NETS 213: CROWDSOURCING AND HUMAN COMPUTATION

Quality Control Part 1

Classification System for Human Computation

Motivation

Quality Control

Aggregation

Human Skill

Process Order

Task-request Cardinality

Quality Control

Crowdsourcing typically takes place through an open call on the internet, where anyone can participate. How do we know that they are doing work conscientiously? Can we trust them not to cheat or sabotage the system? Even if they are acting in good faith, how do we know that they're doing things right?

Different Mechanisms for Quality Control

Aggregation and redundancy

Embedded gold standard data

Reputation systems

Economic incentives

Statistical models

ESP Game

"think like each other"

Player 1 guesses: purse
Player 1 guesses: bag
Player 1 guesses: brown

Success! Agreement on "purse"

Player 2 guesses: handbag

Player 2 guesses: purse
Success! Agreement on "purse"

Rules

- Partners agree on as many images as they can in 2.5 minutes
- Get points for every image, more if they agree on 15 images
- Players can also choose to pass or opt out on difficult image
- If a player clicks the pass button, a message is generated on their partner's screen; a pair cannot pass on an image until both have passed

Taboo Words

- Players are not allowed to guess certain words
- Taboo words are the previous set of agreed upon words (up to 6)
- Initial labels for an image are often general ones (like “man” or “picture”)
- Taboo words generate more specific labels and guarantee that images get several different labels

ARTigo

ABOUT ARTIGO
BLOG / /
HIGHSCORE

SEARCH

- SIMPLE SEARCH
- ADVANCED SEARCH

GAMES

- ARTIGO GAME
- ARTIGO TABOO
- KARIDO
- TAG A TAG BETA
- COMBINO BETA

Tags
NACHDENKLICH MAN

Game stats

- For 4 months in 2003, 13,630 people played the ESP game, generating 1,271,451 labels for 293,760 different images
- 3.89 labels/minute from one pair of players
- At this rate, 5,000 people playing the game 24 hours a day would label all images on Google (425,000,000 images) with 1 label each in 31 days
- In half a year, 6 words could be associated to every image in Google's index

ESP's Purpose is Good Labels for Search

- Labels that players agree on tend to be "better"
- ESP game disregards the labels that players don't agree on
- Can run the image through many pairs of players
- Establish a threshold for good labels (permissive = 1 pair agrees, strict = 40 agree)

Are they any good?

- Are these labels good for search?
- Is agreement indicative of better search labels?
- Is cheating a problem for the ESP game?
- How do they counter act it?

Google Image Labeler BETA

Google Image Labeler

time left

01:09

score

140

passes

0

label pass

Your partner has suggested 9 labels.

zoom out

off-limits

hand
red

my labels

print
handprint

Original Evaluation

Pick 20 images at random that have at least 5 labels

15 people the images and agreed on labels

Do these have anything to do with the image?

Dog

Leash

German

Shepard

Standing

Canine

When is an image done?

When it accumulates enough keywords not to be fun anymore

System notes when an image is repeatedly passed

Can re-label images at a future date to see if their labels are still timely and appropriate

Pre-recorded game play

The server records the timing of a session between two people

Each side can be used to play with a single player in the future

Especially useful when game is gaining in popularity

Cheating in ESP

Partners cannot communicate with each other, so cheating is hard

Could propagate a strategy on a popular web site ("Let's always type A")

Randomly paired players and pre-recorded game play make it hard

Ground Truth

Ability to produce labels of expert quality

Measure the quality of labels on an authoritative set

How good are labels from non-experts compared to labels from experts?

Fast and Cheap – But is it Good?

Snow, O'Conner, Jurafsky and Ng (2008)

Can Turkers be used to create data for natural language processing?

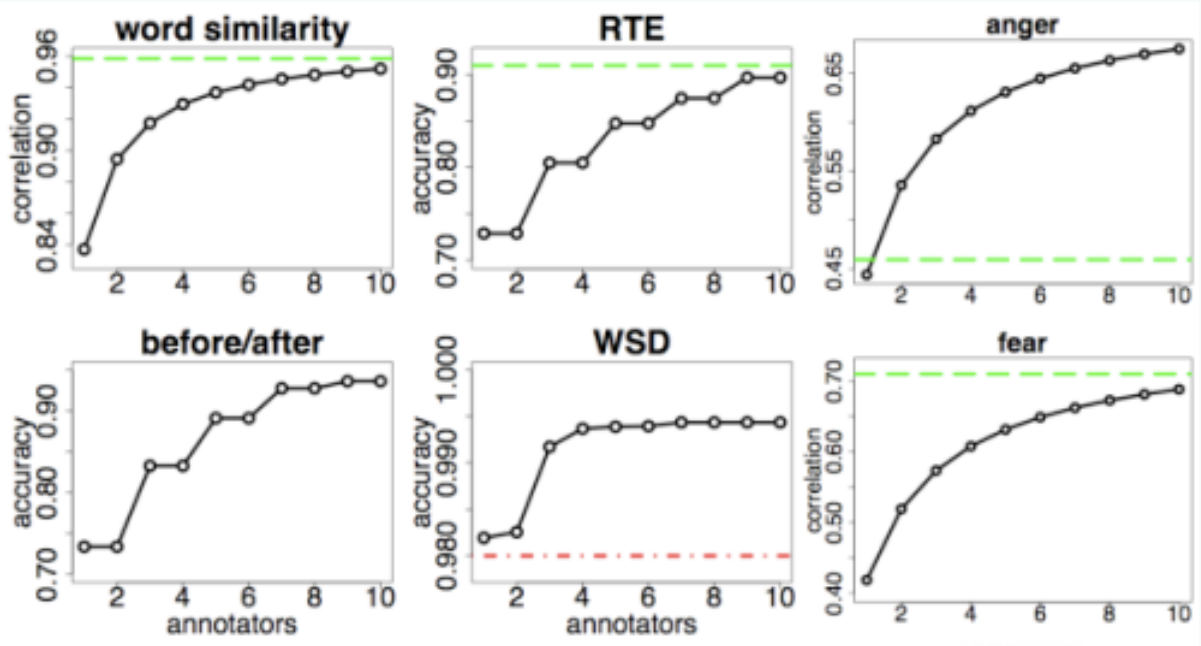
Measured their performance in a series of well-designed experiments

NLP Annotation

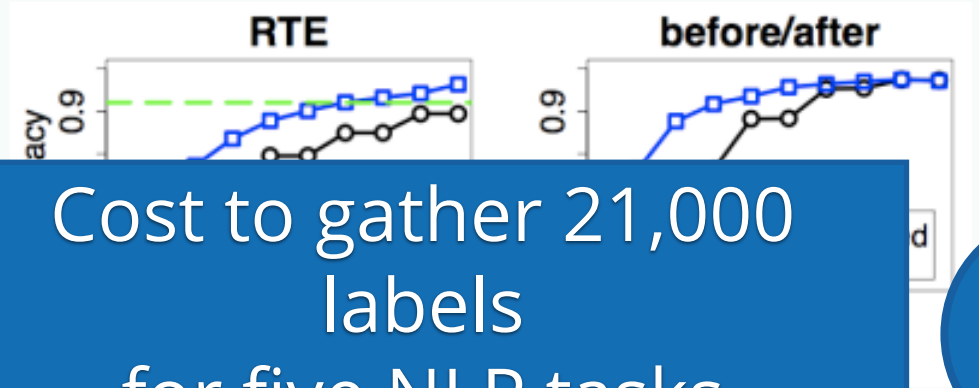
Affect Recognition	fear(“Tropical storm forms in Atlantic”) > fear(“Goal delight for Sheva”)
Word Similarity	sim(lad, boy) > sim(rooster, noon)
Textual Entailment	if “Microsoft was established in Italy in 1985” then “Microsoft was established in 1985”?
Word Sense	“a bass on the line” v. “a funky bass line”
Temporal Annotation	*collapsed* happens before *explosion* in: “The building collapsed in the wake of the massive explosion.”

Snow *et al* (EMNLP 2008)

NLP Annotation



Combine non-expert judgments for high correlation with experts



Cost to gather 21,000 labels for five NLP tasks

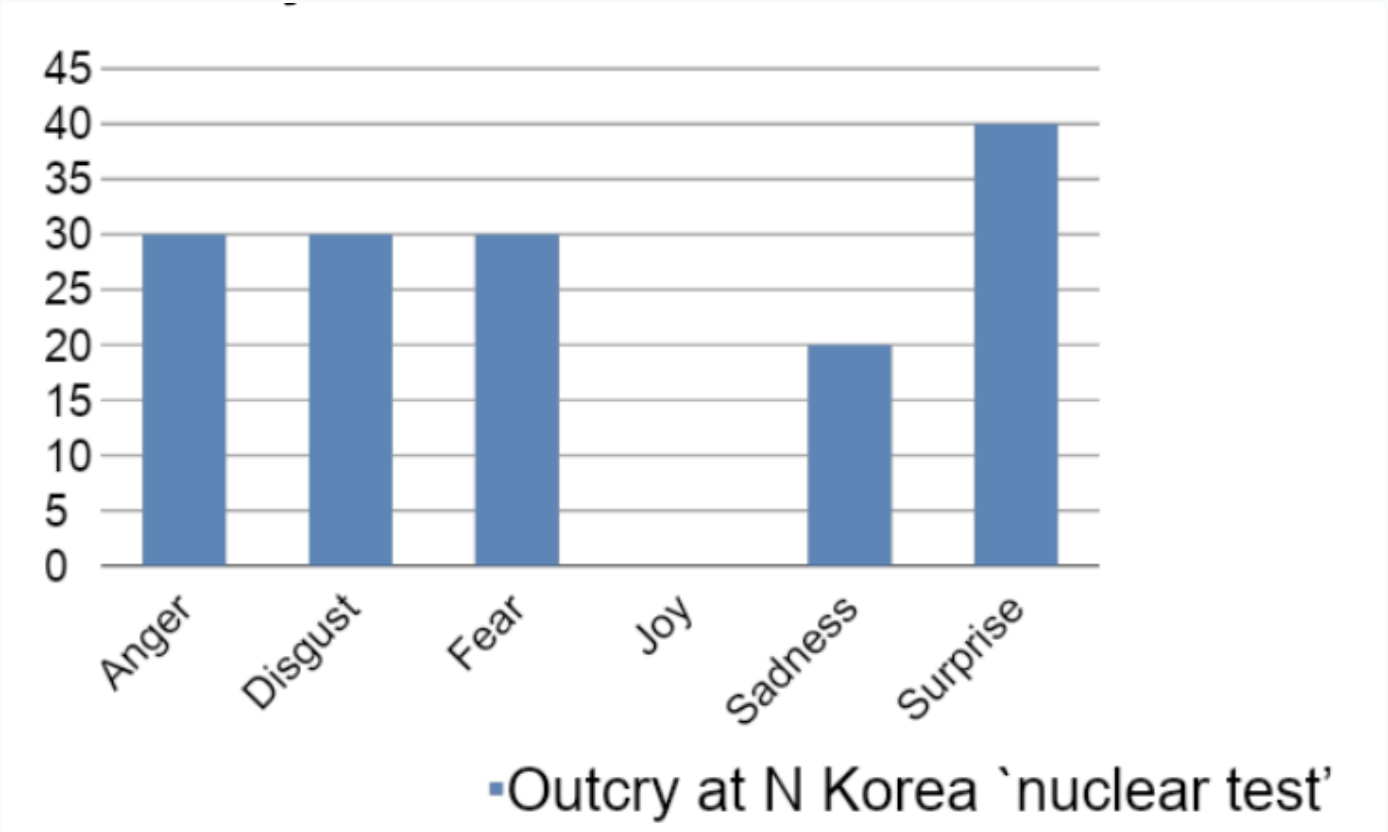
Weight w/ small amount of gold standard data for

$ 25. 82

Affect Recognition

Turkers are shown short headlines

Given numeric scores to 6 emotions



Affect Recognition Goals

Sentiment Analysis – enhance the standard positive/negative analysis with more nuanced emotions

Computer assisted creativity – generate text for computational advertising or persuasive communication

Verbal expressively for speech-to-text generation – improve the naturalness and effectiveness of computer voices

Word Similarity

Give a subjective numeric score about how similar a pair of words is

30 pairs of related words like {boy, lad} and unrelated words like {noon, string}

Used in psycholinguistic experiments

sim(lad, boy) > sim(rooster, noon)

Word Sense Disambiguation

Read a paragraph of text, and pick the best meaning for a word

Robert E. Lyons III was appointed **president** and chief operating officer…

1) executive officer of a firm, corporation, or university
2) head of a country (other than the U.S.)
3) head of the U.S., President of the United States

Recognizing Textual Entailment

Decide whether one sentence is implied by another

Is “Oil prices drop” implied by “Crude Oil Prices Slump”?

Is “Oil prices drop” implied by “The government announced that it plans to raise oil prices”?

Temporal Annotation

Did a verb mentioned in a text happen before or after another verb?

It just blew up in the air, and then we saw two fireballs go down to the water, and there was smoke coming up from that.

Did *go down* happen before/after *coming up?*

Did *blew up* happen before/after *saw?*

Experiments

These data sets have existing labels that were created by experts

We can therefore measure how well the workers' labels correspond to experts

What measurements should we use?

Correlation

Headline	Expert	Non-expert
Beware of peanut butter pathogens	37	15
Experts offer advice on salmonella	23	10
Indonesian with bird flu dies	45	39
Thousands tested after Russian H5N1 outbreak	71	80
Roots of autism more complex than thought	15	20
Largest ever autism study identifies two genetic culprits	12	22

Kendall tau rank correlation coefficient

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{1/2\ n^*(n-1)}$$

Headline	Expert	Non-expert
Beware of peanut butter pathogens	37	15
Experts offer advice on salmonella	23	10

Concordant > >

Kendall tau rank correlation coefficient

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{1/2\ n^*(n-1)}$$

Headline	Expert	Non-expert
Experts offer advice on salmonella	23	10
Largest ever autism study identifies two genetic culprits	12	22

discordant > <

Kendall tau rank correlation coefficient

$$\tau = \frac{\text{(number of concordant pairs) - (number of discordant pairs)}}{1/2\ n^*(n-1)}$$

$$\tau = \frac{11 - 4}{15} = 0.46$$

Fast and Cheap – But is it Good?

Snow, O'Conner, Jurafsky and Ng (2008)

Can Turkers be used to create data for natural language processing?

Measured their performance in a series of well-designed experiments

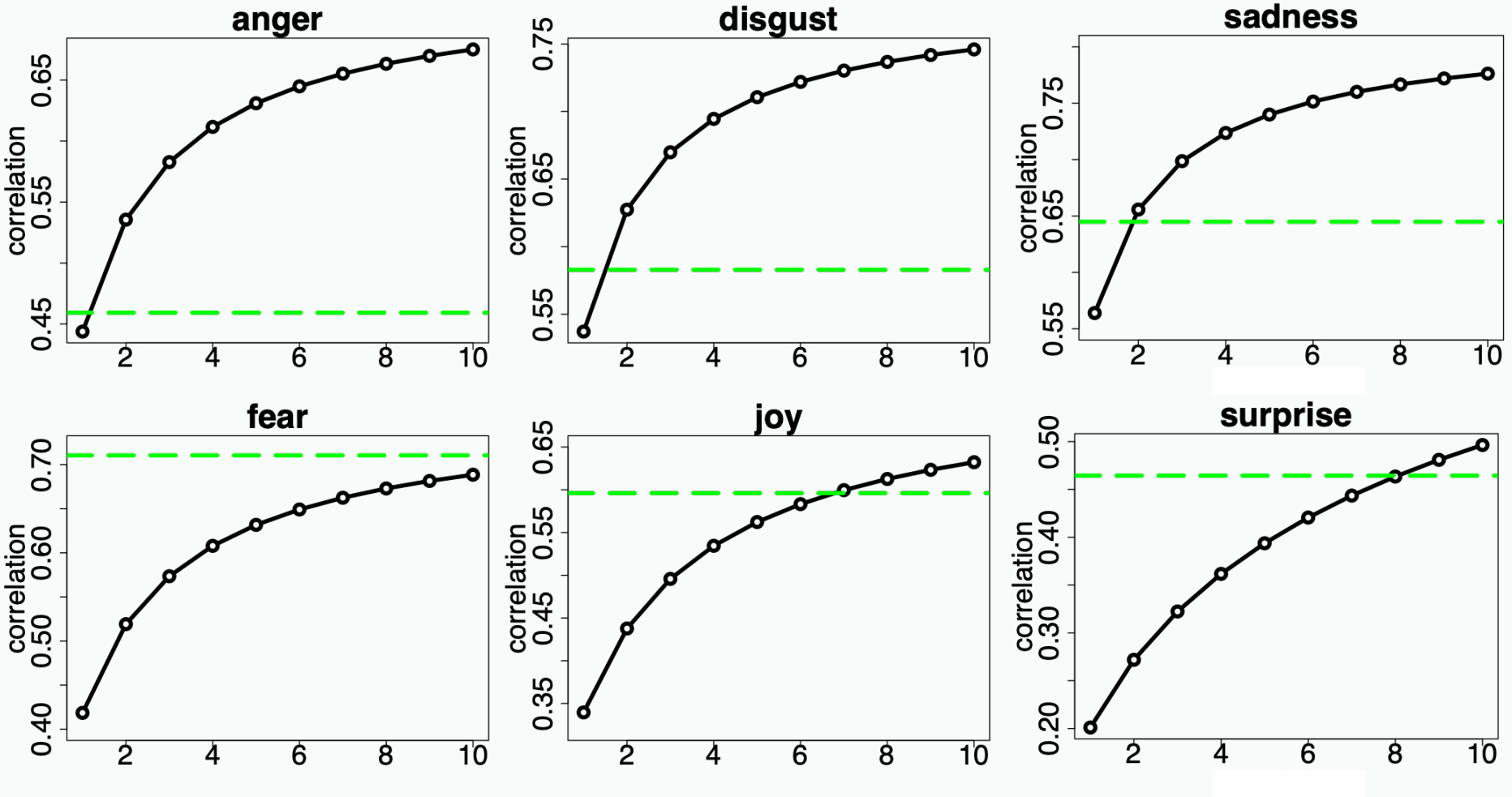
Experiments galore

- Calculate a correlation coefficient for each of the 5 data sets by comparing the non-expert values against expert values
- In most cases there were multiple annotations from different experts – this let's us establish a topline
- Instead of taking a single Turker, combine multiple Turkers for each judgment

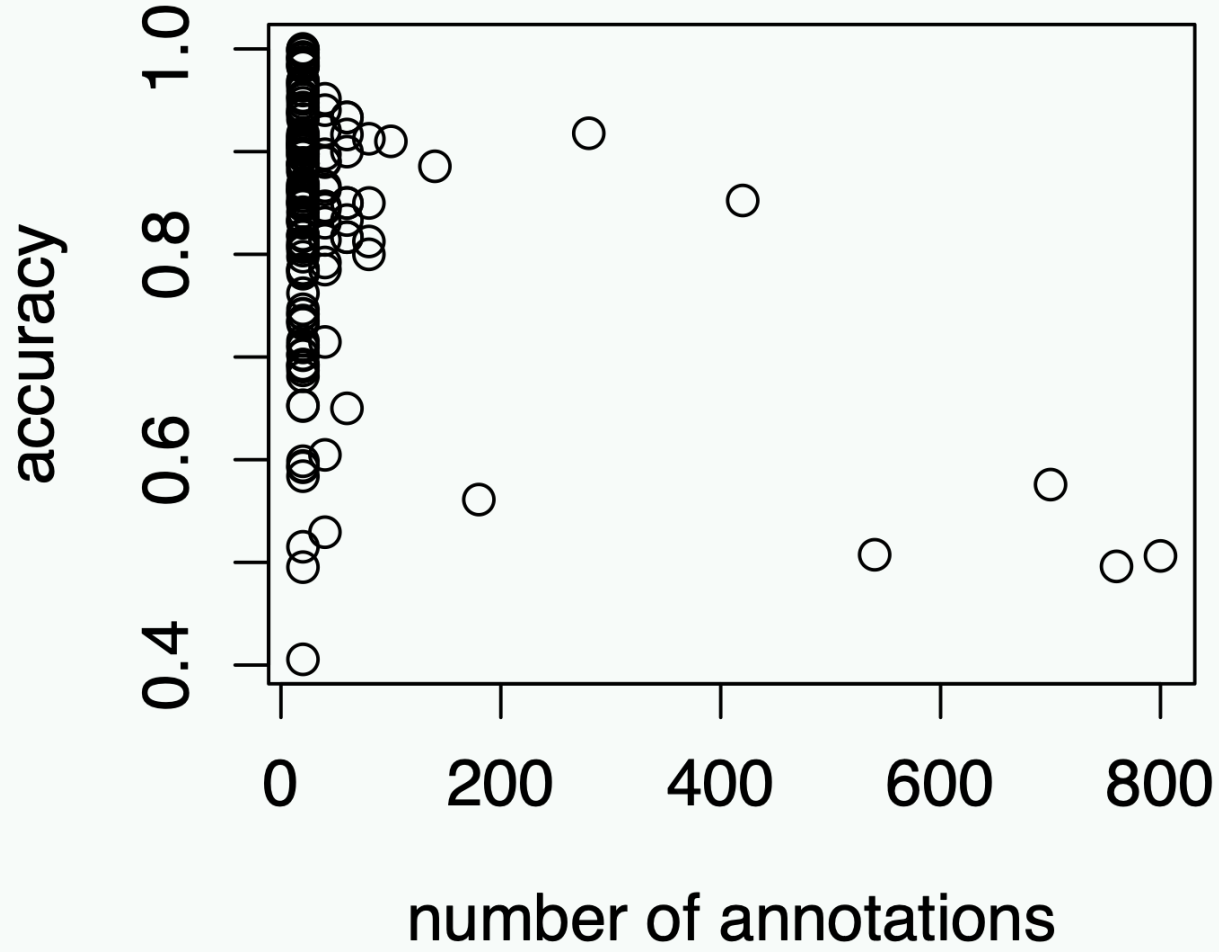
Sample sizes

Task	Labels
Affect Recognition	7000
Word Similarity	300
Recognizing Textual Entailment	8000
Word Sense Disambiguation	1770
Temporal Ordering	4620
Total	**21,690**

Agreement with experts increases as we add more Turkers



Accuracy of individual annotators



Calibrate the Turkers

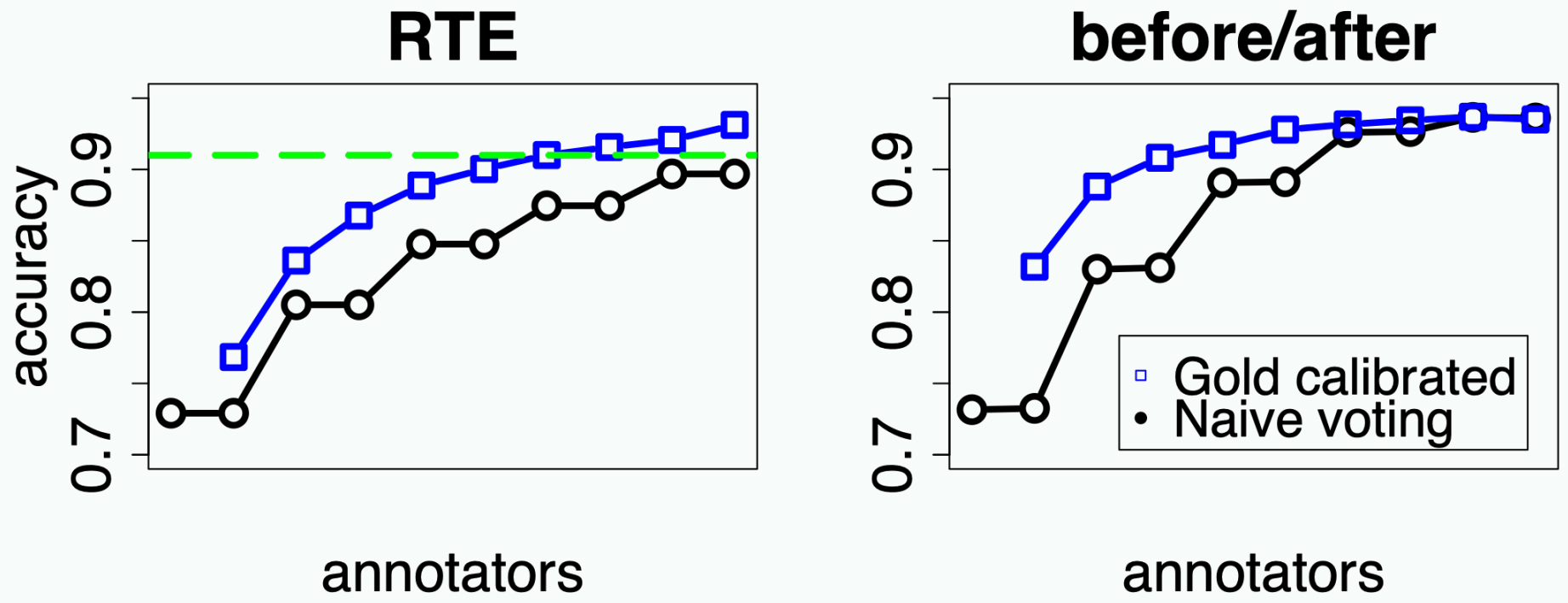
Instead of counting each Turker's vote equally, instead weight it

Set the weight of the score based on how well they do on gold standard data

Embed small amounts of expert labeled data alongside data without labels

Votes will count more for Turkers who perform well, and less for those who perform poorly

Weighted votes



Limitations?

Embedding gold standard data and weighted voting seems like the way to go

What are its limitations?

Limitations

- Requires objective answers – it is difficult to measure accuracy of subjective responses
- Applies mainly to structured data like multiple choice questions – things like content generation / free text responses can’t be calibrated in the same way
- Higher costs – requires creation of gold standard data by experts, requires multiple Workers to do each item

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. Varied hourly vs. per-unit pay.

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

Sorted

Payment scheme

Everyone: $0.10 for doing training examples and filling out a survey

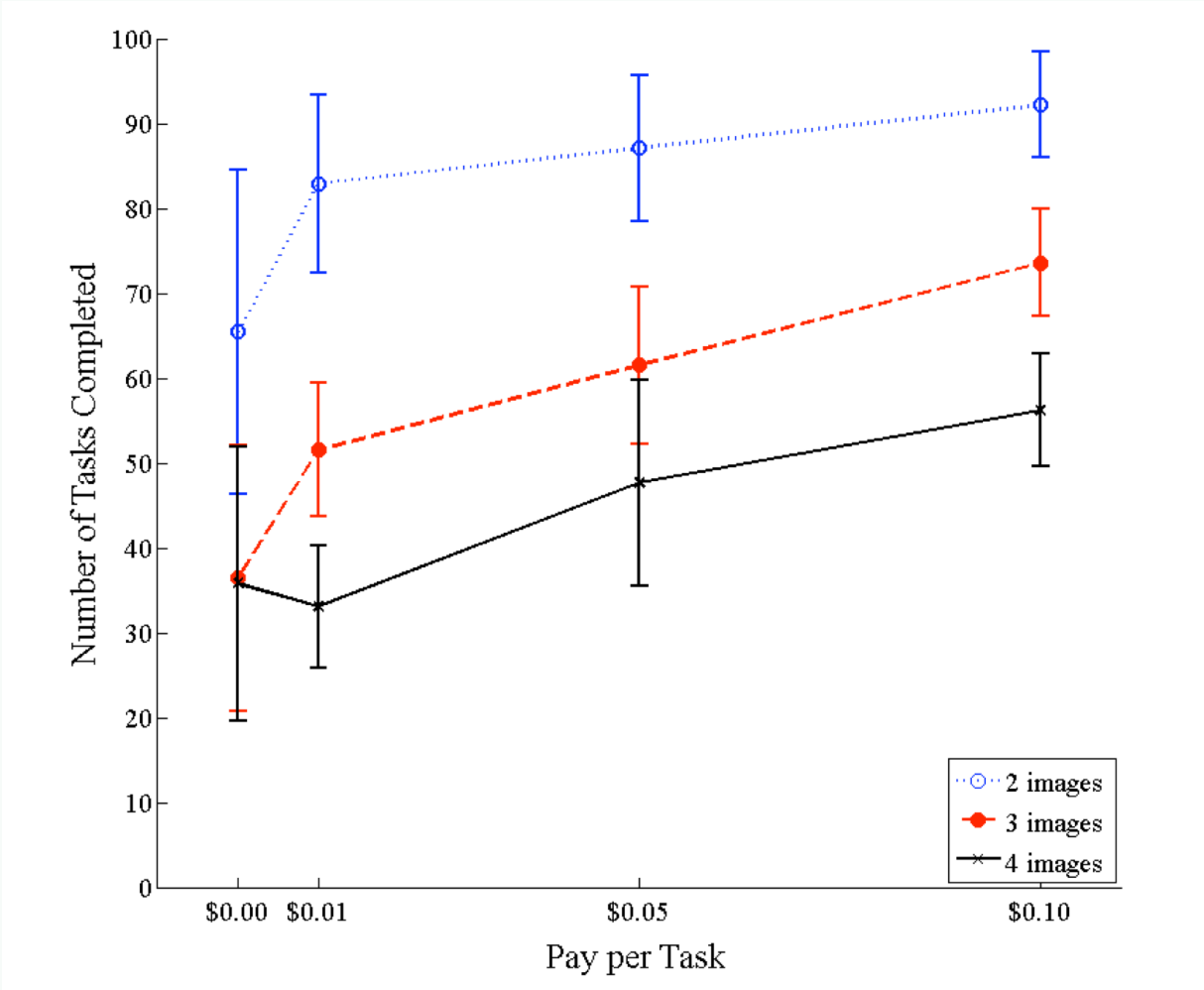
Payment levels: nothing, 1¢, 5¢, 10¢ per set

Num images per set (independent of payment): 2, 3, 4

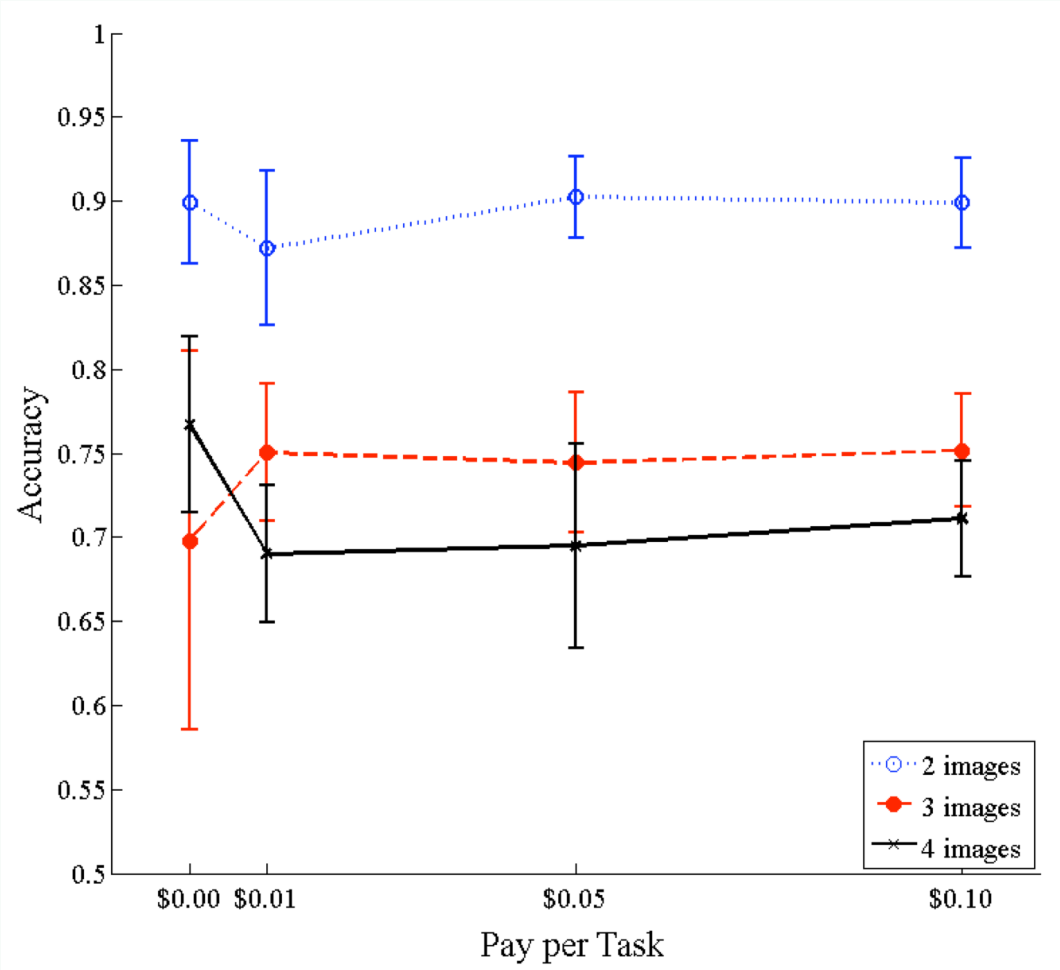
Each person sorted up to 99 sets of images, could end participation at any point and get paid for what they did

611 subjects sorted a total of 36,425 image sets

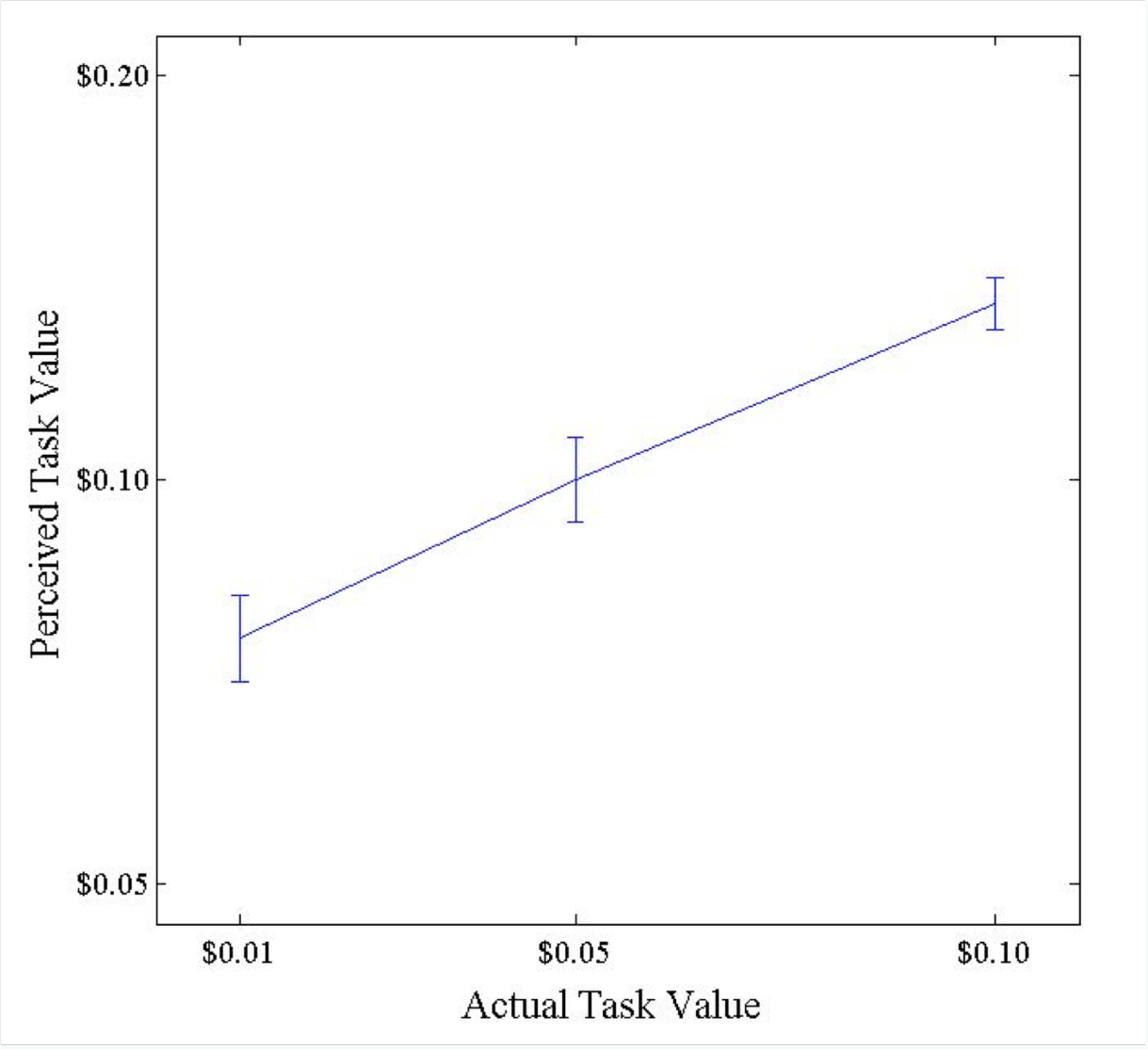
Number of tasks done



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

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)

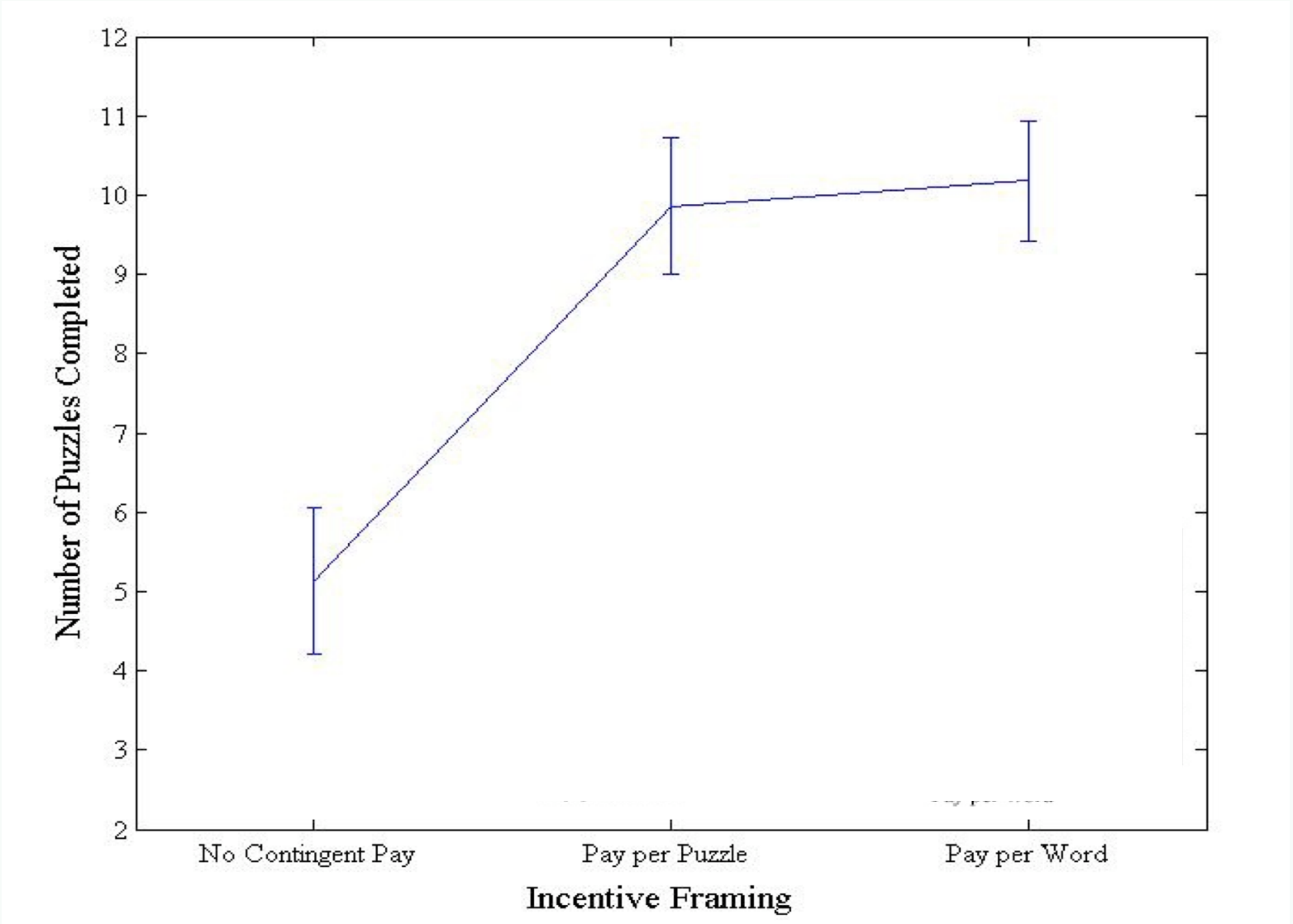
Subjects were told that they would be paid either on a per-grid basis or a per-word basis, or not told anything

quantity = number of puzzles completed
quality = fraction of words found per puzzle

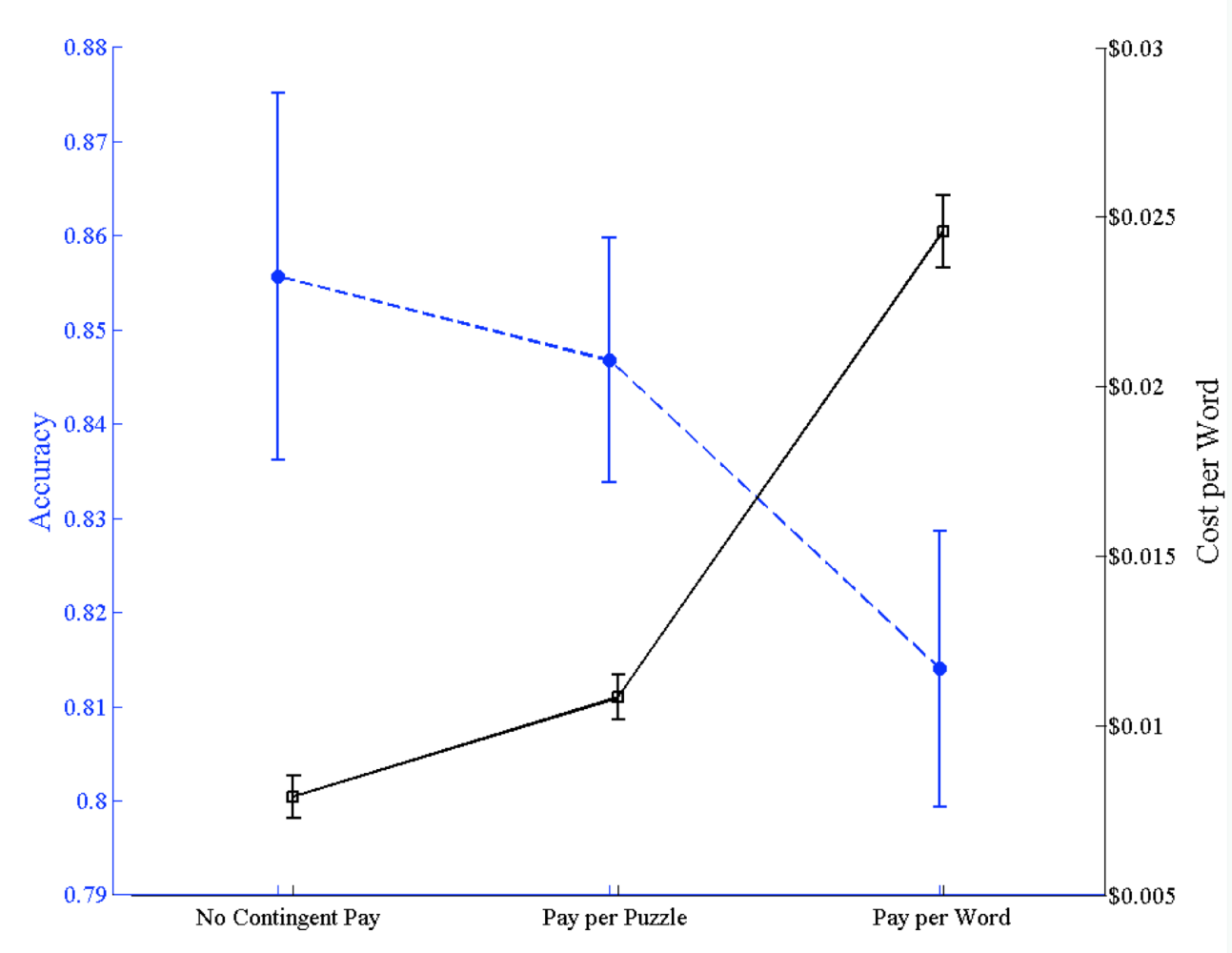
Participants could do up to 24 puzzles

320 subjects solved 2736 puzzles, finding 23,440 words

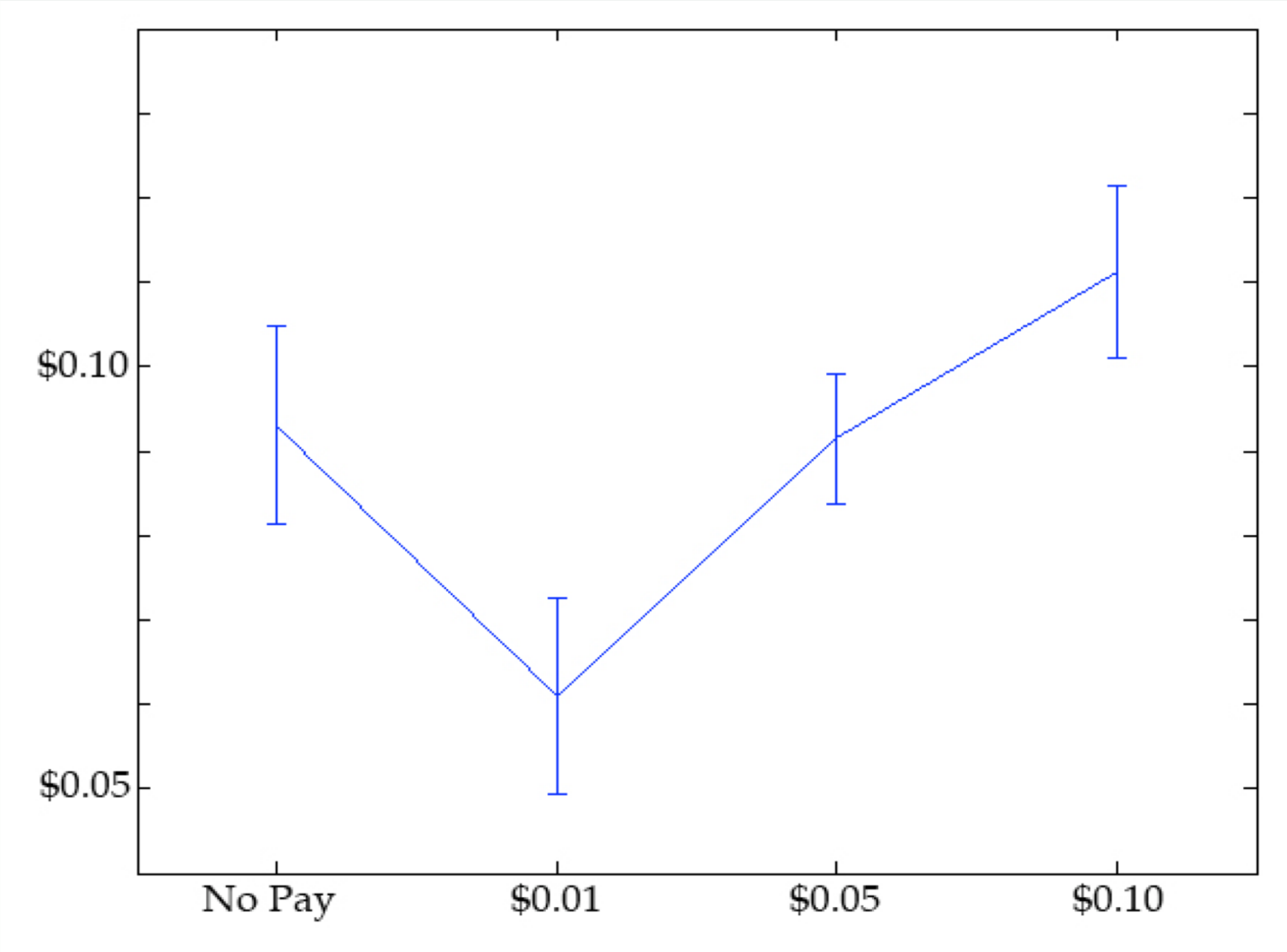
Fun v. pay


12
11
10
9
8
7
6
5
4
3
2
Number of Puzzles Completed
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?