Are writing questions in math fair?

Lex Konnelly, Nathan Sanders,
Jason Siefken, and Pocholo Umbal

University of Toronto

6th Northeastern Conference on Research
in Undergraduate Mathematics Education

15 October 2022

UNIVERSITY OF
TORONTO

Overview of the presentation

Writing in Math

Types of Math Questions

Linear Algebra

Non-prose Questions

- Find a *unit* vector that is orthogonal to $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

- Draw a line $\ell \subseteq \mathbb{R}^2$ that *cannot* be expressed as a span.

- Let $X = \{\vec{a}, \vec{b}, \vec{c}\} \subseteq \mathbb{R}^3$. The statement "span$(X) = \mathbb{R}^3$" is (a) Must be true, (b) Must be false, or (c) Depends on the vectors $\vec{a}, \vec{b}, \vec{c}$.

Prose Questions

- Explain whether or not there are vectors $\vec{a}, \vec{b}, \vec{c} \in \mathbb{R}^2$ such that $\{\vec{a}, \vec{b}, \vec{c}\}$ is *linearly independent*.

- ...
  Explain using complete English sentences, whether X must be a linearly independent set. Your explanation must (i) ...

Are writing questions fair?

Main Question

Are prose questions (writing questions) fair?

- All things being equal: Maybe?
  - Explanations $\neq$ Proofs.
  - Little training on grading "writing" among mathematicians.
- All things are **not** equal:
  - International students/non-native speakers.

Context

Authors' Beliefs

Mathematical writing should be a core learning objective.

Prose/writing questions are a valuable learning/assessment tool.

Linear Algebra Context

- Large classes ($\sim$200/section, $\sim$1200/semester)
- **Many international students** ($\sim$50%)

- Partnership with the *Writing Integrated Teaching* program
  - ▶ Consultations for instructors
  - ▶ Provides funding/training for TAs

Assessing (dis)advantage

- Divide questions into **prose**, **non-prose**, and "mixed" types.
- Compute the *non-prose advantage* for each student.

Non-prose advantage

A student's average across non-prose questions *minus* their average across prose questions.

- Compare mean non-prose advantage across groups.

We interpret the *difference in* non-prose advantage across groups as a measure of fairness.

The Data

Survey

A survey asked students:

1. their use of English as a home language
2. which languages they are fluent in
3. their self-assessed proficiency in academic English writing
4. their living situation.

We combined this with

- Prose/non-prose averages from two midterms ("mixed" type questions excluded)
- Institutional data on gender & international-student status

Data

- $n = 463$ took both midterms and had usable survey data
- 56% international students
- Positive **non-prose advantage** across all students: $\sim 2.8\%$
- Similar trends across both midterms (only aggregated data is presented)

Methods

Underlying Assumption

If prose questions are *linguistically* fair, students' non-prose advantage should not be sensitive to their language background or demographics like international status.

Analysis

Exploratory $\rightarrow$ step-down regression procedure.

Step-down Regression Summary

- Start with many predictors.
- Iteratively remove the least descriptive and insignificant predictor.
- See what significant predictors remain.

Finds the "best model" as measured by the Akaike information criterion [Aka74]

Factors

Factors:

- **native English speaker**
- **multilingualism**
- **self-assessed writing proficiency**
- **living situation** (on campus, with family, or off-campus without family)
- **gender**
- **international student status**

To predict

- midterm average
- prose questions' average
- non-prose questions' average
- non-prose advantage (Our "fairness" measure)

Results

Aggregate results

|	mean %	(SD)
Overall midterm	66.9	(17.7)
Prose	64.6	(21.0)
Non-prose	67.4	(18.1)
Non-prose advantage	2.83	(15.0)

Significant factors for scores

For **midterm** average and average on **non-prose** questions, there were three significant factors out of all those tested:

- living situation
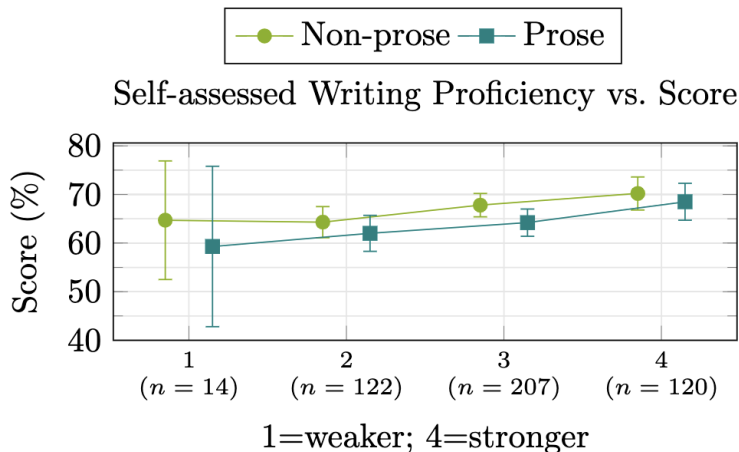- self-assessed writing proficiency
- gender

For average on **prose** questions, there were two significant factors:

- living situation
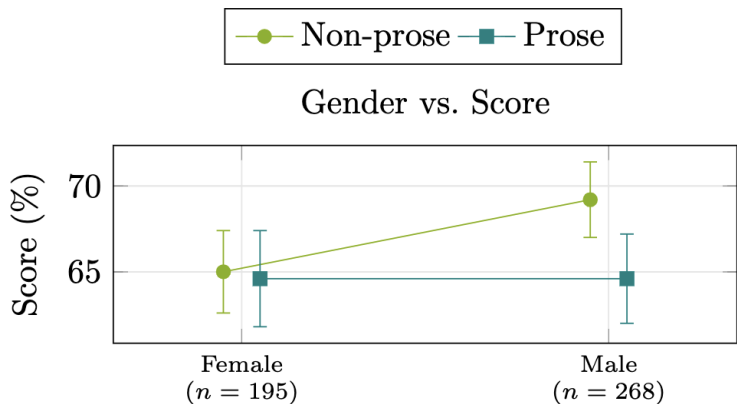- self-assessed writing proficiency

Significant factors for scores

Living Situation vs. Score

Significant factors for scores

Self-assessed Writing Proficiency vs. Score
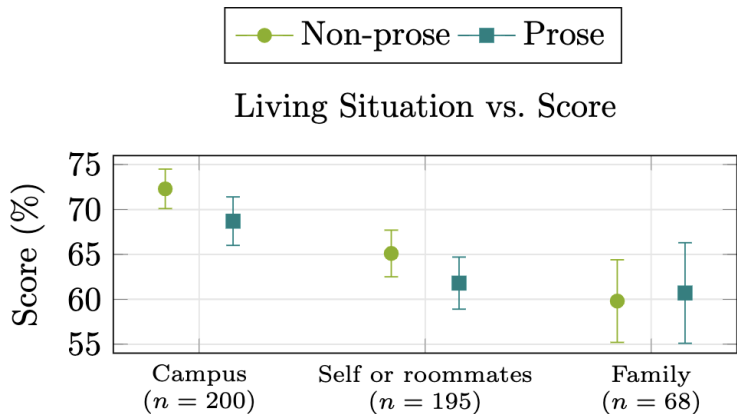
Significant factors for scores

Significant factors for non-prose advantage

The **non-prose advantage** for each group can be seen in the graphs as the distance between the non-prose and prose questions.

The higher the non-prose score (the lime green dot) is above the prose score (the dark green square), the greater the non-prose advantage.

For two of these factors (living situation and self-assessed writing proficiency), non-prose advantage is the **same across groups**.

Significant factors for non-prose advantage

Living situation: students living on campus performed the highest overall on both question types, but the non-prose advantage is about the same for those students and those living off campus alone or with roommates.

Self-assessed writing proficiency: students who rated themselves higher similarly performed higher on both question types, but the non-prose advantage did not vary by group.

Living Situation vs. Score

Self-assessed Writing Proficiency vs. Score

1=weaker; 4=stronger
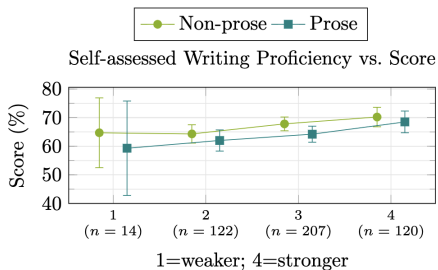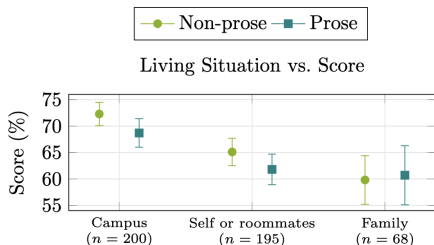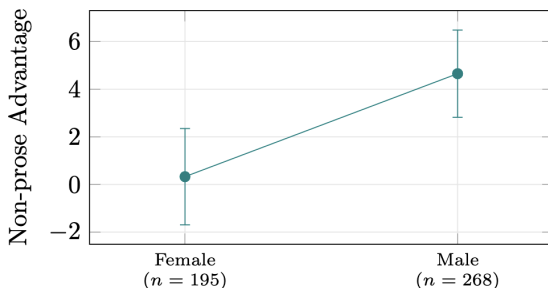
Significant factors for non-prose advantage

Only gender turns out to be a significant factor for non-prose advantage.

Gender: female students perform about the same on both types of questions, while male students perform about the same as female students on prose questions but much higher on non-prose questions.

Gender vs. Non-prose Advantage

Discussion

Language-related factors

The original motivation for the study was to see if **language background** might correlate to disproportionately lower scores on prose questions.

For example, we might expect to find a bias against non-native English speakers, with them performing disproportionately worse on prose questions, resulting in a higher non-prose advantage.

Language-related factors

Surprisingly, **none of the language-related factors** (native English proficiency, multilingualism, international status, and self-assessed writing proficiency) are significant factors for non-prose advantage.

And of these, only self-assessed writing proficiency had any correlation to a difference in scores at all, and that was in the obvious way (higher self-assessment correlates with higher scores across the board).

Language-related factors

The lack of linguistic bias between prose and non-prose questions is an important and valuable result.

It suggests that **writing questions are indeed fair**, as least linguistically. They can be asked and evaluated in a math course in ways that do not disproportionately disadvantage students in the linguistic minority.

Language-related factors

Why might this be the case?

First, the **nature of the grading** may have helped minimize biases from the graders.

- student tests were anonymized, so graders would not be influenced by student names
- answers were graded according to a rubric to guide graders to focus only on content rather than form
- many of the graders were not native English speakers themselves, so they may been less attuned to linguistics errors or more likely to overlook them

Language-related factors

Second, to be admitted to the university, incoming students who do not speak English as a native language must still demonstrate a **minimum level of fluency** by passing a standardized English test (e.g., TOEFL, IELTS, etc.).

It may be the case that, when appropriate marking rubrics are used, the university's entry requirements put non-native speakers on a level playing field with native speakers, at least for the amount of English required for these questions.

Language-related factors

However, it is possible that students are instead impacted by the **use of English throughout the course**: all course instruction was conducted in English, and all midterm questions were written in English.

Thus, students with less proficiency in English could end up with lower performance on all questions regardless of type, due to less effective learning from lectures and/or more difficulty in understanding the midterm questions.

Language-related factors

If these overall effects are strong enough, they could overshadow any effect of language ability that might be specific only to performance on prose questions.

Regardless, whatever barriers that students may face due to their language background, it seems that prose questions are viable supplements and alternatives to the non-prose questions traditionally used in math courses.

Factors with an impact

Two factors (living situation and self-assessed writing proficiency) do affect the absolute scores for both non-prose and prose questions, but they essentially do so equally, so there is no difference in non-prose advantage.

This means that there are differences between students in these groups, but those differences are not exacerbated or ameliorated by the use of prose questions.

Living situation

Why does living situation matter? Why should students living on campus perform better than other students overall?

- no commute, so they have more time to study and are more likely to attend lecture
- easier access to campus resources (office hours, writing centers, study groups, etc.)
- high cost of campus residence means on-campus students are more likely to come from families of higher socioeconomic status

More research is needed!

Self-assessed writing proficiency

Why does self-assessed writing proficiency matter?

Writing skills and math skills likely co-vary. That is, students who (think they can) write better are also better at answering math questions.

This may simply correlate to academic ability in a broader sense: students who rate themselves highly on writing may be the students who do well in all their school subjects.

As with living situation, this affects the raw scores on the two question types roughly equally, so this effect cancels out when looking at non-prose advantage.

Gender

Gender is the only factor to have an effect on non-prose advantage. So, why does gender matter?

Gender differences in STEM courses is an extensively studied subject [EB16], which is why we included it as a variable of interest.

We find that overall midterm scores are statistically significantly higher for male students (68.3%) than for female students (64.9%), which fits known patterns in much previous work.

Gender

However, when we analyze the midterm scores by question type (prose versus non-prose), we find that male and female students perform **equally on prose questions** (64.6% each).

So gender differences in overall performance on the midterm tests are due to differences in the non-prose questions only, with male students scoring an average of 69.2% compared to female students with 65.0%.

Consequently, this shows up as a difference in non-prose advantage, and interestingly, **gender was the only factor** in our models that had a statistically significant effect on non-prose advantage.

Gender

Research shows that the gender gap in math is likely not due to innate differences between genders [Fri89, Hyd14]. Instead, it seems that a more likely culprit is a difference in **self-efficacy**, that is, a student's confidence in their own abilities [LHP10, KSRP14].

So perhaps male students are less confident on prose questions than on non-prose questions, while female students are equally confident on both types of questions.

Again, more research is needed!

What do prose questions measure?

Since we are seeing some difference between question types, we might wonder what prose questions are measuring.

Two reasonable hypotheses:

(i) prose questions and non-prose questions both measure the same underlying mathematical skills

(ii) these two question types measure distinct skills

Maybe both are true! More research is needed to draw strong conclusions, but we argue that our results support hypothesis (i).

What do prose questions measure?

If hypothesis (ii) were correct, we might expect students with a stronger background in English to have a smaller non-prose advantage than those with a different language background.

However, we see no such relationship between non-prose advantage and any linguistic traits.

Meanwhile, self-assessed writing proficiency positively correlates to student performance on *both* prose and non-prose questions, suggesting that self-assessed writing proficiency may actually correspond to a student's general academic ability rather than writing ability specifically.

What do prose questions measure?

Further evidence for hypothesis (i) comes from comparing the two midterms to each other.

If the skills for each question type were different, we might expect to see differences in improvement over time, since different skills may develop at different rates.

However, our results are consistent for both midterms separately: the same factors are significant or not, and to the same extent. Student performance did improve between the midterms, but **uniformly** across the board for all groups and both question types.

Limitations

As with most educational research, there are many possible confounding variables that may offer alternative interpretations of our data.

- Data on language status and ability is self-reported, and we have evidence of at least some confusion of the survey questions based on student responses.
- There is a selection effect related to who took the survey: about 65% of all students in the course took the survey, and those who did not take the survey scored 8% lower on their midterms overall ($p < 0.001$), indicating they are a distinct population.

Limitations

- Only data from students who took both midterms *and* completed the course was analyzed. Students who missed a midterm or dropped the course may show different results.
- Our university has somewhat unusual demographics, with a large proportion of international students as well as many domestic students raised with languages other than English in the home.

Final thoughts

Prose questions do appear to be linguistically fair, if done well.

Our results bring up questions:

- What causes gender differences in non-prose advantage?
- What explains results about living situation?
- Are there hidden populations we missed?

- How much did classroom culture/marking rubrics/etc. affect this result? Will others get the same?

Authors' Belief

Writing questions provide a rich window into student thinking and can be incorporated without fear of penalizing students based on their linguistic background.

Thank you!

References

[Aka74] Hirotugu Akaike, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control **19** (1974), no. 6, 716–723.

[EB16] Sarah L. Eddy and Sara E. Brownell, *Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines*, Physical Review Physics Education Research **12** (2016), no. 2, 020106.

[Fri89] Lynn Friedman, *Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks*, Review of Educational Research **59** (1989), no. 2, 185–213.

[Hyd14] Janet Shibley Hyde, *Gender similarities and differences*, Annual Review of Psychology **65** (2014), 373–398.

[KSRP14] M. Allison Kanny, Linda J. Sax, and Tiffani A. Riggers-Piehl, *Investigating forty years of STEM research: How explanations for the gender gap have evolved over time*, Journal of Women and Minorities in Science and Engineering **20** (2014), no. 2, 127–148.

[LHP10] Sara M. Lindberg, Janet Shibley Hyde, and Jennifer L. Petersen, *New trends in gender and mathematics performance: A meta-analysis*, Psychological Bulletin **136** (2010), no. 6, 1123–1135.