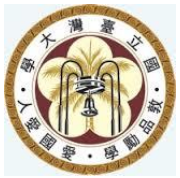
Carnegie
Mellon

Evaluating Human and Automated Generation of Distractors for Diagnostic Multiple-Choice Cloze Questions to Assess Children's Reading Comprehension

Yi-Ting Huang
Department of
Information Management
National Taiwan University

Jack Mostow
Project LISTEN
School of Computer Science
Carnegie Mellon University



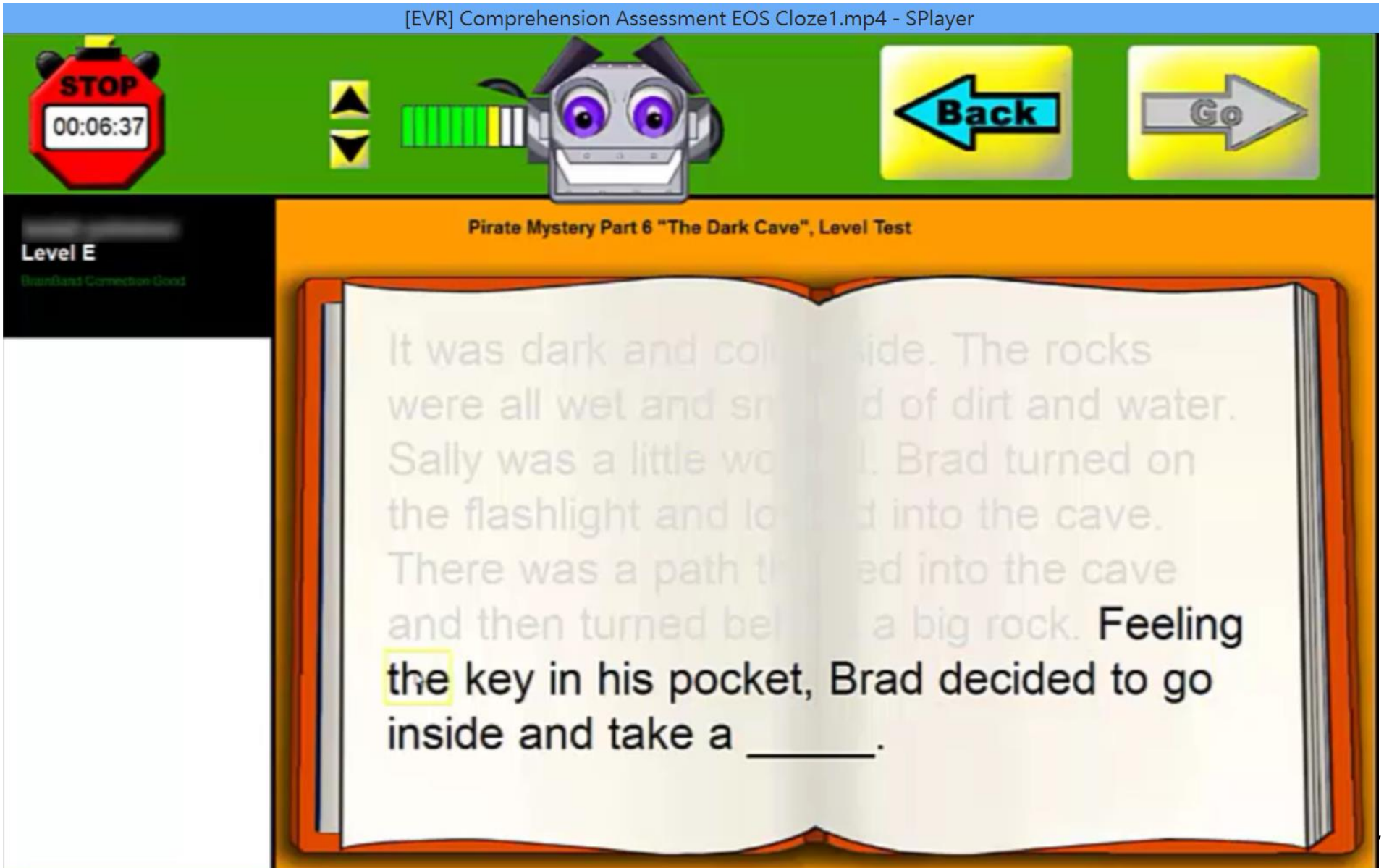
AIED 2015

Outline of talk

I. Problem
II. Experiment
III. Results
IV. Conclusion

Carnegie Mellon

DQGen



I. Problem: diagnostic assessment of children's reading comprehension

Comprehension involves multiple cognitive processes

- Syntactic: recognize grammatical structure
- Semantic: interpret meaning of phrase or sentence
- Intersentential: integrate information from previous context

DQGen (Jang '12) assesses reading comprehension

- Inserts multiple-choice cloze questions to answer while reading
- Each distractor tests a different cognitive process:
  - Ungrammatical distractor tests syntax
  - Nonsensical distractor tests semantics
  - Plausible distractor tests intersentential processing

How DQGen picks distractors: Example

Some of those cells patrol your body. They are hungry, and they eat germs! Some stop the trouble germs make. Others make antibodies. They stick to germs. That helps your body find and kill

a) are — **Type: ungrammatical**

b) intestines — **Type: nonsensical**

c) terrorists — **Type: plausible**

d) germs — **Type: correct**
Pick: original text word

Why automate? **cheap, scalable, systematic!**

This paper: how evaluate distractors?

Previous work:

- Ask experts to rate them (Liu ’05, Goto ’10, Gates ’11)
- Analyze student responses to questions (Mitkov ’06, ’09)
- Compare to human-written distractors, blind to source (Pino ’08)
- Estimate time to generate with/without system (Mitkov ’06)

This paper:

- Compare to human-written distractors for same questions
- Ask human judges to categorize choice, blind to source and type
- Measure % categorized as intended type
- Measure time to categorize or write

II. Experiment design

27 education researchers via experiment website

Categorize choices for questions 1..8

Write distractors for questions 9..16

II. Experiment design

27 education researchers via experiment website

Categorize choices for questions 1..8

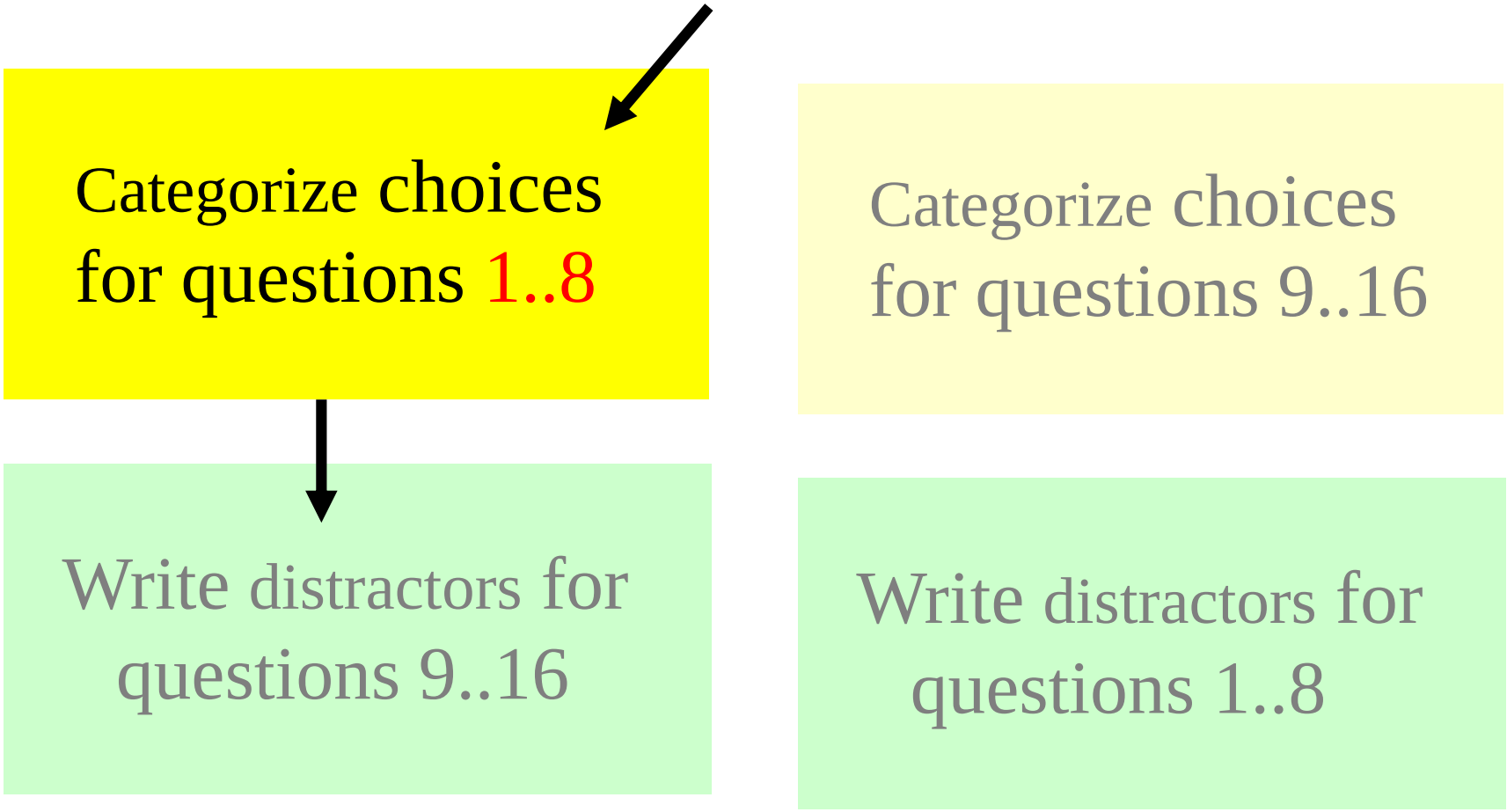
Categorize choices for questions 9..16

Write distractors for questions 9..16

Write distractors for questions 1..8

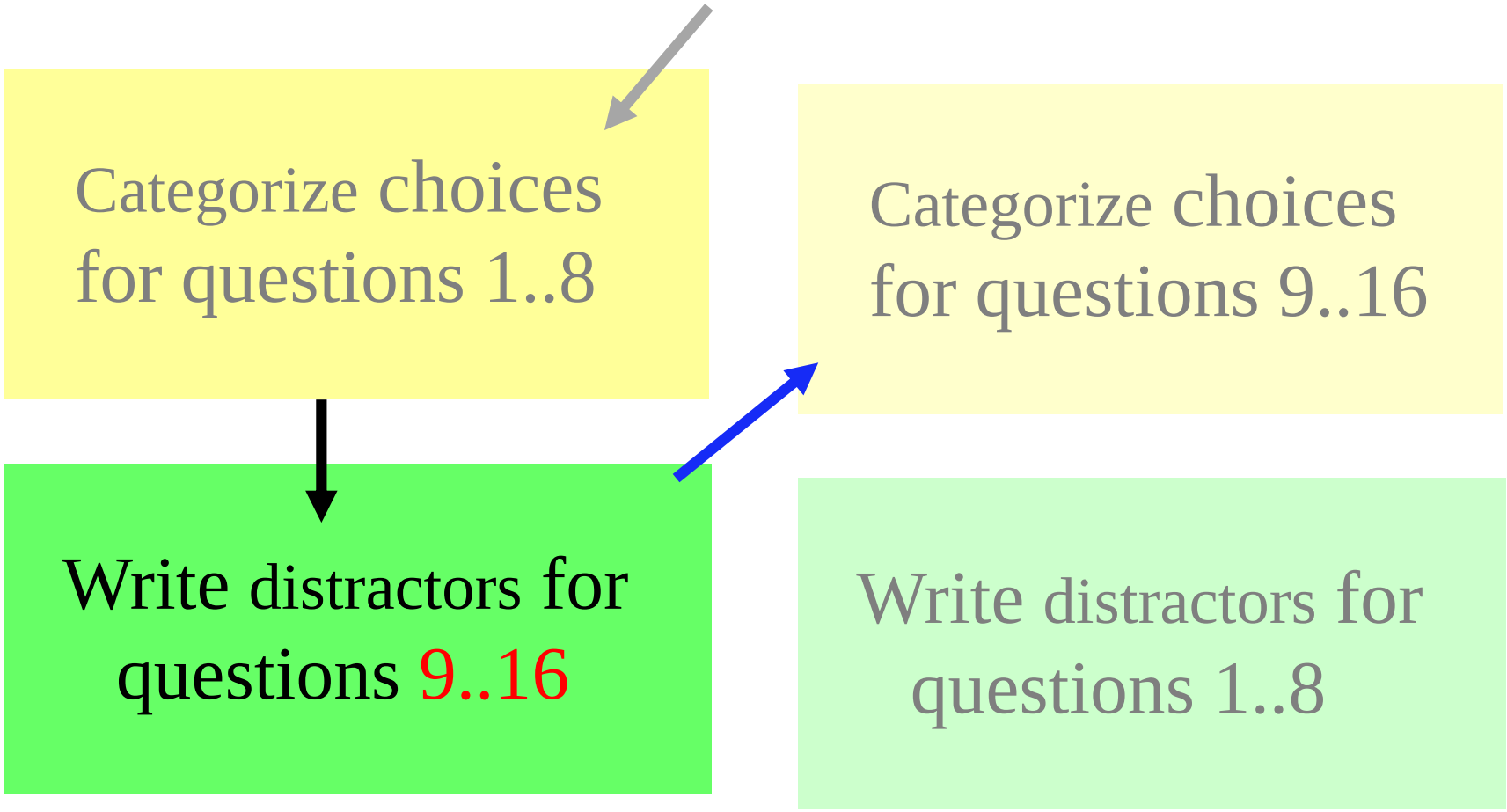
II. Experiment design

27 education researchers via experiment website



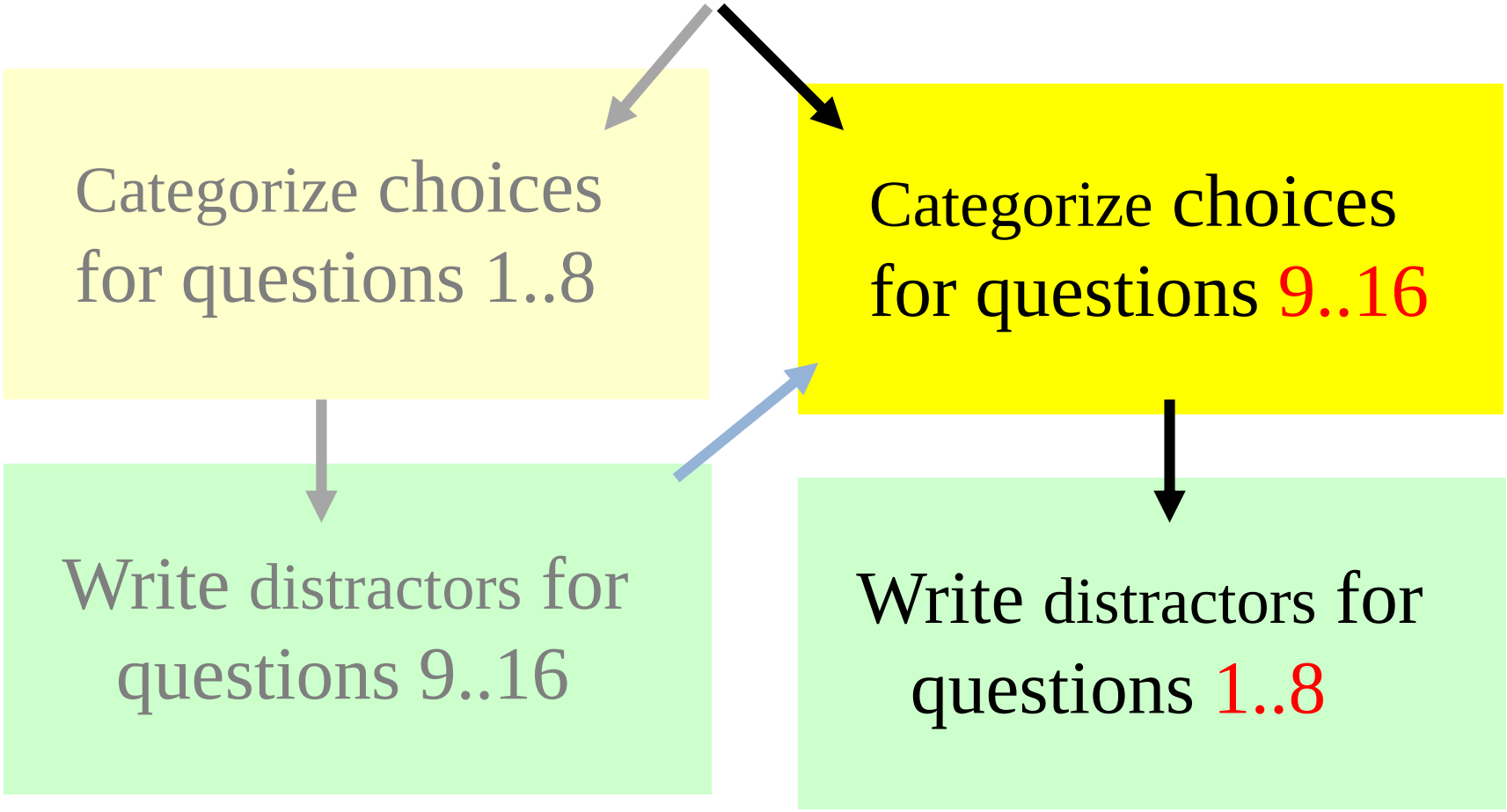
II. Experiment design

27 education researchers via experiment website



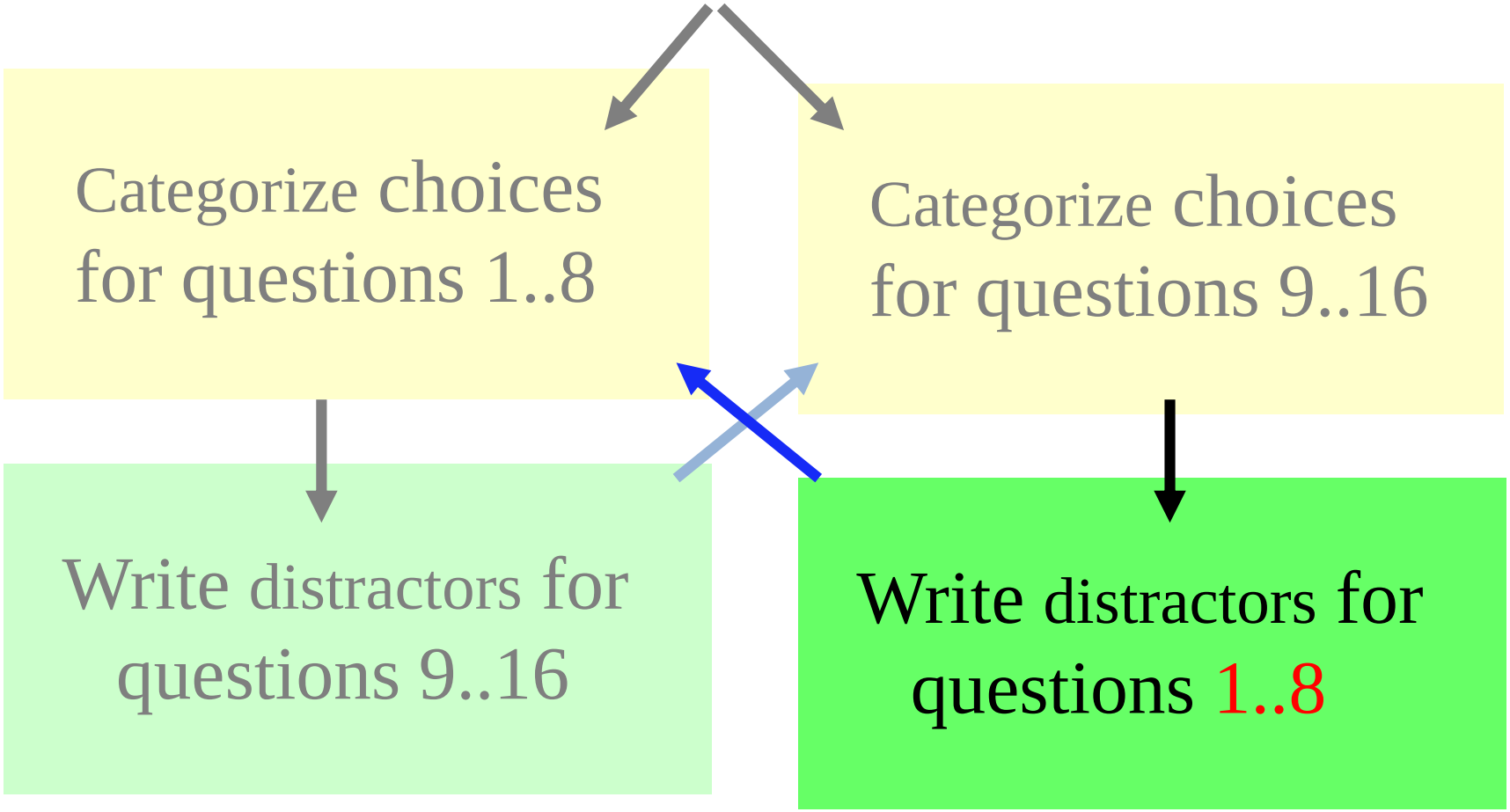
II. Experiment design

27 education researchers via experiment website



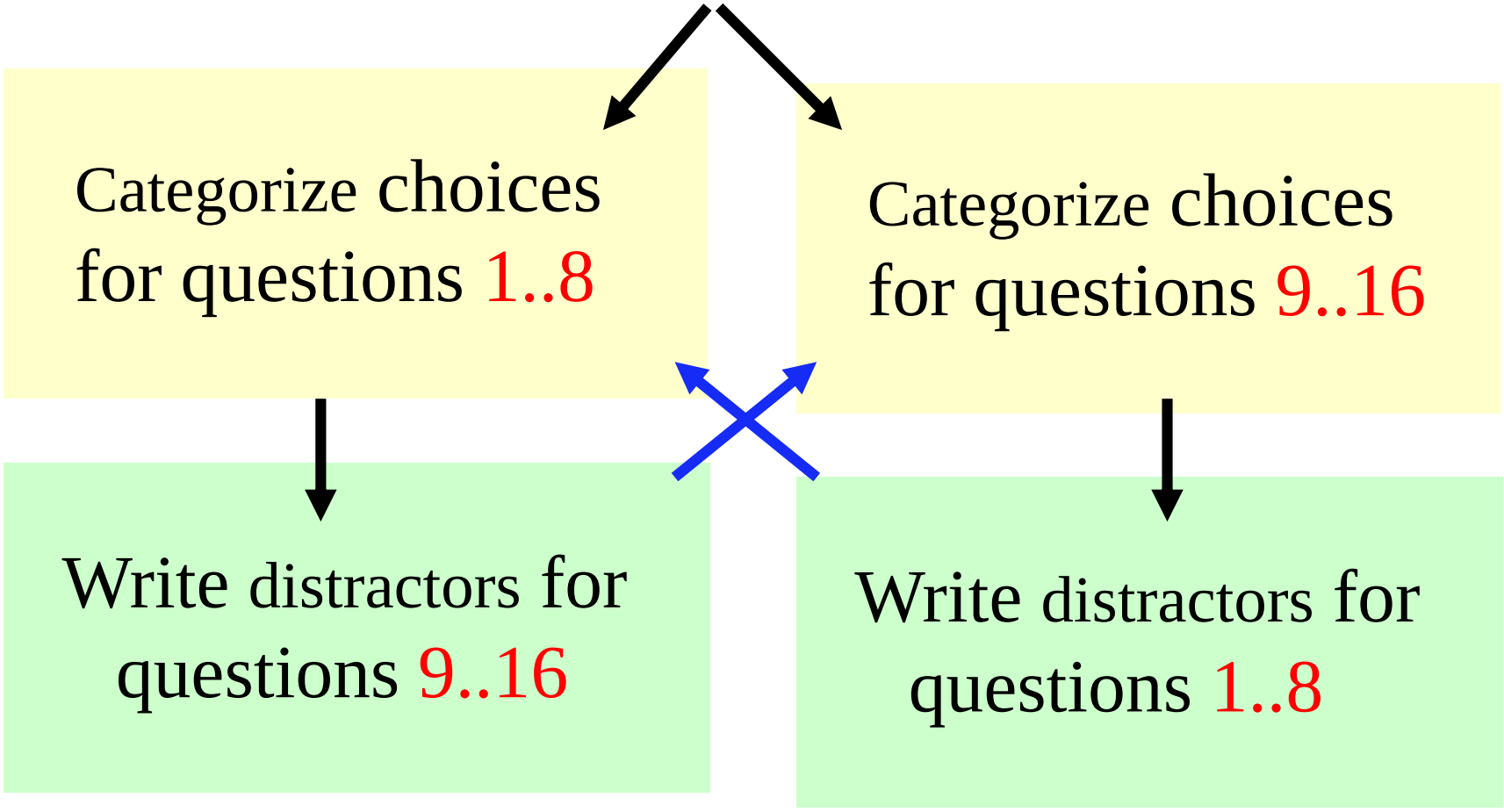
II. Experiment design

27 education researchers via experiment website

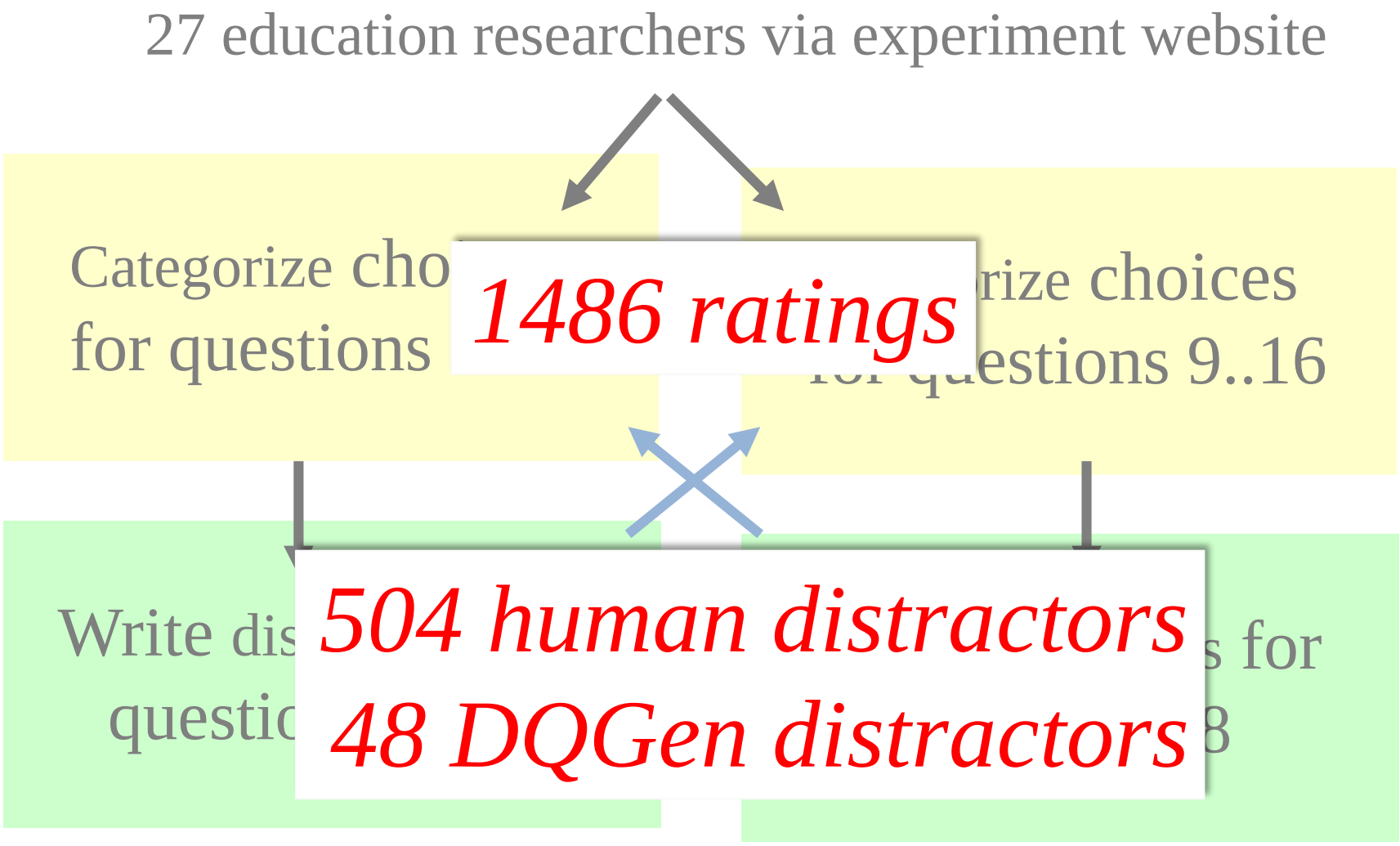


II. Experiment design

27 education researchers via experiment website



II. Experiment design



Project LISTEN

Name:

Email:

Start

Thank you for helping our study. We hope you find it fun and interesting!

– Yi-Ting Huang and Jack Mostow

Project LISTEN

Thank you for helping our research by doing two tasks: *rating* (the first task) and *designing* (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension.

In the first task, you will read 4 texts containing a total of 8 cloze items, and some different candidate completions of each item. You will classify each completion as Correct, Plausible, Nonsensical, or Ungrammatical. For example:

Some of those cells patrol your body.
They are hungry, and they eat germs!
Some stop the trouble germs make.
Others make antibodies.
They stick to germs.
That helps your body find and kill _____.

(1) are --> **Ungrammatical**
(2) intestines --> **Nonsensical** (but grammatical)
(3) terrorists --> **Plausible** (meaningful by itself but incorrect given the preceding text)
(4) germs --> **Correct**

Please classify each choice on its own merits, independently of the others.

Project LISTEN

Thank you for helping our research by doing two tasks: *rating* (the first task) and *designing* (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension.

In the first task, you will read 4 texts containing a total of 8 cloze items, and some different candidate completions of each item. You will classify each completion as Correct, Plausible, Nonsensical, or Ungrammatical. For example:

Some of those cells patrol your body.
They are hungry, and they eat germs!
Some stop the trouble germs make.
Others make antibodies.
They stick to germs.
That helps your body find and kill _____.

(1) are --> **Ungrammatical**
(2) intestines --> **Nonsensical** (but grammatical)
(3) terrorists --> **Plausible** (meaningful by itself but incorrect given the preceding text)
(4) germs --> **Correct**

Please classify each choice on its own merits, independently of the others.

Project LISTEN

Thank you for helping our research by doing two tasks: *rating* (the first task) and *designing* (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension.

In the first [illegible] you will read 4 texts containing a total of 8 cloze items, and some different candidate completions of [illegible] classify each completion as Correct, Plausible, Nonsensical, or

Thank you for helping our research by doing two tasks: *rating* (the first task) and *designing* (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension.

(1) are --> Ungrammatical
(2) intestines --> Nonsensical (but grammatical)
(3) terrorists --> Plausible (meaningful by itself but incorrect given the preceding text)
(4) germs --> Correct

Please classify each choice on its own merits, independently of the others.

Project LISTEN

Thank you for helping our research by doing two tasks: *rating* (the first task) and *designing* (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension.

In the first task, you will read 4 texts containing a total of 8 cloze items, and some different candidate completions of each item. You will classify each completion as Correct, Plausible, Nonsensical, or Ungrammatical. For example:

Some of those cells patrol your body.
They are hungry, and they eat germs!
Some stop the trouble germs make.
Others make antibodies.
They stick to germs.
That helps your body find and kill _____.

(1) are --> **Ungrammatical**
(2) intestines --> **Nonsensical** (but grammatical)
(3) terrorists --> **Plausible** (meaningful by itself but incorrect given the preceding text)
(4) germs --> **Correct**

Please classify each choice on its own merits, independently of the others.

Project LISTEN

Thank you for helping our research by doing two tasks: *rating* (the first task) and *designing* (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension.

In the first task, you will read 4 texts containing a total of 8 cloze items, and some different candidate completions of each item. You will classify each completion as Correct, Plausible, Nonsensical, or Ungrammatical. For example:

Some of those

In the first task, you will read 4 texts containing a total of 8 cloze items, and some different candidate completions of each item. You will classify each completion as Correct, Plausible, Nonsensical, or **Ungrammatical**. For example:

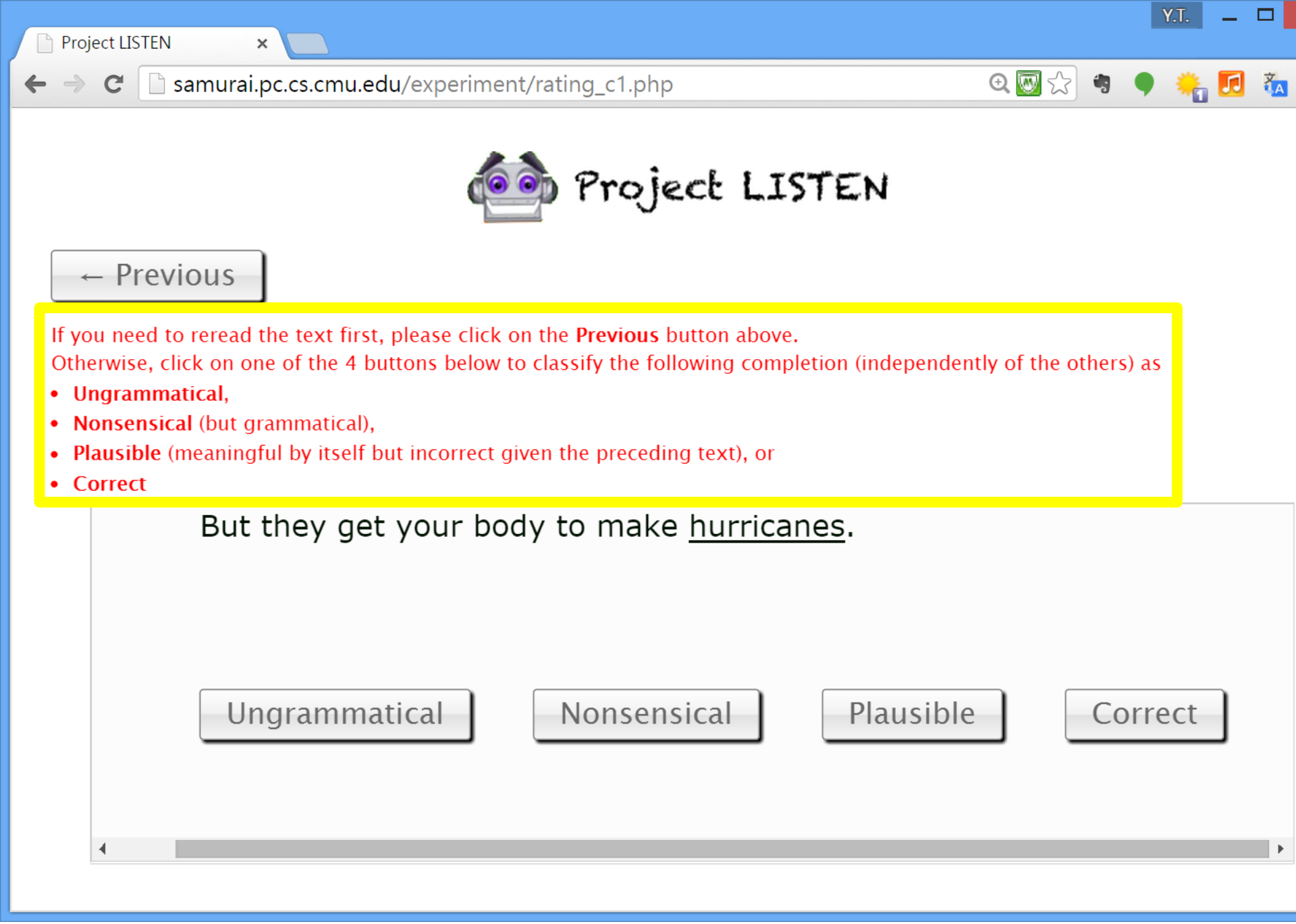
Please classify each choice on its own merits, independently of the others.

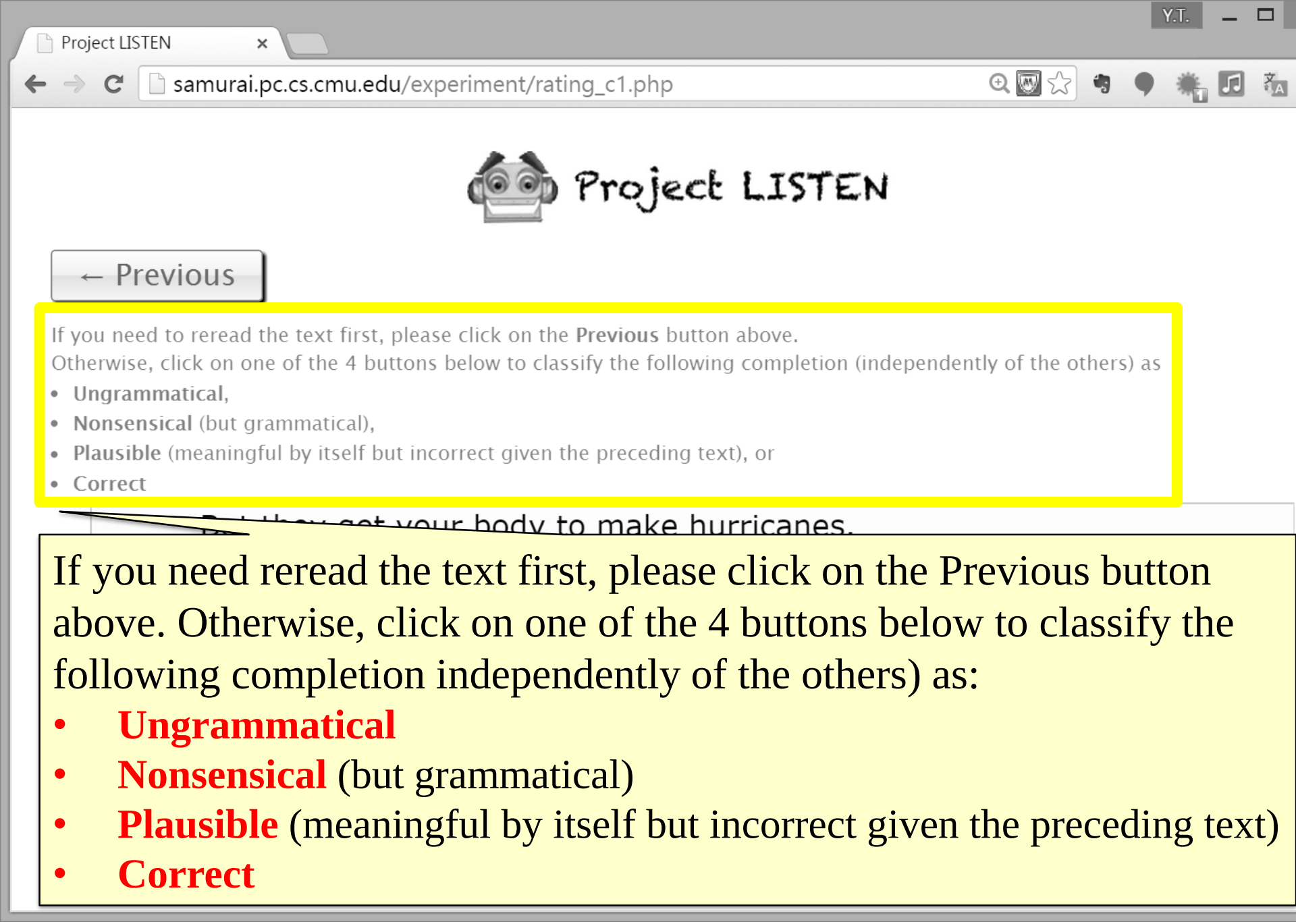
Tiny Invaders

GETTING YOUR SHOTS

Getting shots helps your immune system. Shots are filled with dead or weak germs. These germs do not make you sick.

Next


Project LISTEN
samurai.pc.cs.cmu.edu/experiment/rating_c1.php
Project LISTEN
← Previous
If you need to reread the text first, please click on the Previous button above.
Otherwise, click on one of the 4 buttons below to classify the following completion (independently of the others) as
• Ungrammatical,
• Nonsensical (but grammatical),
• Plausible (meaningful by itself but incorrect given the preceding text), or
• Correct
But they get your body to make hurricanes.
Ungrammatical
Nonsensical
Plausible
Correct



If you need reread the text first, please click on the Previous button above. Otherwise, click on one of the 4 buttons below to classify the following completion independently of the others) as:

- **Ungrammatical**
- **Nonsensical** (but grammatical)
- **Plausible** (meaningful by itself but incorrect given the preceding text)
- **Correct**

Project LISTEN

samurai.pc.cs.cmu.edu/experiment/rating_c1.php

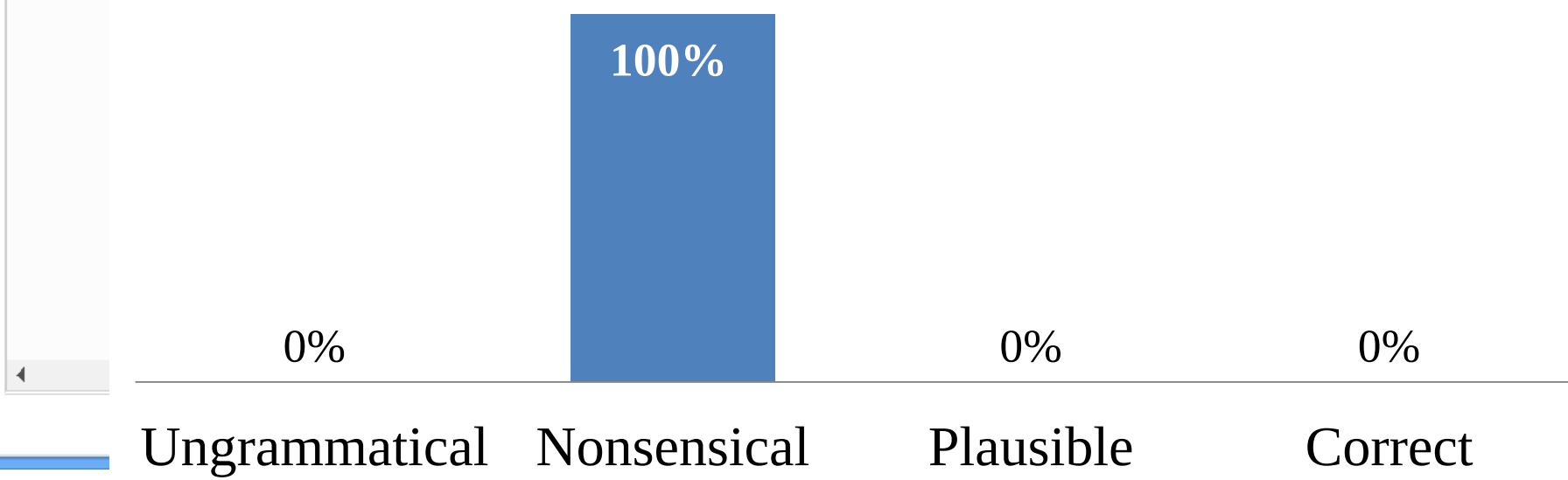
Project LISTEN

← Previous

If you need to reread the text first, please click on the **Previous** button above.
Otherwise, click on one of the 4 buttons below to classify the following completion (independently of the others) as

- **Ungrammatical**,
- **Nonsensical** (but grammatical),
- **Plausible** (meaningful by itself but incorrect given the preceding text), or
- **Correct**

But they get your body to make hurricanes.



Project LISTEN

samurai.pc.cs.cmu.edu/experiment/rating_c5.php

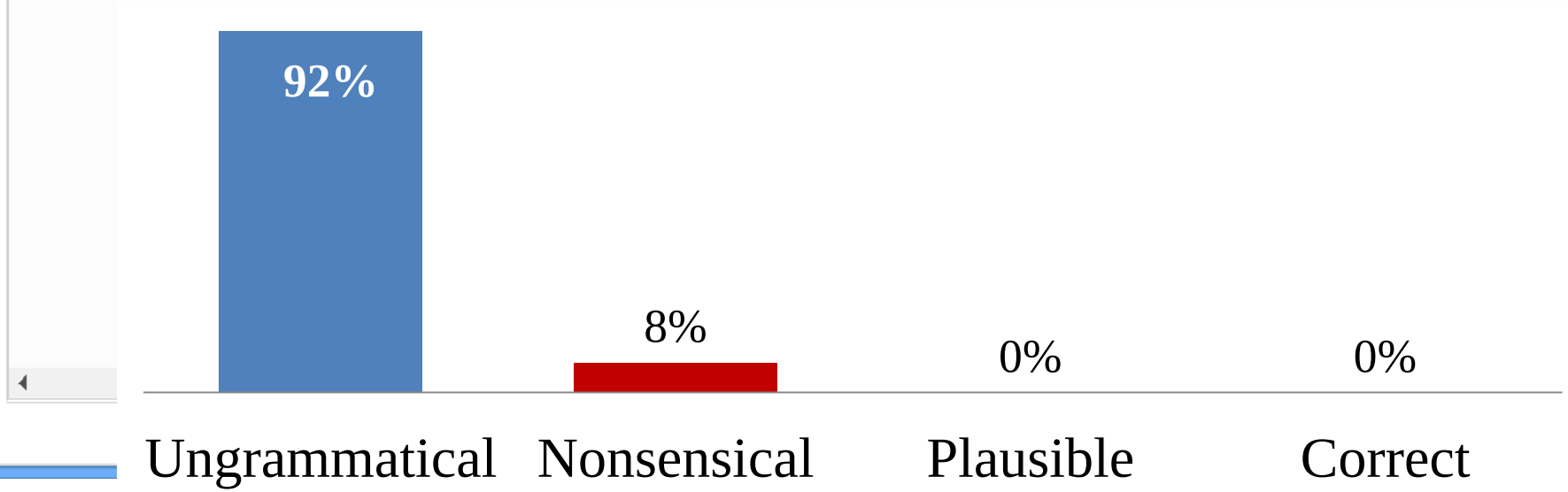
Project LISTEN

← Previous

If you need to reread the text first, please click on the **Previous** button above.
Otherwise, click on one of the 4 buttons below to classify the following completion (independently of the others) as

- **Ungrammatical**,
- **Nonsensical** (but grammatical),
- **Plausible** (meaningful by itself but incorrect given the preceding text), or
- **Correct**

But they get your body to make with.



Project LISTEN

samurai.pc.cs.cmu.edu/experiment/rating_c3.php

Project LISTEN

← Previous

If you need to reread the text first, please click on the **Previous** button above.
Otherwise, click on one of the 4 buttons below to classify the following completion (independently of the others) as

- **Ungrammatical**,
- **Nonsensical** (but grammatical),
- **Plausible** (meaningful by itself but incorrect given the preceding text), or
- **Correct**

But they get your body to make cholesterol.

Ungrammatical | Nonsensical | Plausible | Correct

Project LISTEN

samurai.pc.cs.cmu.edu/experiment/read_qid9.php

Tiny Invaders

GETTING YOUR SHOTS

Getting shots helps your immune system. Shots are filled with dead or weak germs. These germs do not make you sick.

Next

But they get your body to make cholesterol.

Project LISTEN

samurai.pc.cs.cmu.edu/experiment/rating_c3.php

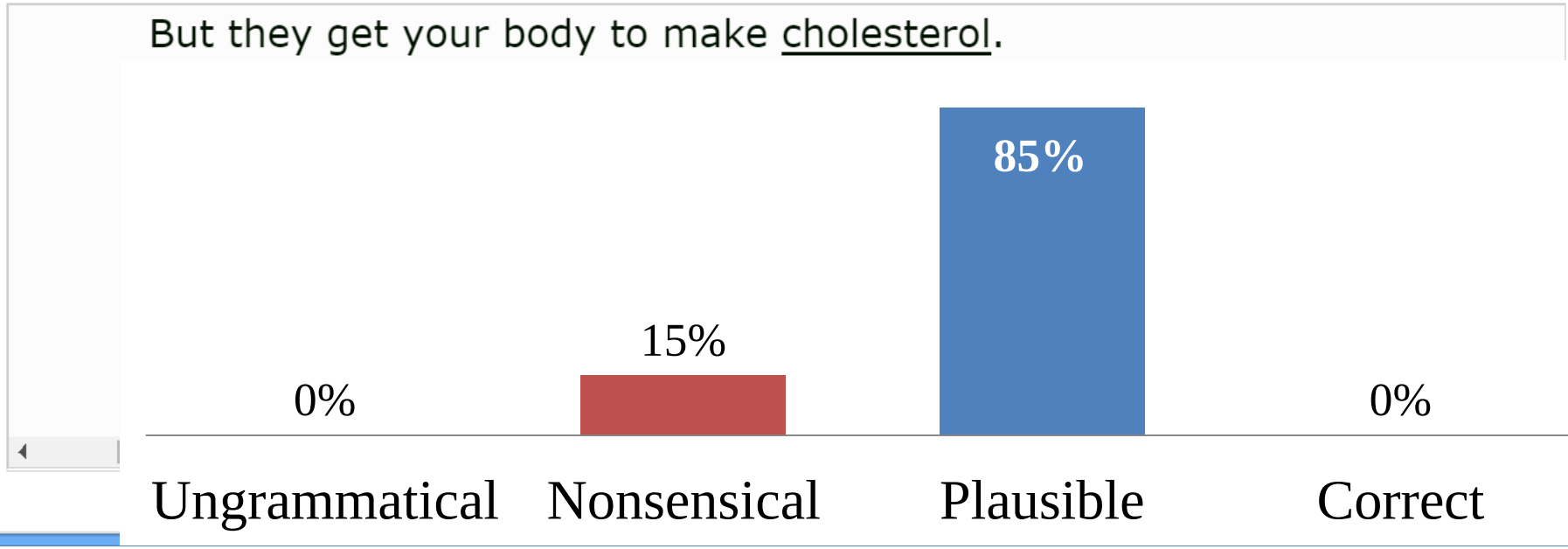
Project LISTEN

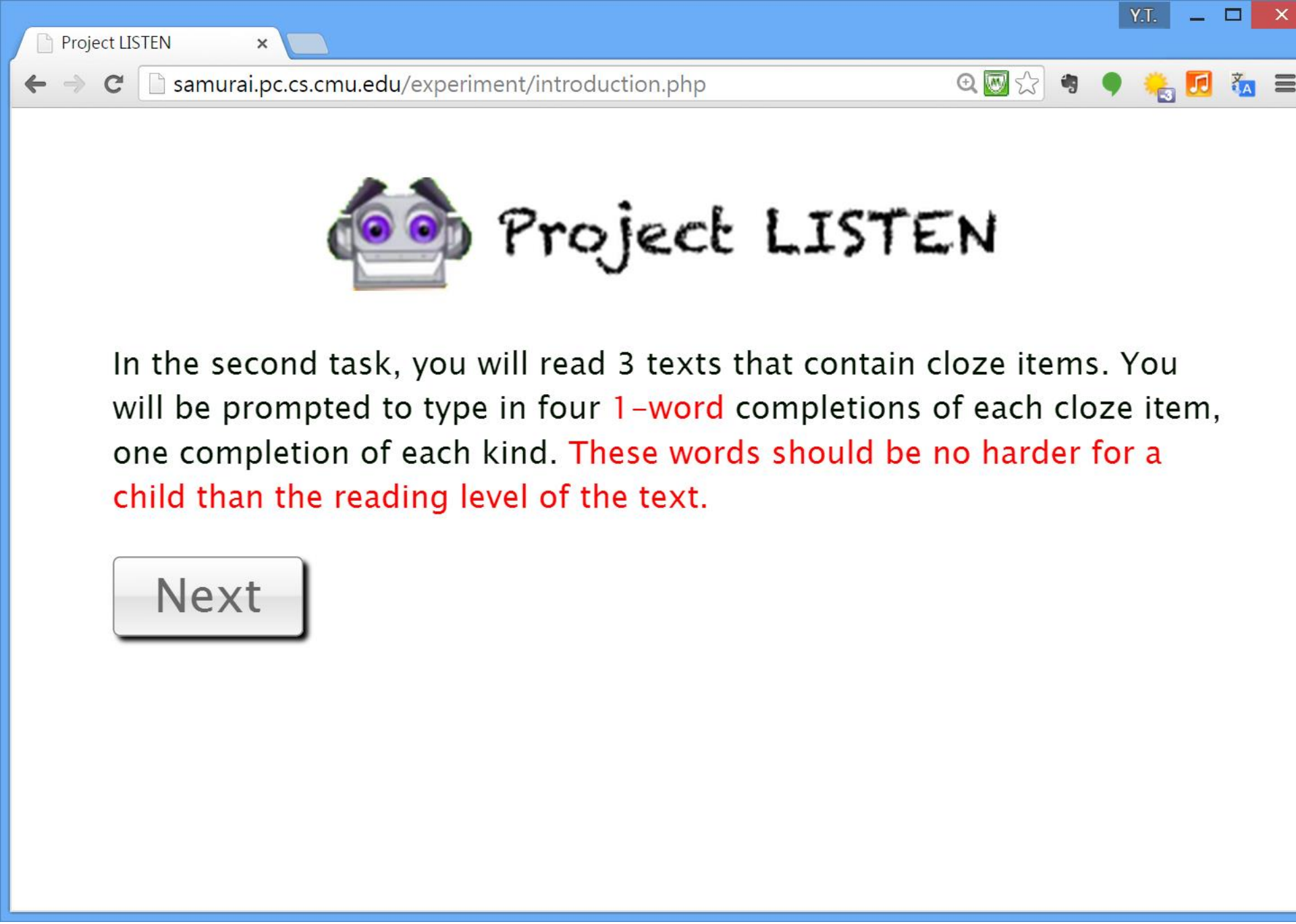
← Previous

If you need to reread the text first, please click on the **Previous** button above.
Otherwise, click on one of the 4 buttons below to classify the following completion (independently of the others) as

- **Ungrammatical**,
- **Nonsensical** (but grammatical),
- **Plausible** (meaningful by itself but incorrect given the preceding text), or
- **Correct**

But they get your body to make cholesterol.

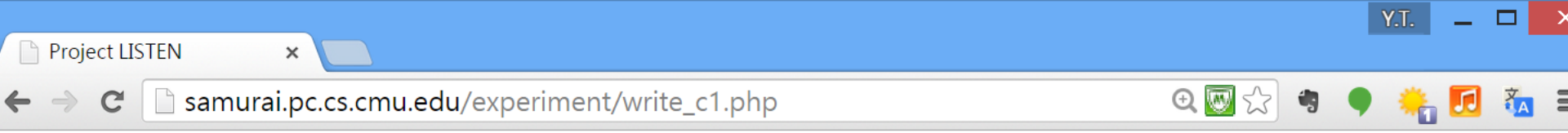



Y.T.
Project LISTEN
samurai.pc.cs.cmu.edu/experiment/introduction.php
Project LISTEN
In the second task, you will read 3 texts that contain cloze items. You will be prompted to type in four 1-word completions of each cloze item, one completion of each kind. These words should be no harder for a child than the reading level of the text.
Next

Food Groups

Then, there is the milk food group.
Milk, cheese, yogurt, and ice cream are in the milk food group.
Try to eat two to four servings of something in the milk food group every day.

Next



Project LISTEN

← previous

Ungrammatical

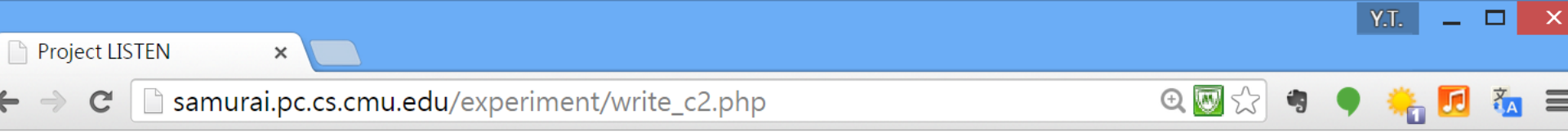
Please type in a word that makes the completed sentence **Ungrammatical.**

A serving in the milk food group means a glass of milk or a cup of

next →

Humans: after, also, angry, are, because, Down, first, happy, healthy, quickly, red, the, they, well

DQGen: eat



Project LISTEN

Nonsensical

← previous

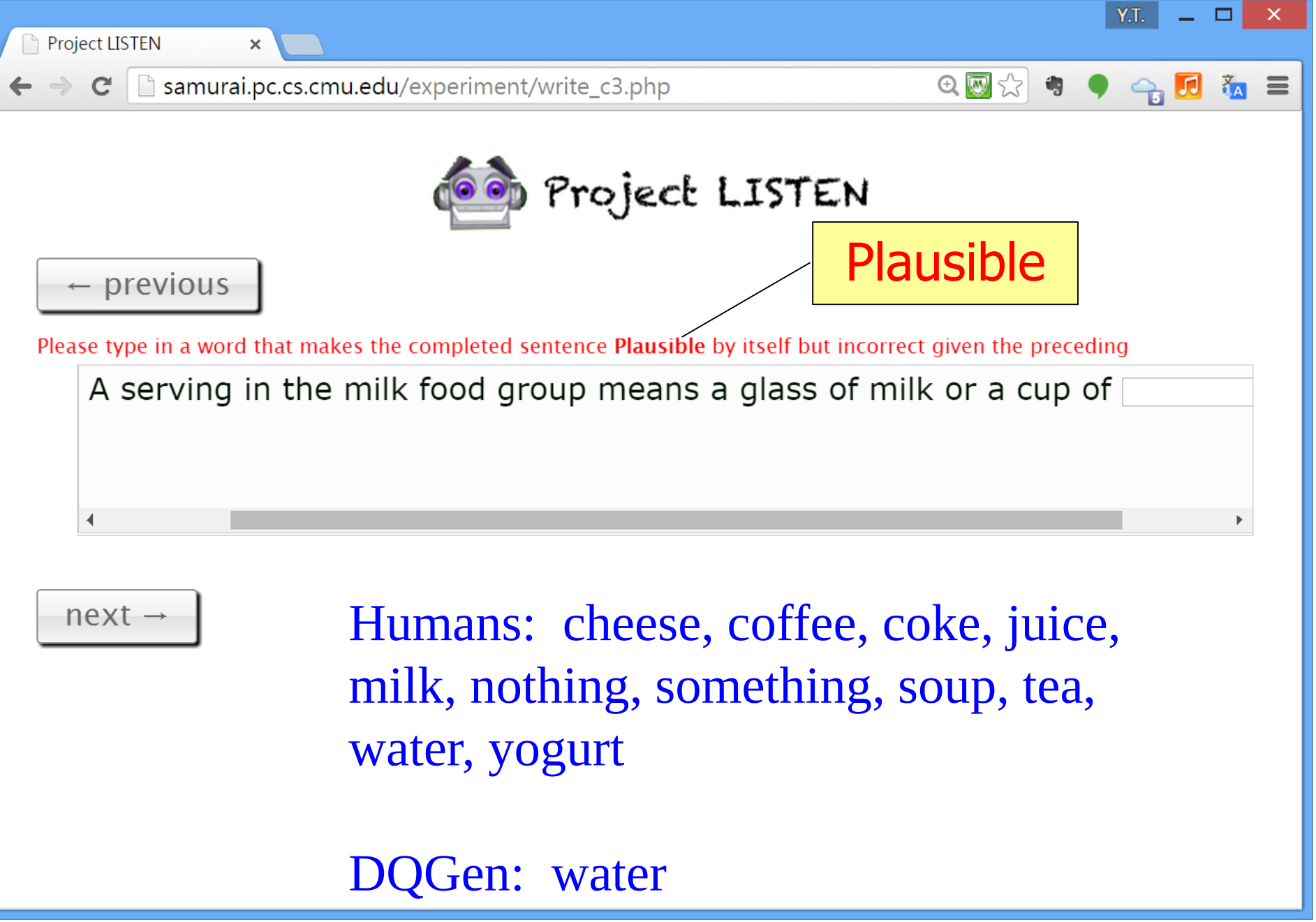
Please type in a word that makes the completed sentence **Nonsensical.**

A serving in the milk food group means a glass of milk or a cup of

next →

Humans: air, balloons, barley, cats, kids, moon, nails, serving, soda, sugar, tapioca, tree, trees

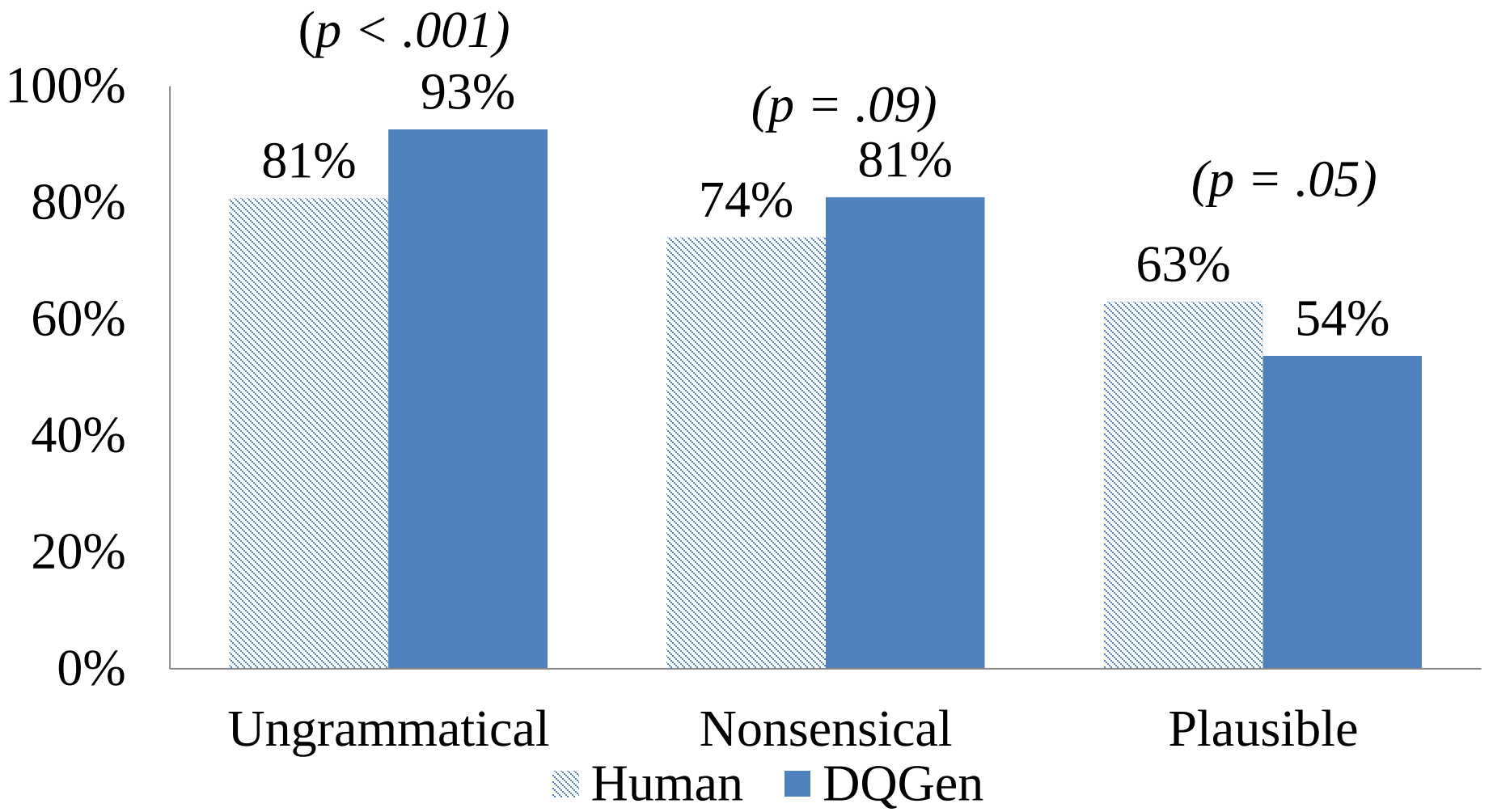
DQGen: football



Humans: cheese, coffee, coke, juice, milk, nothing, something, soup, tea, water, yogurt

DQGen: water

III. Results: % categorized as intended type



(based on 1486 ratings by 27 judges of 16 correct answers, 48 DQGen distractors, and 504 human distractors; statistical details in paper)

Why performance was lower for plausible distractor?

- **Too plausible:**
  '*our people have always mustered the determination to construct from these crises the pillars of our democracy.' [history]*

- **Not plausible enough:**
  '*It took six great big strong guys to load it all into the computer.' [truck]*

III. Results: potential speedup

Timestamped logs show time to categorize or write

	Time to generate	% categorized as intended
DQGen alone	Negligible	76%
Human alone	19 sec	73%
DQGen + (perfect) vetting + (imperfect) rewrite	10 sec = 5 sec to vet + 19 sec when rewrite	92%

DQGen + human vetting would be faster <u>and</u> better!

IV. Conclusion

Evaluation methodology more controlled

- Compared distractors for same questions
- Evaluated 3 types: ungrammatical, nonsensical, plausible
- Measured % of categorizations that match intended type
- Measured time to categorize vs. write

Compared DQGen vs. human distractors

- DQGen > Human: ungrammatical, nonsensical
- DQGen < Human: plausible

Potential payoff of DQGen + human

- Cut human authoring time in half
- Increase from 73% to 92% of distractors matching intended type

Questions?

My main question after this talk is _____ .