowledge is inherently multi-dimensional and different types of questions may measure different aspects of knowledge [16]. In the literature, reliability is often measured with a correlation coefficient. As explained by Ebel and Frisbie,

> The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained independently from members of the same group. [8]

To measure reliability, we perform an analysis of variance using the lmer package in R [1]. This statistical procedure allows us to estimate how much of the observed difference in scores actually represents differences in student ability, and how much is attributable to noise (e.g., the idiosyncratic nature of each individual question). In line with the seminal works of Hoyt and Cronbach, we represent the variation in student ability using random effects [12, 6].

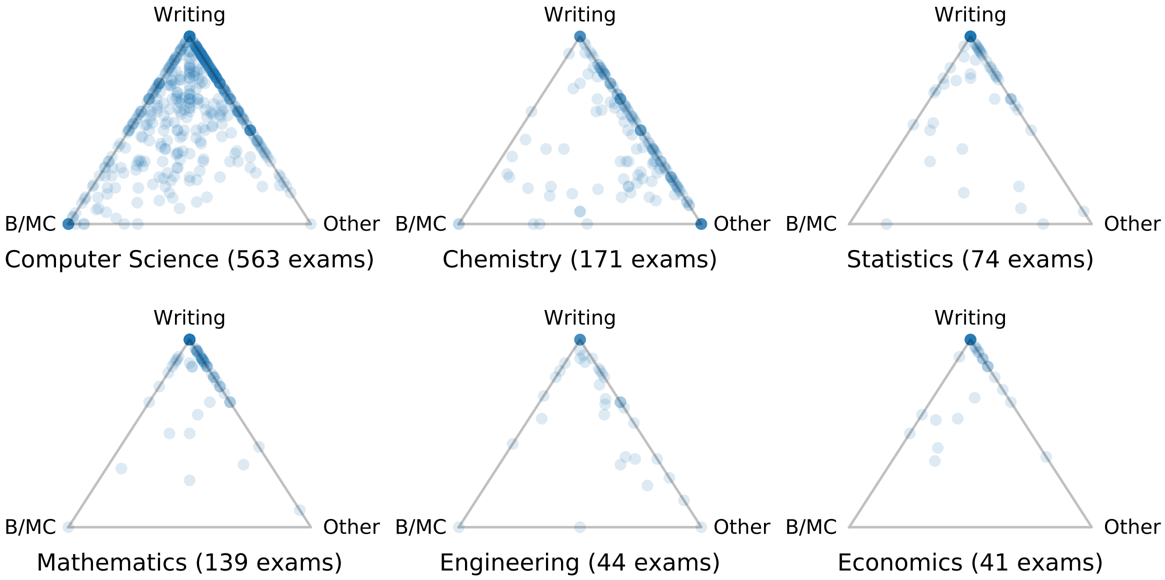


Figure 9. Exams plotted based on the mix of question types they are composed of, divided by topic.

Our analysis is based on two different model specifications. In the first, we model the score of student s on question q of type t as the sum of three terms.

$$\text{model 1: } score_{s,q,t} = G_t + K_{s,t} + u_{s,q,t}$$

Here, G_t represents the global average score for questions of type t. $K_{s,t}$ represents student knowledge, meaning student s's underlying ability to answer questions of type t correctly. Meanwhile, $u_{s,q,t}$ represents idiosyncratic noise that results from the particular combination of student s with question q. We assume that each $K_{s,t}$ is drawn independently from a mean-zero distribution with variance σ_K^2 and each $u_{s,q,t}$ is drawn independently from a mean-zero distribution with variance σ_u^2. Hence, σ_K^2 corresponds to the strength of the signal we wish to measure, while σ_u^2 corresponds to the strength of the noise.

Our second model specification is similar to the first, but we include an extra model level to represent exams. Let subscript a denote the assignment that question q belongs to. Then we represent the score for student s as follows:

$$\text{model 2: } score_{s,q,a,t} = G_t + K_{s,t} + A_{s,a} + u_{s,q,k,t}$$

The new term, $A_{s,a}$, represents the degree to which student s is specifically suited for assignment a. For example, a student may study more for some exams than others, and we may want our model to account for this possibility. Some exams may also be written in a way that is more or less tailored to a particular student. As above, we assume that $A_{s,a}$ is independently drawn from a mean-zero distribution with variance σ_A^2.

In Table 3, we report the results of both models in two ways. First we present the signal-to-noise ratio for each question type, σ_K/σ_u. This metric has the intuitive appeal of directly comparing two sources of variation. If a type of question has a lower signal-to-noise ratio, it will require more questions of that type to achieve a desired level of reliability in exam scores. Second, we present the reliability coefficient that is common in the reliability literature. Standard errors for all metrics were computed using a bootstrap procedure. [1]

	signal-to-noise ratio		reliability coef.	
	model 1	model 2	model 1	model 2
binary choice	0.25	0.19	0.057	0.036
	(0.003)	(0.005)	(0.002)	(0.002)
multiple choice	0.30	0.26	0.084	0.061
	(0.002)	(0.002)	(0.001)	(0.001)
short writing	0.40	0.36	0.138	0.112
	(0.001)	(0.002)	(0.001)	(0.001)
multi-type	0.41	0.36	0.144	0.117
	(0.001)	(0.002)	(0.001)	(0.001)
medium writing	0.42	0.37	0.151	0.120
	(0.001)	(0.002)	0.001	(0.001)
drawing	0.41	0.38	0.144	0.124
	(0.003)	(0.004)	(0.002)	(0.002)
long writing	0.49	0.45	0.194	0.170
	(0.002)	(0.002)	(0.001)	(0.001)

Table 3. Two linear mixed model specifications estimate the reliability of each question type.

Binary choice has the lowest signal-to-noise ratio, followed by multiple choice questions. For questions that involve writing, the amount of signal increases with the length of text written, from short, to medium, to long writing lengths.

As mentioned above, an instructor may compensate for an unreliable question type by writing more questions. As the

[1] A cases bootstrap, resampling at the level of students, without resampling at the level of questions [24].

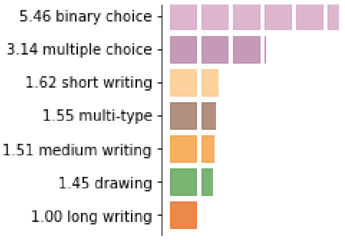


Figure 10. Based on model 2, each row shows how many questions of a particular type are needed to attain the reliability of one long writing question.

number of questions increases by n, the signal-to-noise ratio tends to increase by $\sqrt{n}$. In Figure 10, we present how many questions of each type would be needed to equal the signal-to-noise ratio of a single long written question. These results are based on model 2, though results based on model 1 follow a similar pattern.

At one end of the scale, long writing questions stand out for their unusually high reliability. It takes 45% more questions from the next type, drawing, to equal the same level of reliability. Four question types, including drawing, medium writing, multi-type, and short writing are very comparable in terms of reliability. Finally, multiple choice and especially binary choice questions stand out for especially low reliability. It takes over five binary-choice questions to equal the signal-to-noise ratio of a single long writing question. As another comparison, instructors that switch from multiple choice to any of the constructed response types would only need to create half as many questions to achieve the same reliability.

STUDENT PERFORMANCE ACROSS QUESTION TYPES

We are interested in the degree to which a student's performance on one type of question is correlated to performance on other question types. Such a correlation analysis can inform a discussion of whether one question type is an effective substitute for others. It can also reveal subgroups of students that perform well on one type of question but poorly on another.

Existing studies are commonly based on the raw correlation between scores on different question types. This is the case, for example, when a linear regression is used to predict scores on one type of question based on scores on another type [2]. Unfortunately, the idiosyncratic noise associated with individual questions will tend to bias raw correlations downward (as well as slope coefficients). Moreover, because we have data on more than two question types, bias due to noise may affect different question types to different extents.

In contrast to the standard approach, we apply a linear mixed model to account for the effects of noise. Our model is the same as described above, except we assume that a student's knowledge across types may be correlated. Specifically, we assume that each student has a knowledge vector, $(K_{i,1}, K_{i,2}, \ldots, K_{i,m})$, which is drawn from a distribution with covariance matrix C. The resulting correlation matrix estimates how much correlation there is among students' underlying ability to answer each type of question.

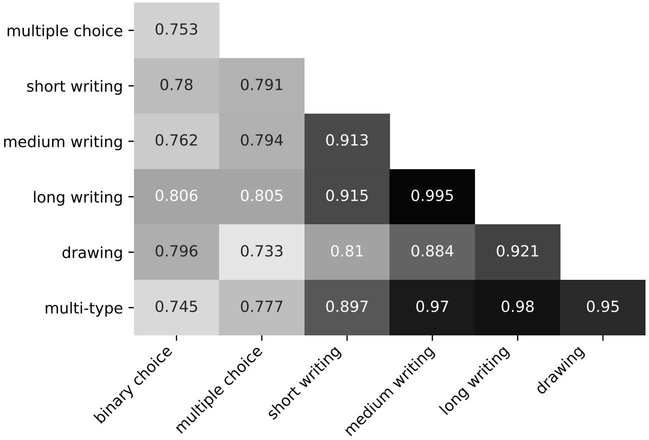


Figure 11. Estimated correlations of student ability by question type.

Student abilities for different questions types have generally high correlations, exceeding 70 percent for all pairs we examined. Some question types can be seen as more central than others, featuring high correlations with most other question types. Unsurprisingly, multi-type ability fits into this category. The writing types all have high correlations with each other. Interestingly, drawing ability is highly correlated with all writing types, especially long writing. This might reflect the highly expressive nature of these questions.

In contrast, multiple-choice (and binary-choice) questions stand out for having unusually low correlations with all other question types. Although students that are good at multiple-choice questions still tend to be good at other types of questions, the relationship is not especially strong. One stylized interpretation is that there is a substantial number of students that are good at answering multiple-choice questions, but bad at other types, and vice-versa [16]. Surprisingly, multiple-choice ability and binary-choice ability are not highly correlated, indicating that these question types may test different dimensions of knowledge, despite their similar format.

For another view of the correlation data, we perform a principal component analysis. One way to understand this technique is to imagine plotting each student as a point in a multi-dimensional space, with a separate dimension corresponding to each question type. The principal components of variation are the directions in this space along which students tend to vary the most. The eigenvalue associated with each principal component indicates the magnitude of variation in that direction. The top three principal components are graphed in Figure 12.

The first principal component simply indicates that some students are better than others for all question types, and this component is dominant with an eigenvalue of 6.09. We may think of this as distinguishing generally high-scoring students from generally low-scoring students. The next two principal components have similar eigenvalues 0.37 and 0.27 indicating similar magnitudes of variation. Component number 2 distinguishes students that are good at binary choice and multiple choice questions compared to those that are good at the more

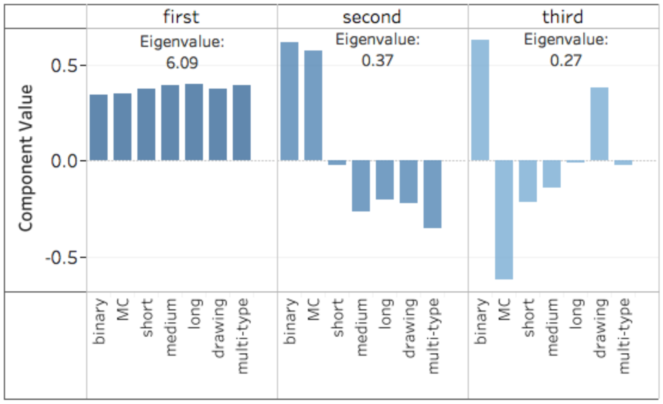


Figure 12. Top three principal components of student ability.

expressive question types. Once again, this supports the notion that a substantial number of students score well on multiple choice, but not constructed response questions, or vice-versa. Component number 3 distinguishes students that are good at multiple choice questions from those that are good at binary choice questions. This might indicate that these questions are applied to test different kinds of knowledge [3].

DISCUSSION AND CONCLUSION

At the conception of this study, we set out to answer a very basic question: how do university professors write exams today? Many of our more ambitious goals, such as helping instructors write better exams, require an accurate answer to this seemingly straightforward question. By working with an educational company, we gained a unique vantage point from which to observe the modern practice of exam writing. We are able to observe a large number of exam questions in the wild, with minimal impact to the existing practice that instructors were already following. Moreover, our annotation system provides a uniquely detailed accounting of what questions types are prevalent.

Our exploratory analysis reveals some common patterns in the way exams are ordered. While the start of an exam has a relatively high number of binary and multiple choice questions, the balance tilts towards open-ended questions by the end. One commonly heard strategy for ordering exams is to place harder questions near the end. Although the literature only provides support for this pattern in limited circumstances, we see it show up in question scores for every subject in our data. In particular, statistics exams stand out for having the greatest decline in scores from the start of an exam to the end.

We observed considerable differences in student scores across different question types. Binary choice questions garnered the highest mean score of 78%, while long writing questions fell on the other extreme with a mean score of 66%. In general, the more open-ended questions tended to result in lower scores. A closer look at multiple choice questions confirmed that scores tend to be lower for questions with more answer choices.

One place our study may benefit current practice is through our estimation of question reliability. Although other authors have applied linear mixed models to this area, we are able to apply this technique in an observational setting across seven question categories. Moreover, our large sample size allows for precise estimates, which we confirm with bootstrapping. Instructors may be particularly interested to note that it takes over three multiple choice questions (and over 5 binary questions) to equal the reliability of a single long writing question.

Finally, to the best of our knowledge, our study is the first to extend a linear mixed model so it represents student affinity for different question types. While it is common for studies to estimate the correlation between scores on multiple choice questions and one other type, interpreting the results is complicated by the fact that there is no universal benchmark for what constitutes a high or a low correlation. By measuring correlations among seven categories, we are able to compare correlations against each other. From this perspective, multiple choice and binary choice questions stand out, having unusually low correlations with all other question types. Interestingly, these variables also have low correlation with each other, suggesting that they test different types of knowledge or ability.

Further analyzing our results with a principal component analysis, we note that students vary considerably in terms of whether they are able to answer multiple choice questions well, or more expressive questions well. One possible explanation is that some students are inherently disadvantaged by the multiple choice format and would perform better if exams were converted to free response types. However, another explanation is that the types of knowledge that instructors tend to test with multiple choice questions are very different from the types that they tend to test with free-response questions. It would then be no surprise to find a low correlation between scores on these question types, even though the format is not the cause. In a related study, Thissen et al. study a set of multiple choice and essay questions that are explicitly designed to test the same knowledge. Applying a factor analysis, they argue that any differences in the knowledge tested by these question types is minor [23]. This suggests that the variation we find between selected choice questions and free-response questions is likely due to the different situations in which instructors choose to employ these questions.

Acknowledgements The authors thank Arjun Singh for discussions about this work, and Gradescope for supplying the data and annotation services for this project.

REFERENCES

1. Bates, D., Mächler, M., Bolker, B., and Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*, 1 (2015), 1–48.

2. Becker, W. E., and Johnston, C. The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record 75*, 4 (1999), 348–357.

3. Bridgeman, B., and Morgan, R. Success in college for students with discrepancies between performance on multiple-choice and essay tests. *Journal of Educational Psychology 88*, 2 (1996), 333.

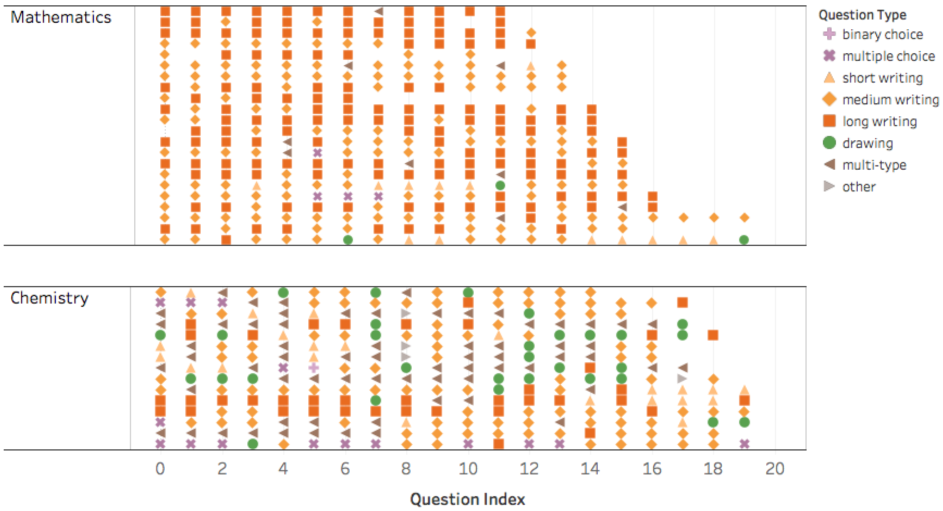


Figure 13. Questions in sequence order, color and shape coded by type, for exams in Math and Chemistry between length 10 and 20 questions.

4. Chidomere, R. C. Test item arrangement and student performance in principles of marketing examination: A replication study. *Journal of Marketing Education 11*, 3 (1989), 36–40.

5. Cronbach, L. J. An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology 32*, 7 (1941), 533.

6. Cronbach, L. J. Coefficient alpha and the internal structure of tests. *psychometrika 16*, 3 (1951), 297–334.

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

8. Frisbie, D., and Ebel, R. Essentials of educational measurement, 1991.

9. Frisbie, D. A. Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. *Journal of Educational Measurement 10*, 4 (1973), 297–304.

10. Gronlund, N. E., and Waugh, C. K. *Assessment of Student Achievement, Ninth Edition*. Pearson, 2009.

11. Hambleton, R. K., and Traub, R. E. The effects of item order on test performance and stress. *The Journal of Experimental Education 43*, 1 (1974), 40–46.

12. Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika 6*, 3 (1941), 153–160.

13. Laffitte Jr, R. G. Effects of item order on achievement test scores and students' perception of test difficulty. *Teaching of Psychology 11*, 4 (1984), 212–214.

14. Marso, R. N. Test item arrangement, testing time, and performance. *Journal of Educational Measurement 7*, 2 (1970), 113–118.

15. Nicol, D. E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and higher Education 31*, 1 (2007), 53–64.

16. Palmer, E. J., and Devitt, P. G. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? research paper. *BMC Medical Education 7*, 1 (2007), 49.

17. Pettijohn, T. F., Sacco, M. F., et al. Multiple-choice exam question order influences on student performance, completion time, and perceptions. *Journal of Instructional Psychology 34*, 3 (2007), 142–150.

18. Plake, B. S., Ansorge, C. J., Parker, C. S., and Lowry, S. R. Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance. *Journal of Educational Measurement 19*, 1 (1982), 49–57.

19. Rauch, D. P., and Hartig, J. Multiple-choice versus open-ended response formats of reading test items: A two-dimensional irt analysis. *Psychological Test and Assessment Modeling 52*, 4 (2010), 354–379.

20. Scouller, K. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education 35*, 4 (1998), 453–472.

21. Stecklein, J. Essay tests: why and how. *Bulletin on classroom testing. University of Minnesota Bureau of Institutional Research 2* (1955).

22. Stecklein, J. What is a good test? In *Classroom Test Construction*, T. Covin, Ed. MSS Information Corporation, 1974.

23. Thissen, D., Wainer, H., and Wang, X.-B. Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? an analysis of two tests. *Journal of Educational Measurement 31*, 2 (1994), 113–123.

24. Van der Leeden, R., Meijer, E., and Busing, F. M. Resampling multilevel models. In *Handbook of multilevel analysis*. Springer, 2008, pp. 401–433.