Quality Control - part 2

Crowdsourcing and Human Computation

Instructor: Chris Callison-Burch

Website: crowdsourcing-class.org

Different Mechanisms for Quality Control

- Aggregation and redundancy
- Embedded gold standard data
- **Economic incentives**
- Reputation systems
- Statistical models

Does pay impact quality?

- Economic theory holds that workers are rational actors
- Will choose to improve their performance in response to a scheme that rewards improvements with financial gain
- Example: executive compensation tied to stock price

Different pay schemes

- Lazear studied of workers who installed windshields on a production line
- Switched from pay per hour to pay per unit during a year and a half
- Individual productivity for workers who started in the hourly rate and switched to the per-unit scheme increased by 20%
- Conclusion: performance-based pay schemes can elicit improved performance

Is that the whole story?

- Sometimes financial incentives can undermine “intrinsic motivation”. This can lead to poorer outcomes.
- For complex tasks, performance pay can encourage workers to focus only on the aspects of their jobs that are actively measured
- Can also lead to employees avoid taking risks, thereby hampering innovation

Financial Incentives and the "Performance of Crowds"

- Experiment with economic incentives on Amazon Mechanical Turk
- An exciting tool for behavioral research, since you can recruit thousands of participants from a real labor market

Impact of compensation

- Does compensation change the quantity of work performed (output)?
- Does it change the quality of the work (accuracy)?

Re-order Traffic Images

Unsorted

2.

3.

Sorted

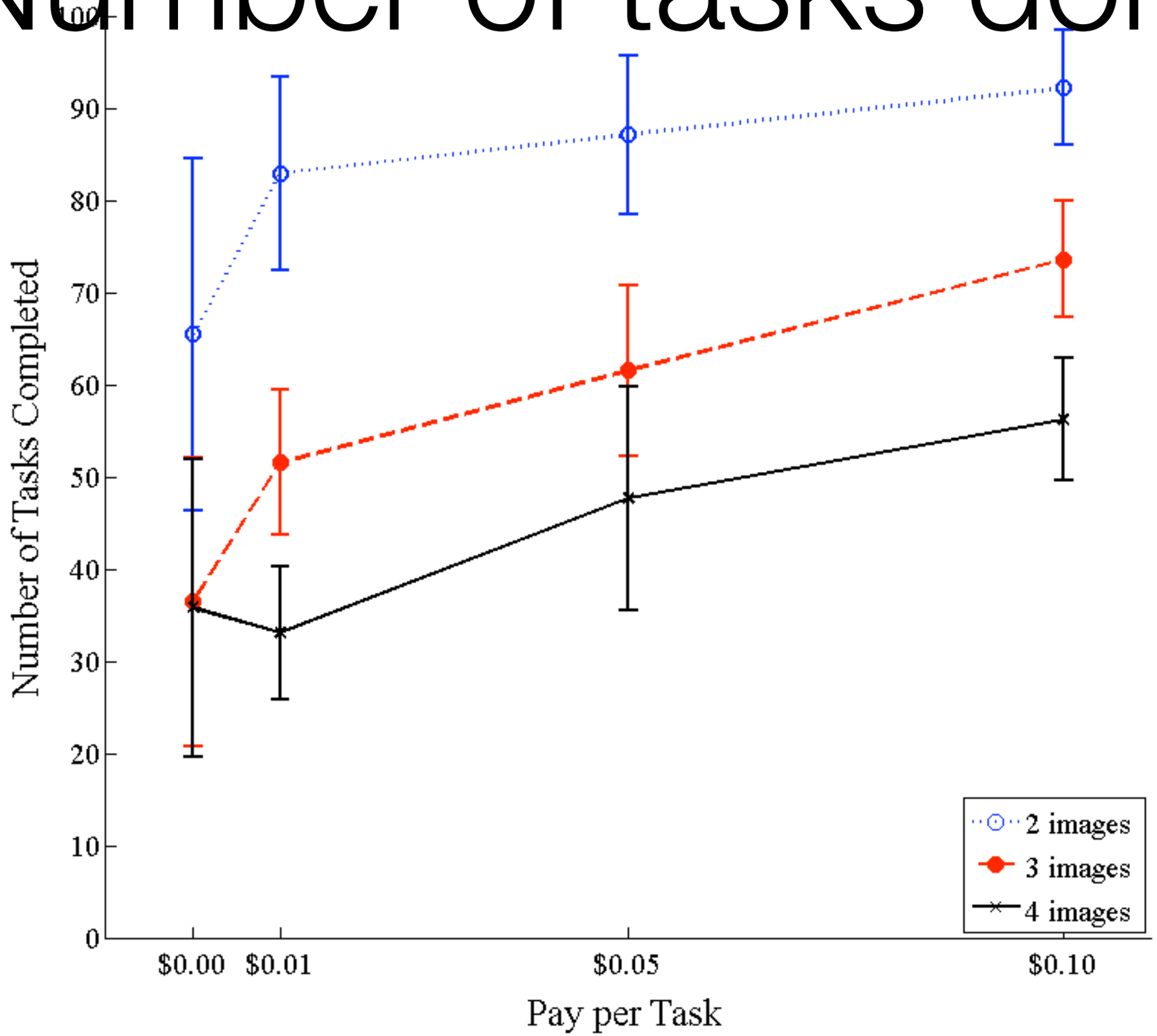
2.

3.

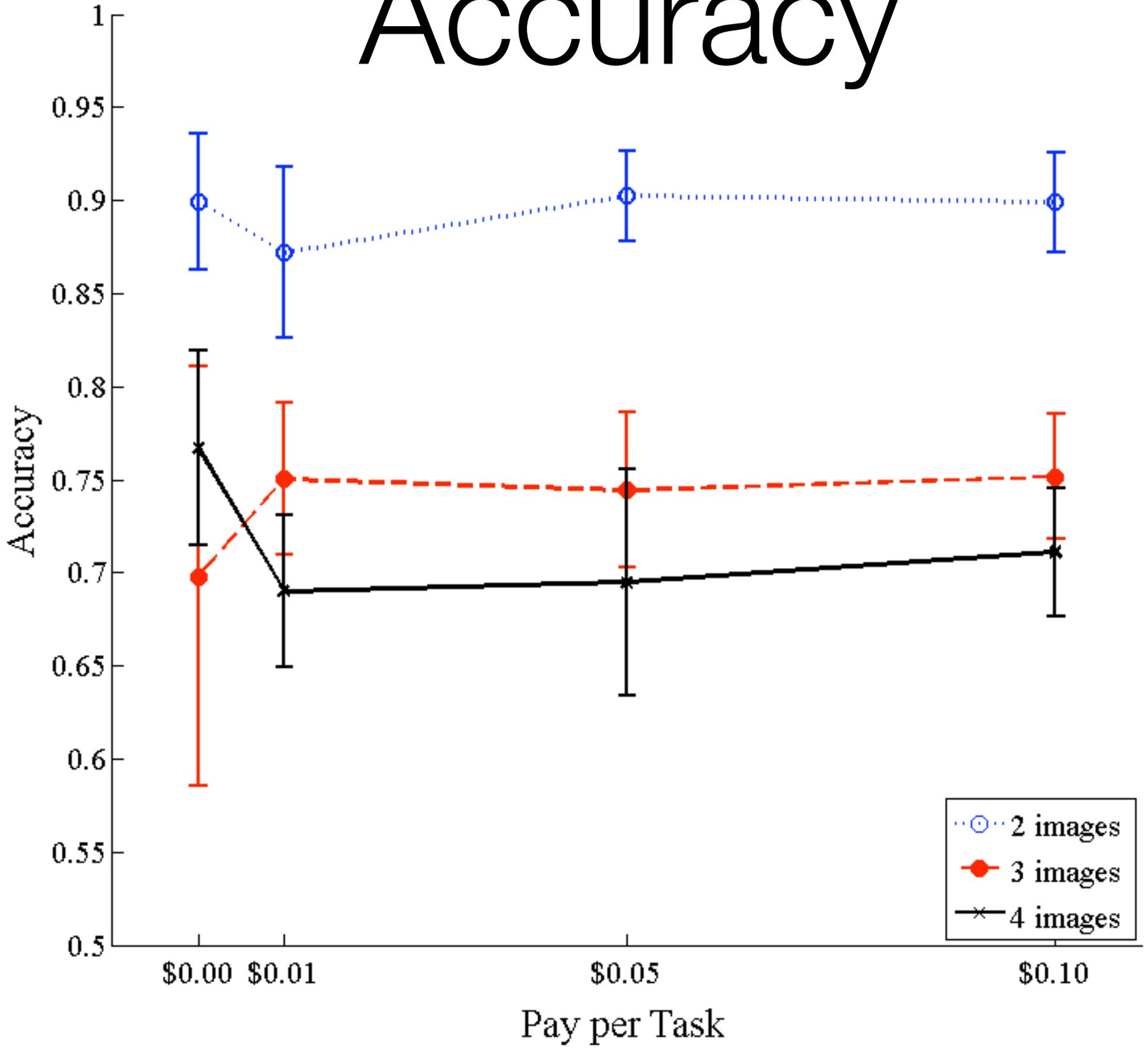
Payment scheme

- Everyone: $0.10 for doing training examples and filling out a survey
- Payment levels: nothing, 1¢, 5¢, 10¢ per set
- Num images per set (independent of payment): 2, 3, 4
- Each person sorted up to 99 sets of images, could end participation at any point and get paid for what they did
- 611 subjects sorted a total of 36,425 image sets

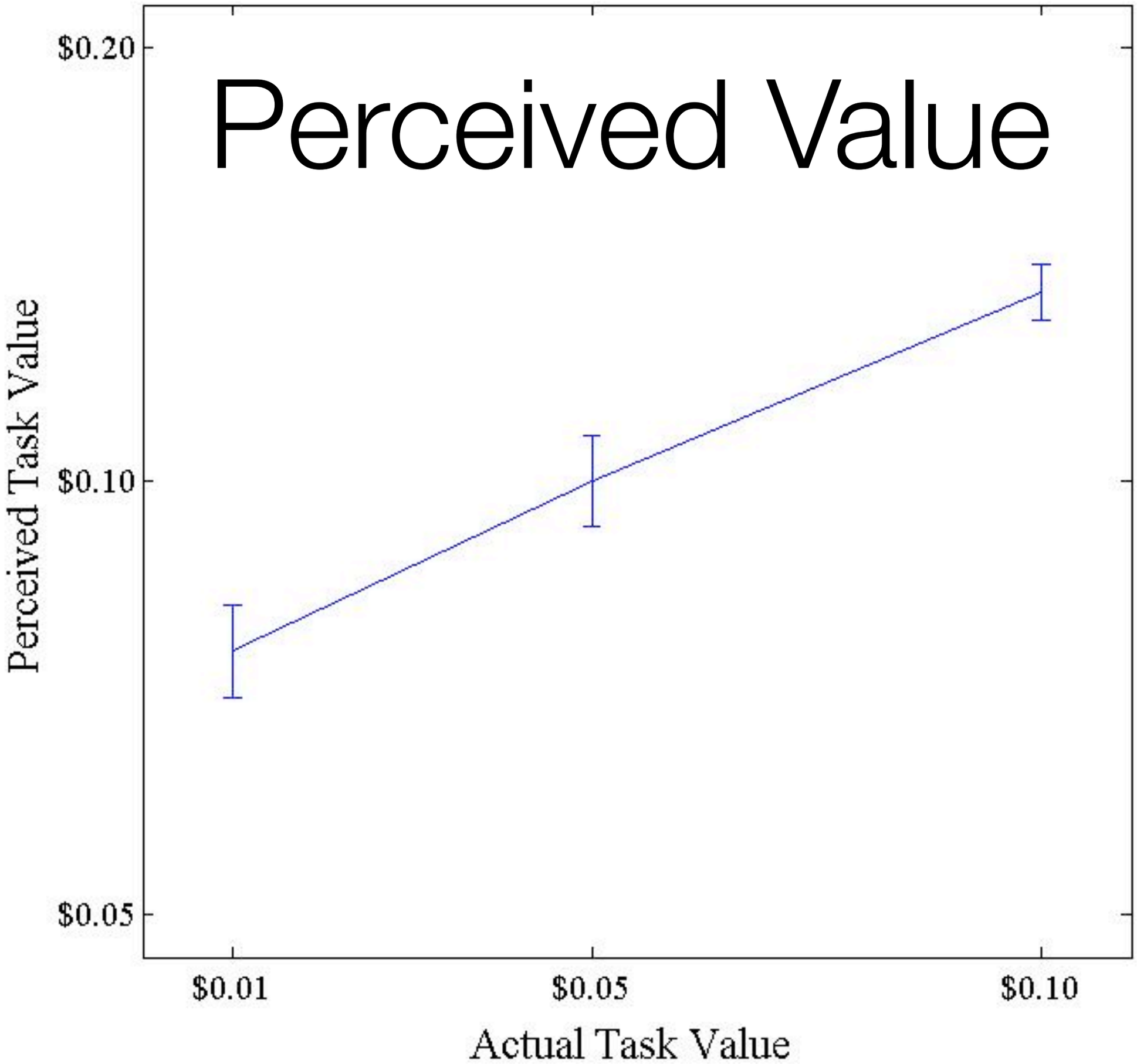
Number of tasks done



Accuracy


1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
$0.00
$0.01
$0.05
$0.10
Pay per Task
2 images
3 images
4 images

Perceived Value



Word Jumble Puzzles

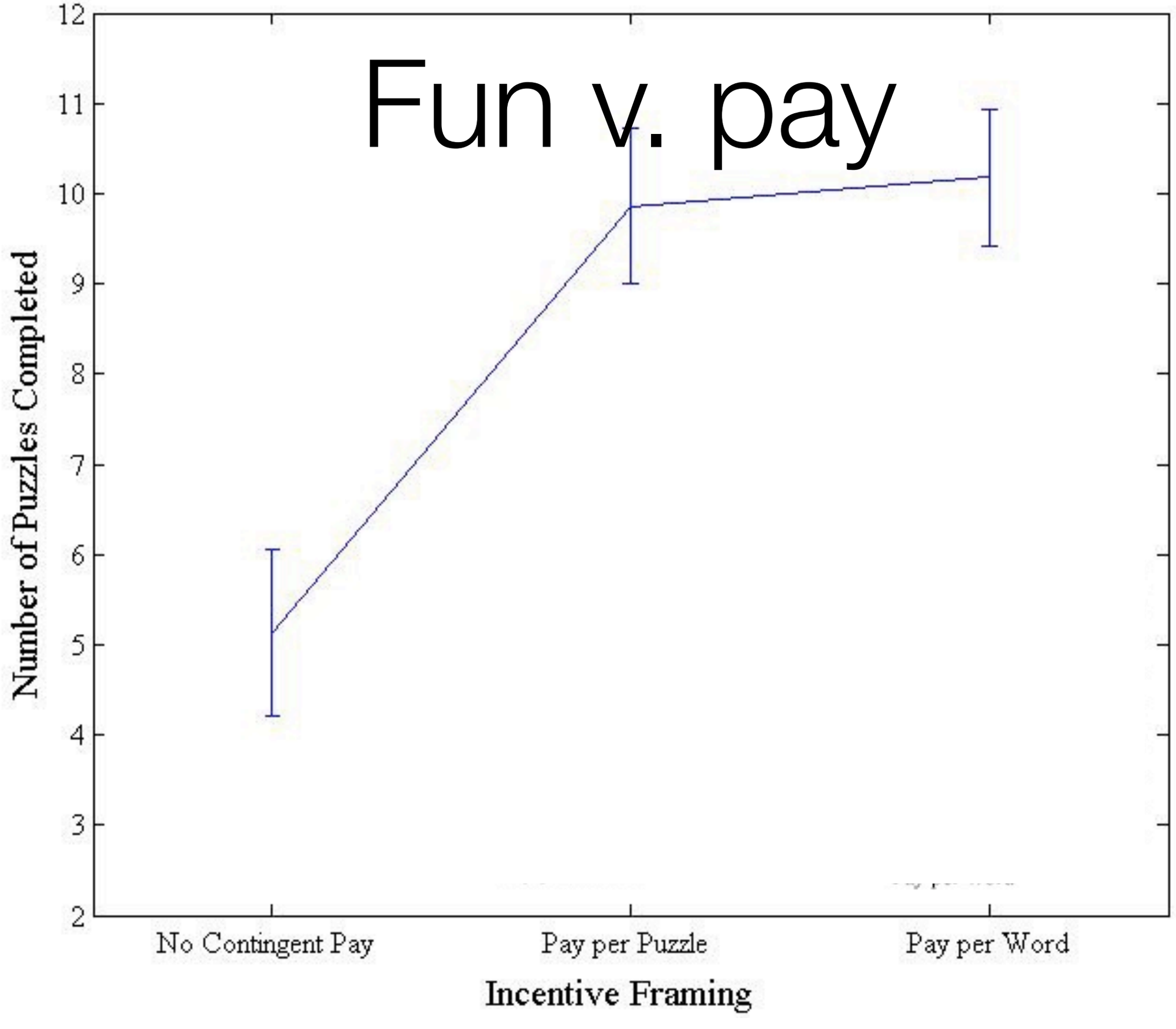
M	A	S	T	E	R	M	O	Z	Q	K	K	W	L	F
R	G	T	R	D	B	U	I	L	D	I	N	G	O	T
D	S	U	C	U	W	R	J	B	M	Q	P	G	L	C
L	Q	R	P	E	T	E	P	M	O	C	A	E	P	F
K	Z	T	C	V	F	T	B	X	W	Q	V	Q	A	I
S	U	L	F	J	D	G	W	U	W	I	C	G	Z	O
W	O	E	O	M	P	M	A	L	R	E	S	U	Z	L
X	V	R	Q	X	O	N	T	T	L	N	U	F	N	W
Y	H	B	A	I	E	N	S	P	L	I	C	N	E	F
E	H	L	D	T	A	V	A	W	I	S	C	E	K	U
N	K	G	K	L	T	T	E	F	Y	B	E	E	C	M
O	P	R	P	B	S	A	L	I	P	J	E	R	N	X
U	V	B	A	Z	J	M	I	D	H	F	D	G	U	V
Q	P	I	A	W	T	F	U	N	U	C	H	G	I	X
Y	K	I	O	T	W	L	O	R	N	O	A	Q	W	A

- Find as many of the of words in a set as you can:
- ACHIEVE, ATTAIN, BUILDING, CHAIR, COMPLETE, GREEN, LAMP, MASTER, MUSIC, PLANT, STAPLE, STEREO, STRIVE, SUCCEED, TURTLE
- Not all of the words listed are in the puzzle!

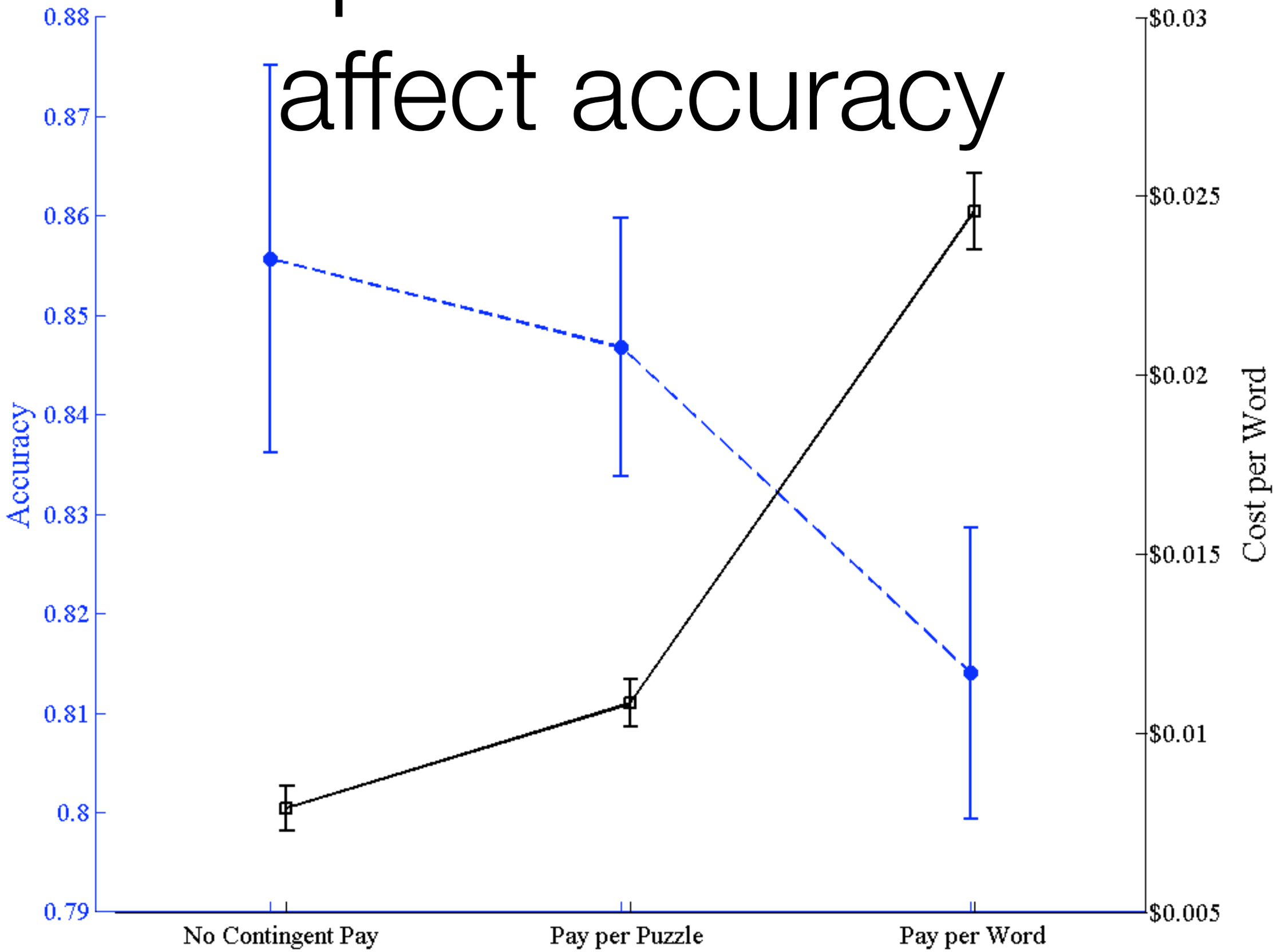
Experimental setup

- Different pay rates (just as before)
- Subjects were told that they would be paid either on a per-grid basis or a per-word basis, or not told anything
- quantity = number of puzzles completed
  quality = fraction of words found per puzzle
- Participants could do up to 24 puzzles
- 320 subjects solved 2736 puzzles, finding 23,440 words

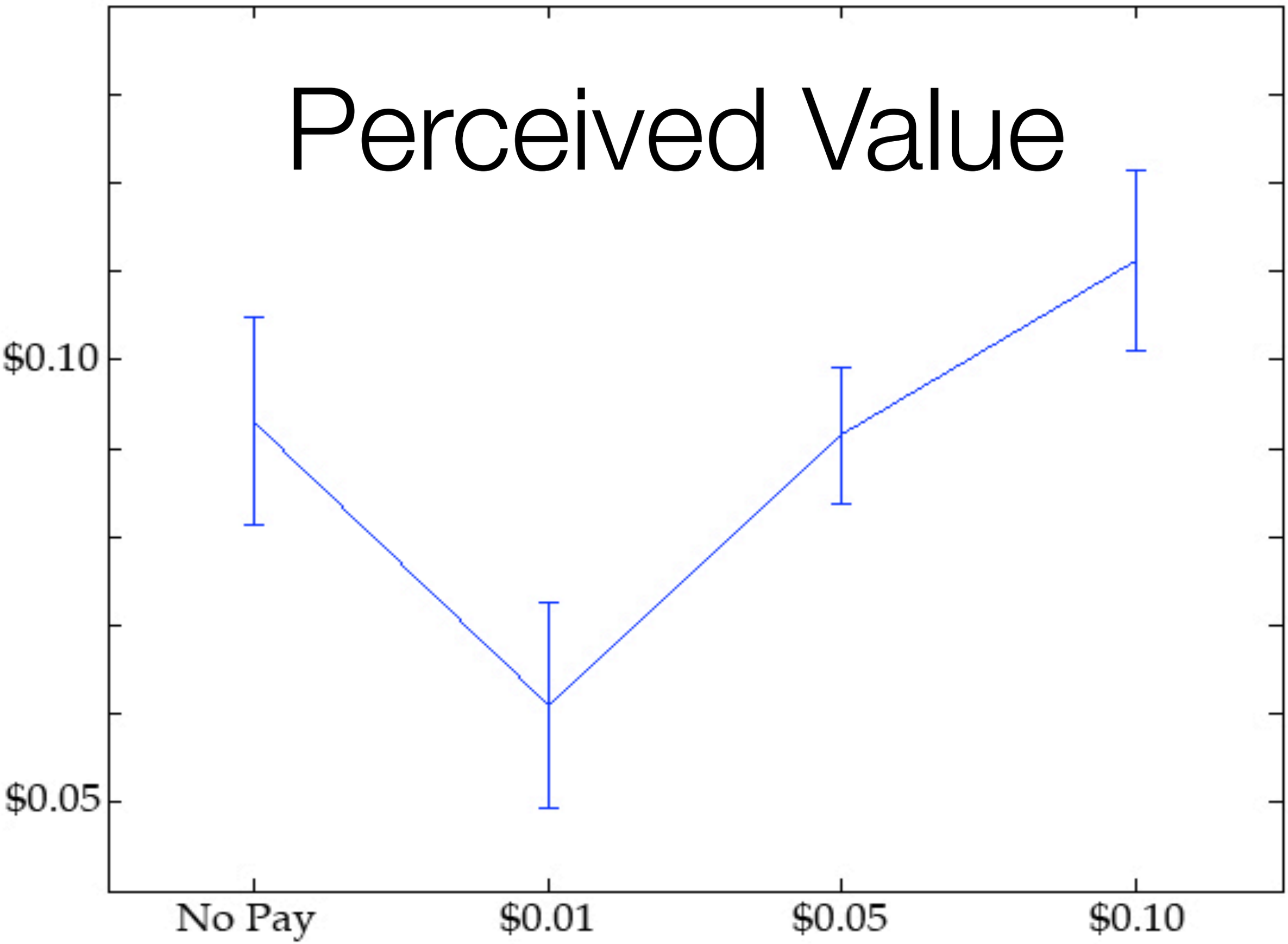
Fun v. pay


Number of Puzzles Completed
12
11
10
9
8
7
6
5
4
3
2
No Contingent Pay
Pay per Puzzle
Pay per Word
Incentive Framing

Compensation doesn't affect accuracy



Perceived Value



Findings

- Paying subjects elicited higher output than gamification, and increasing pay rate yielded even higher output
- However, paying subjects did not affect their accuracy
- Anchoring effects are significant – the reward you set impacts perceived value

Implications for your tasks?

- When you can use non-financial rewards, like intrinsic motivation, do so, since the quality of work will be the same
- When you can’t use intrinsic motivation, it might be in your best interest to pay as little as possible. Your work will be done slower, but quality will be similar.
- Is this fair to workers?

What do you think?

- Is studying workers on Mechanical Turk a valid way of studying other labor markets?
- What possible confounds are there?
- What could we do to control for them?