How do financial incentives affect the crowd work?

— “Financial Incentives and the “Performance of Crowds”

— “Incentivizing High Quality Crowdwork”

— “The Effects of Pay-to-Quit Incentives on Crowdworker Task Quality”

Paper 1: Financial Incentives and the “Performance of Crowds”

- **Goal:**

Study the relationship between financial compensation and performance on web-based crowdsourcing models.

- **Approach:**

Conduct two experiments, Image Ordering and Word Puzzles, on a particular crowd-sourcing platform, Amazon’s Mechanical Turk

Experiment 1: IMAGE ORDERING

Participants are asked to sort a set of images taken from a traffic camera at 2-second intervals into chronological order.

Experiment Design

Three Difficulty Level:

Easy (2 images per set)

Medium (3 images per set)

Hard (4 images per set)

Four payment levels:

No pay

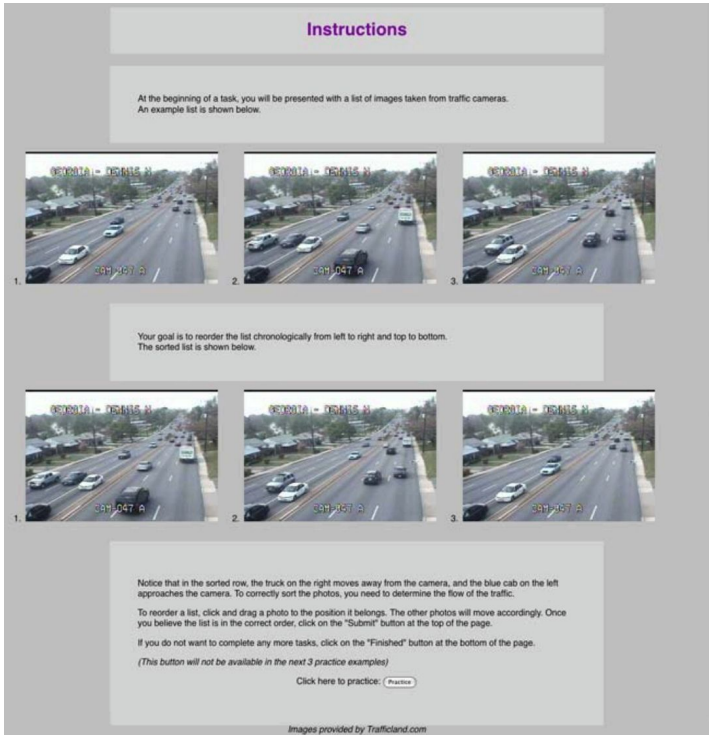
Low pay ($0.01 per image set)

Medium pay ($0.05 per image set)

High pay ($0.10 per image set)

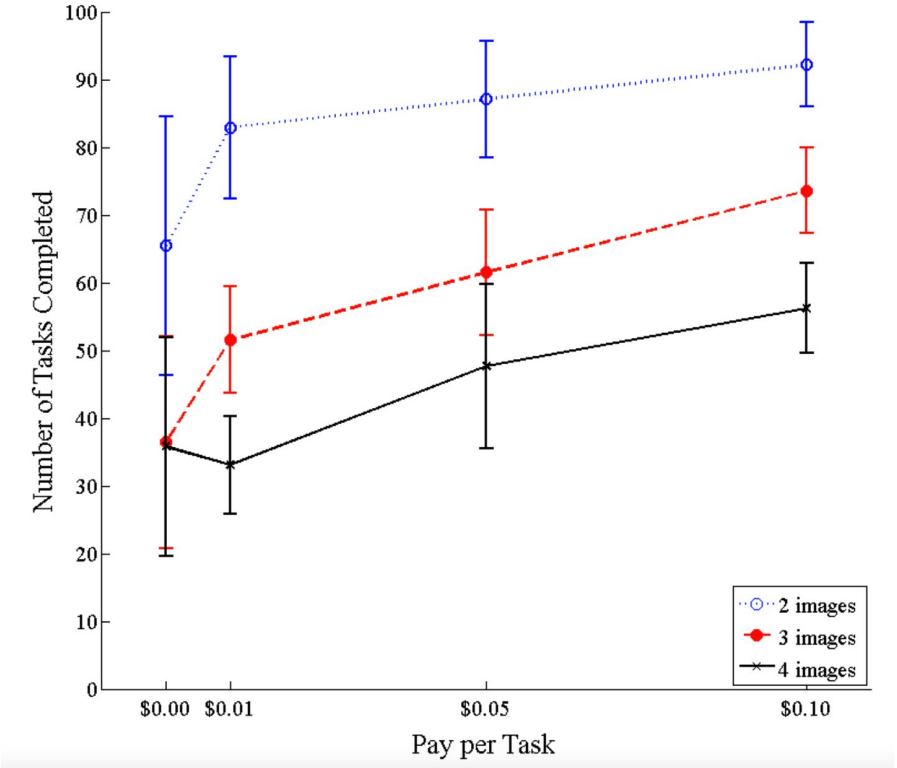
Experiment Implementation

- Task is published as very simple trail with payment of $0.01
- Participants were given information about how much they would be paid after trail, they can choose to continue or quite
- Randomly assigned to a difficulty level.
- Participants could sort any number from 0 to 99 sets of images



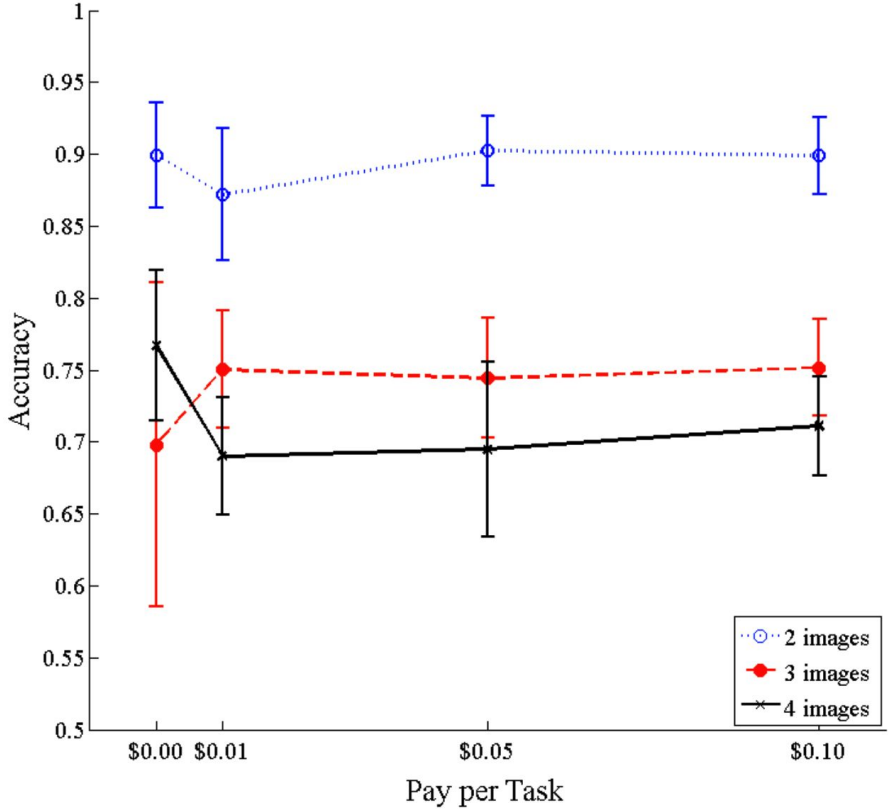
Experiment Results

- that across all difficulty levels, participants chose to complete more tasks on average when the pay was higher.
- that across all payment levels, the number of completed tasks decreased with increasing difficulty.



Experiment Results

- Increasing compensation did not improve accuracy



Experiment 2: WORD PUZZLES

Provide a list of words that might be found in the puzzle.

Participants did not know how many words from the list could actually be found in the puzzle.

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

Hint: Types of Fruit

Answers …

PEAR
GUAVA
POMELO
MANGO
COCONUT
LYCHEE
BANANA
KUMQUAT
POMEGRANATE
STRAWBERRY

DATE
MELON
GRAPE
LEMON
CHERRY
PLUM
TOMATO
ORANGE
LIME
APPLE

Experiment Design

Compensation scheme

1) **pay per puzzle:** paid for every puzzle successfully completed

2) **pay per word:** paid for every word found at a “piece rate”

Four Payment Levels

1) low pay ($ 0.01)

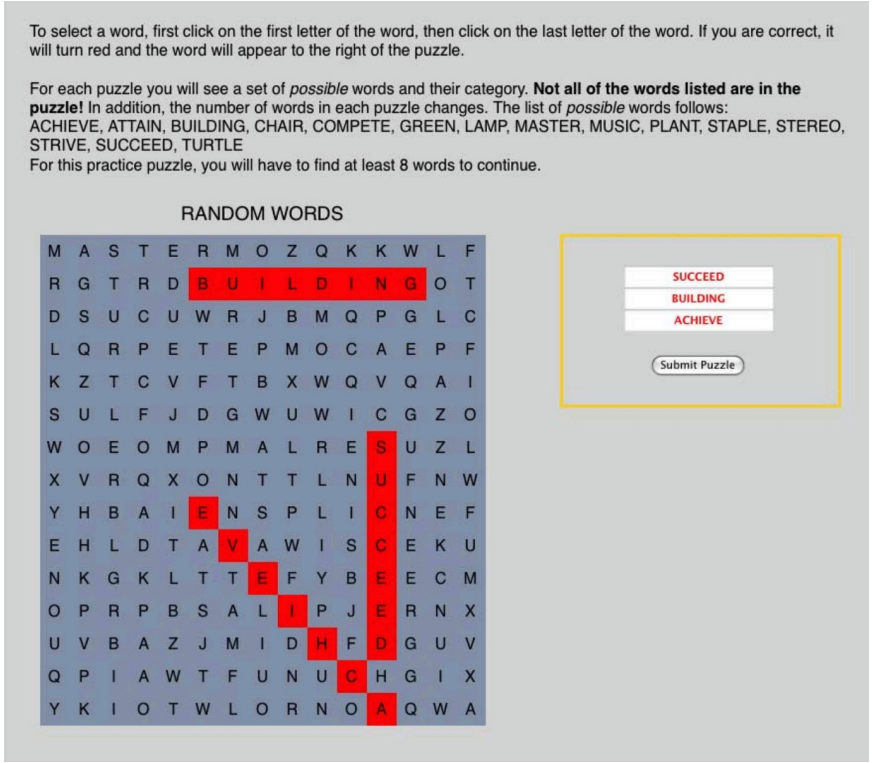
2) medium pay ($ 0.05)

3) high pay ($ 0.1)

4) no pay

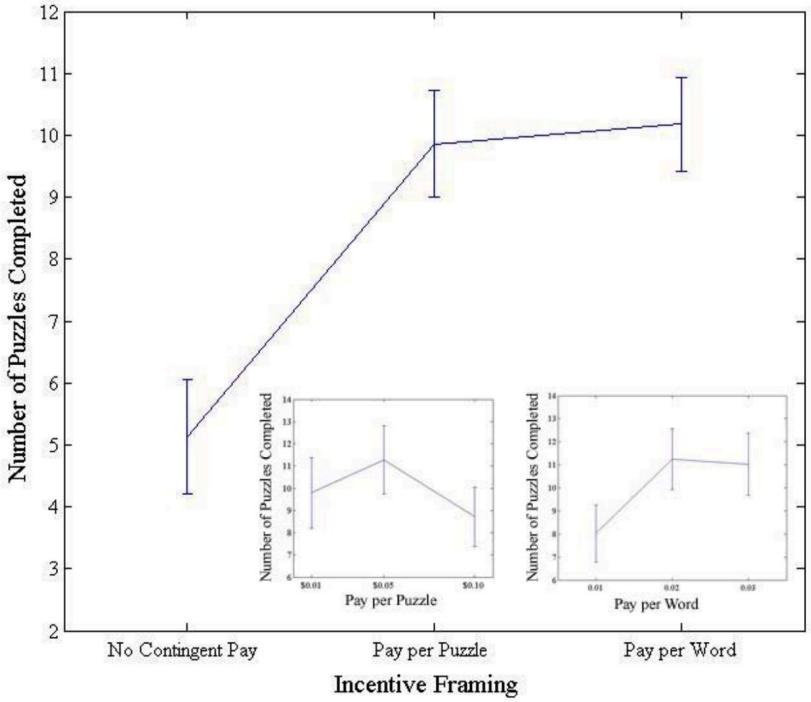
Experiment Implementation

- After task is accepted, they will be informed how much they would be paid, whether the payment would be by puzzle or by word.
- They could complete up to a maximum of 24 puzzles.
- They could then take as much time as they wanted on each puzzle



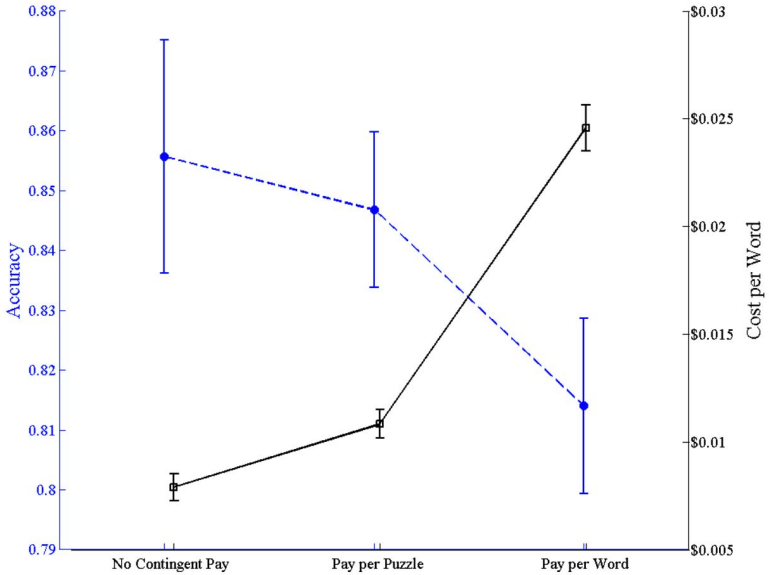
Experiment Results

- Participants who were paid either on a quota or a piece-rate basis completed more puzzles and found more words than participants who were not paid.
- No significant impact of compensation level on quantity of work within the two compensation schemes



Experiment Results

- The level of compensation had no significant effect on the quality of performance
- Those who enjoyed it the most completed 19.3 puzzles on average, compared to the 6.2 puzzles completed by those who only enjoyed it a little.



Participants were paid more in total (solid line) but found fewer words per puzzle (dashed line) in the pay per-word condition than in the pay per-puzzle condition

Conclusion

- first, increased payments increased the quantity of work performed, but not its quality;

- second, that the particular design of the compensation scheme (a quota scheme versus a piece rate, for example) can have a significant effect on quality even to the point where better work can be accomplished for less pay

Recommendations to Employers

- First, when it is possible to use non-financial rewards, such as Harnessing intrinsic motivation, the quality of the work will be as good or better than using financial rewards, and therefore work can be accomplished as effectively for little to no cost.
- Second, when it is not possible to incentivize work through intrinsic motivation (i.e., enjoyable tasks) or through social rewards, it may be in the employer's best interest to offer as little as possible— assuming a large enough crowd exists to make up for the diminished quantity of individual output the low pay would garner. Offering greater reward, in other words, may get the work done faster, but not better.

Discussion

1）Does experiment design in the paper is reasonable or do the conclusions drawn from the paper convince you?

2) Do you think there is anything that could be improved on the content of the experiment or the way the experiment was conducted?

3）How would you improve your the experiments to make the data get from it more objective and convincing？

— “Financial Incentives and the “Performance of Crowds”

— “Incentivizing High Quality Crowdwork”

— “The Effects of Pay-to-Quit Incentives on Crowdworker Task Quality”

Paper 2: Incentivizing High Quality Crowdwork

Goal:

Focus on performance-based payments (PBPs), to understand **when, where,** and **why** PBPs help, identifying properties of the payment, payment structure, and the task itself that make them most effective

Approach:

Design and run four randomized behavioral experiments on the popular crowdsourcing platform Amazon Mechanical Turk to research **whether, when, where,** and **why** does the PBPs work respectively

1: DOES PBP WORK?

- To verify whether PBPs can lead to higher quality crowdwork.
- To determine if there exists what we call an implicit PBP effect:

PBP effect: even if the requester offers a guaranteed payment, MTurk workers have subjective beliefs on the quality of work they must produce in order to receive this payment, and therefore behave as if the payments were (implicitly) performance-based.

Experiment: Proofreading

- workers were asked to proofread an article of 500 to 700 words and correct spelling errors.
- For each article, we randomly inserted 20 typos from a list of common spelling errors.

Experiment Design

Two treatments for the base payment ($ 0.5):

- **Non-Guaranteed:** There were no extra instructions.
- **Guaranteed:** Workers were told they would get paid if they found at least one typo

Bonus Treatment ($ 1.0):

- **No Bonus:** It had no bonus and no mention of a bonus.
- **Bonus for All:** All workers earned a $1 bonus after submitting the HIT.
- **PBP:** Workers earned a $1 bonus if they found 75% of the typos found by the other workers.

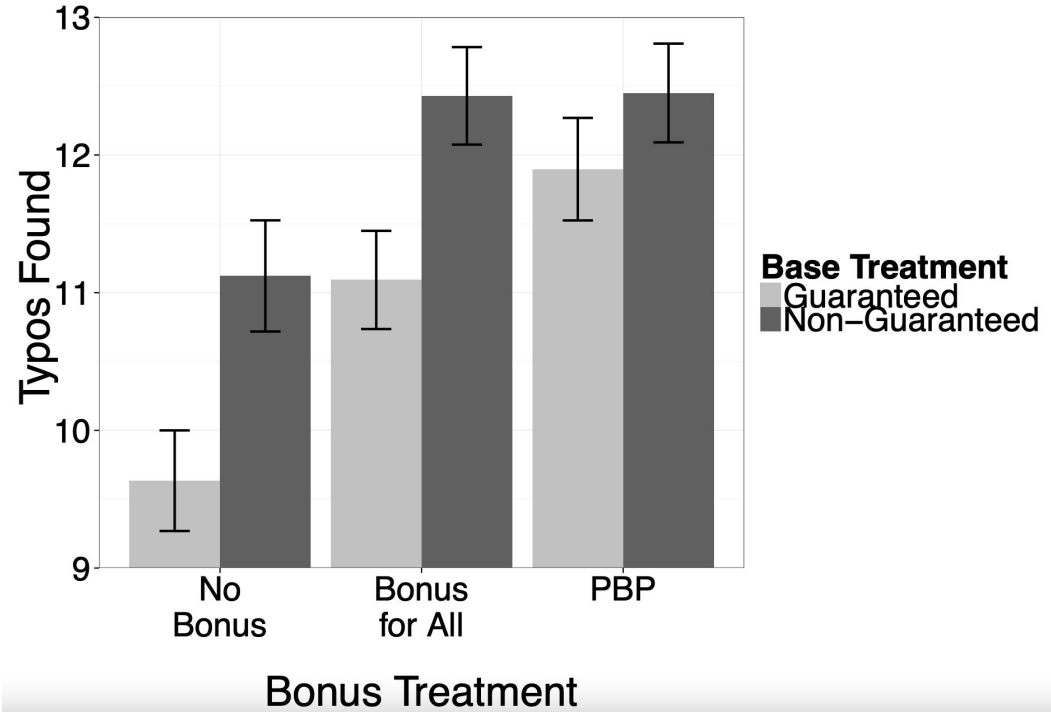
Experiment Implementation

1) Before accepting the HIT, each worker saw a preview consisting of the instructions and the base payment of $0.50 USD.

2) After workers accepted the HIT, they were randomly assigned to different treatments and then shown treatment-specific instructions, when applicable.

Experiment Result

- **PBPs improve quality.**
- All payment schemes may be implicitly performance-based.
- Simply paying more improves quality.
- PBPs can save money compared with high unconditional payments.



2: WHEN DOES PBP WORK?

To understand when PBPs help on changing two parameters of the payment scheme:

1) **the bonus threshold**

2) **the bonus amount**

Experiment Design: Bonus Thresholds

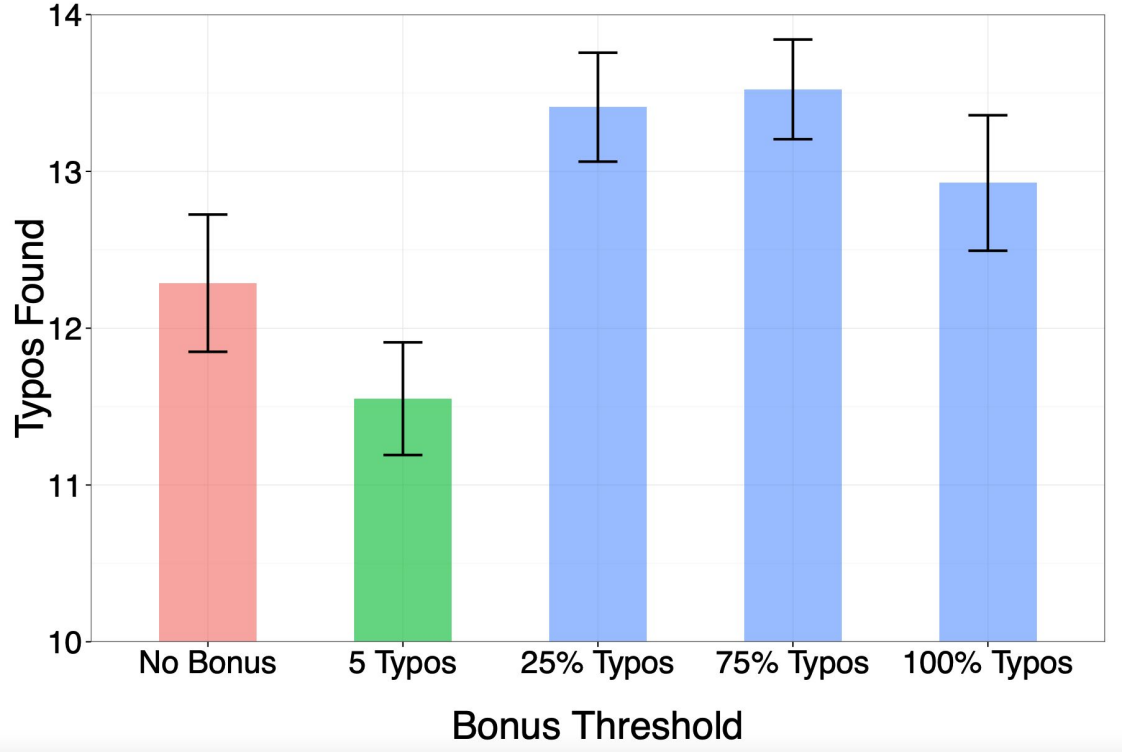
Used the same proofreading task described, with the same base payment of $0.50 and bonus of $1.

Workers were randomly assigned to treatments in which they were told they could earn the bonus if they found:

at least **5 typos** or at least **25%, 75%, or 100% of the typos** found by the other workers

Bonus Thresholds: Results

- **PBPs improve quality for a wide range of bonus thresholds.**
- Setting a threshold anywhere between 25% and 75% would yield similar results, so the improvements from PBPs are not overly sensitive to the threshold.
- if the bonus threshold is set too high, then workers' average performance slightly decreases.



Experiment Design : Bonus Amounts:

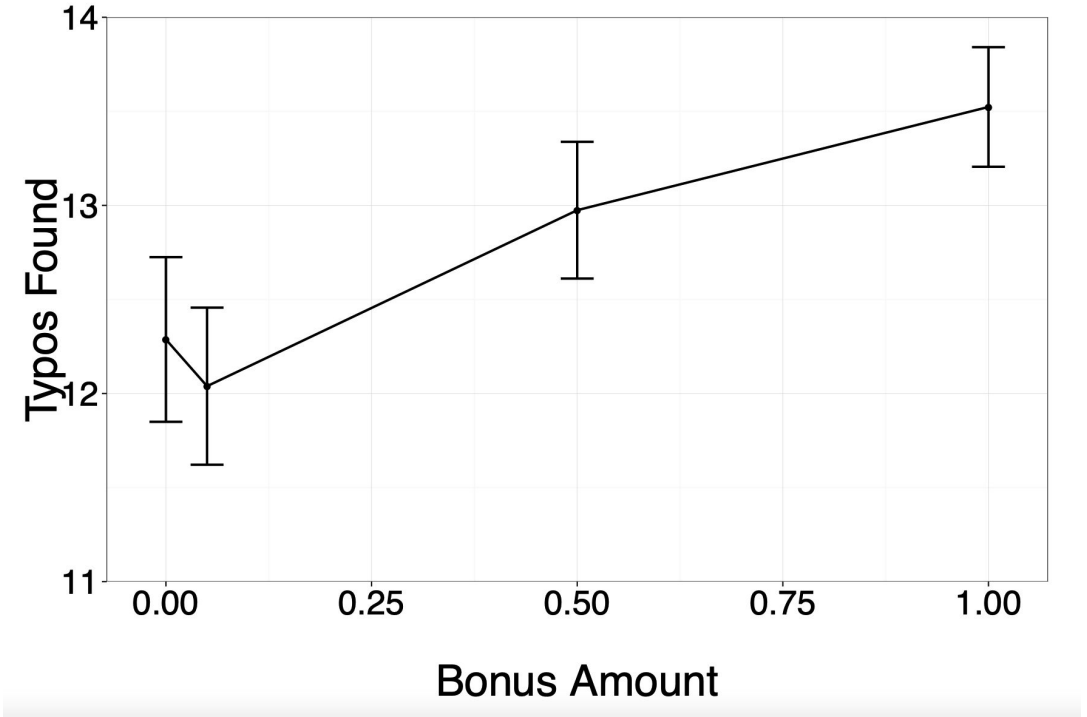
Used the same proofreading task with a base payment of $0.50.

Workers were assigned to treatments in which they could earn either :

$0.05, $0.50, or $1 if they found **75% of the typos** found by other workers.

Bonus Amounts: Results

PBPs lead to higher quality work only when the bonus is significantly large, but increasing the bonus amount has diminishing returns.



3. WHY DOES PBP WORK?

to verify that PBPs are useful in other tasks beyond finding typos.

to explore potential reasons why PBPs work.

Experiment: find identical images

Identical or Not?

- 20 pairs of images
- 10 pairs of identical images
- 10 pairs with minor differences

Experiment: Key properties

- spot more differences if workers spend more time
- work quality = #correct answered pairs

Experiment Design

Low Base

- base payment: $0.50
- no bonus

High Base:

- base payment: $1.50
- no bonus

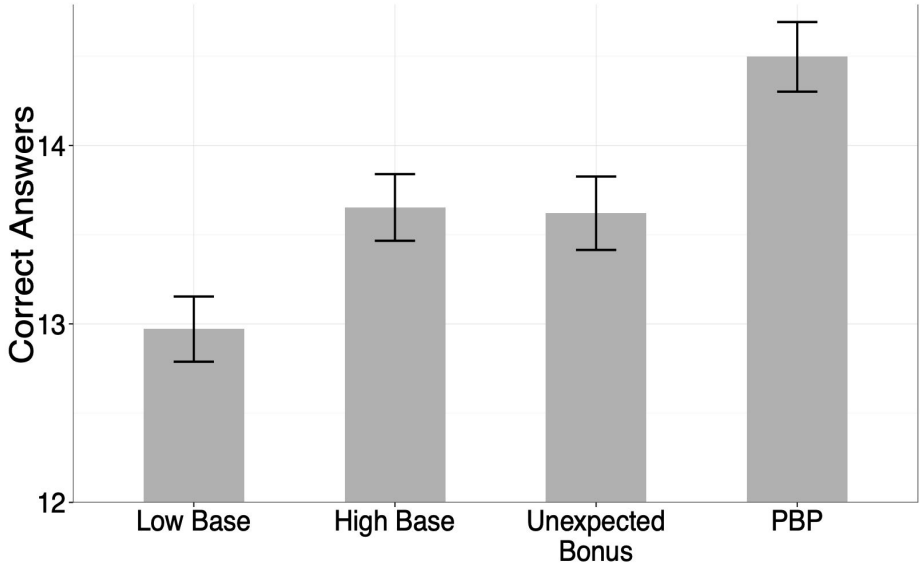
Unexpected Bonus:

- base payment: $0.50
- $1 bonus after accepting the HIT

PBP:

- base payment: $0.50
- $1 bonus for 80% correctness

Result


Correct Answers
14
13
12
Low Base
High Base
Unexpected Bonus
PBP

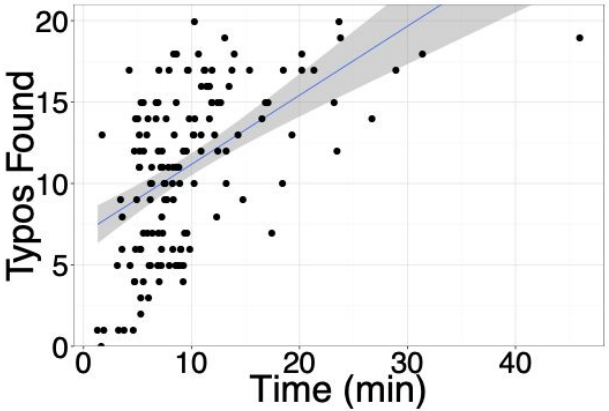
3. Where does PBP work?

- Whether can previous results generalize?
- What properties of a task open up the possibility of performance improvements with PBPs?

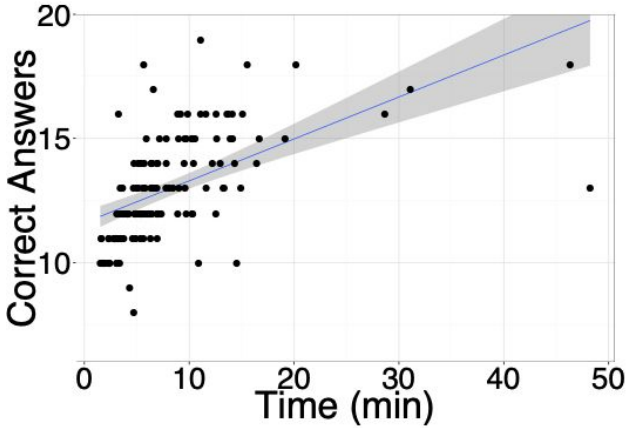
Effort-Responsive Task
=> Effort

PBPs work
= Quality

Previous Experiment Results



(a) Proofreading



(b) Spot-the-Difference

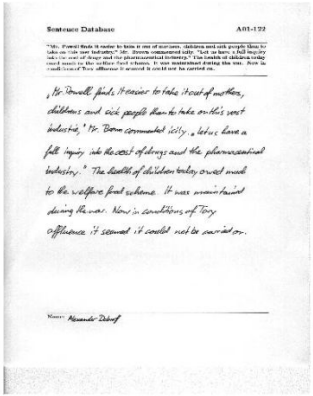
Effort-Responsive Task
=> Effort
=> Time Spent

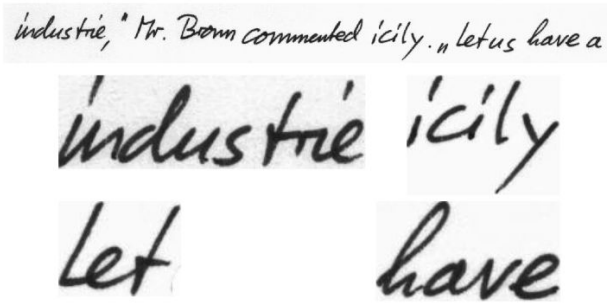
PBPs work
= Quality
= Typos Found / Correct Answers

Experiment: Handwriting Recognition

transcribe two images

- 89 words
- 74 words





IAM Handwriting Database

Experiment Design

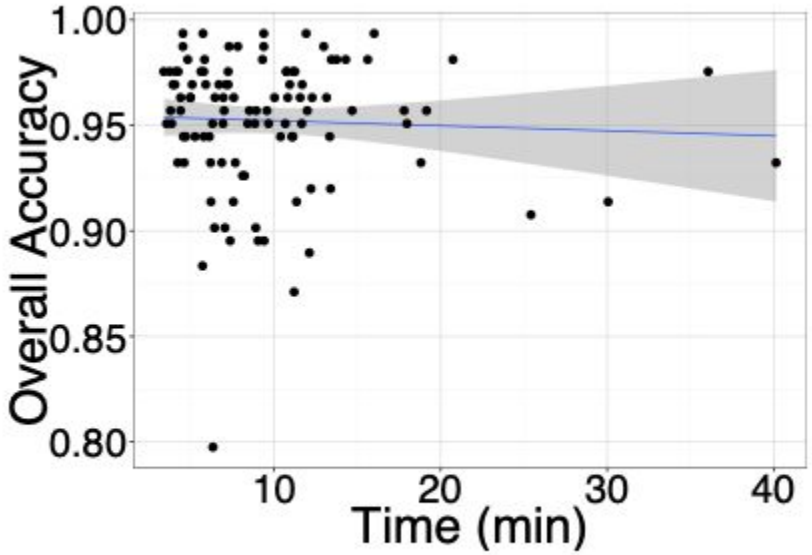
Control

- base payment: $0.50
- no bonus

PBP Treatment:

- base payment: $0.50
- $1 bonus for 90% correctness

Results



(c) Handwriting Rec.

Experiment: Audio Transcription

Transcribe:

- 10 audio clips
- approximately 5 seconds / clip
- heavy accents -> difficulty -> ceiling effect

Experiment Design

Control

- base payment: $0.50
- no bonus

PBP Treatment1 :

- base payment: $0.50
- $1 bonus for 80% correctness

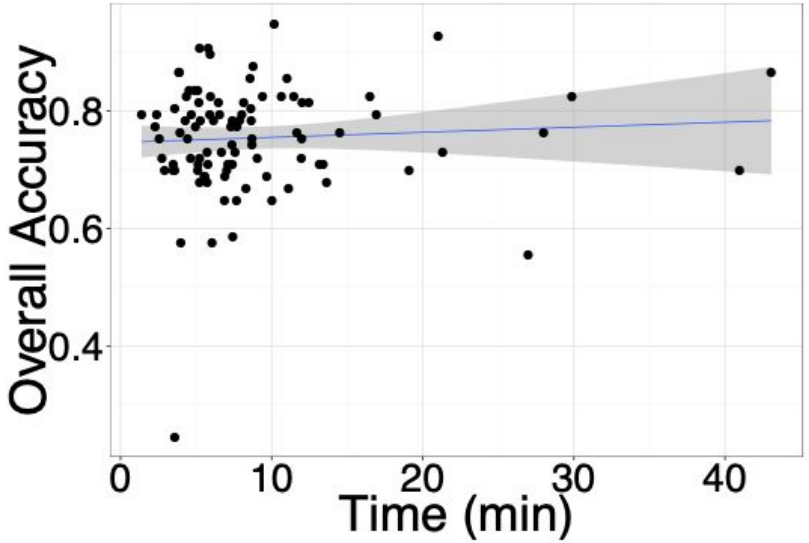
PBP Treatment2 :

- base payment: $0.50
- $1 bonus for 85% correctness

PBP Treatment1 :

- base payment: $0.50
- $1 bonus for 90% correctness

Results



(d) Audio Transcription

Discussion

1) To objectively reflect the impact of payment and bonus on the quality of work in the crowd working settings, what properties or characteristics do you think such tasks need to have? Please briefly explain why.

2) Can you think of other forms incentives besides financial payments that can be use to motivate desired result in crowdsourcing settings? Please briefly describe how will you design the incentives.

3) What are the factors that lead the first paper and the required paper to a different conclusion?

— “Financial Incentives and the “Performance of Crowds”

— “Incentivizing High Quality Crowdwork”

— “The Effects of Pay-to-Quit Incentives on Crowdworker Task Quality”

Paper 3: The Effects of Pay-to-Quit Incentives on Crowdworker Task Quality

Pay-to-Quit

Crowdsourcing

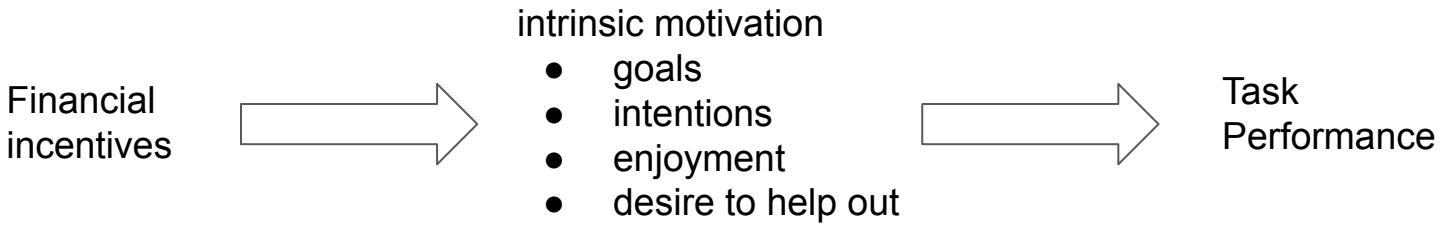
Pay-to-quit Examples

- Zappos
  - $4,000
  - 3%
- Amazon
  - $1,000 per year of employment and up to %5,000
- L18.com
  - one month’s salary
  - 6 / 85
- Riot Games
  - $25,000

Can pay-to-quit be applied for crowdsourcing?

- What factors may induce poorly-performing crowdworkers to quit
- Whether does pay-to-quit inducement and encourage poorly-performing workers to withdraw from the tasks early?
- Whether can pay-to-quit inducements still retain the high-performing workers?
- How are worker demographics related to incentive schemes?

Incentives & Task Performance



Anchoring Effect

Definition:
The anchoring effect is a cognitive bias that describes the common human tendency to rely too heavily on the first piece of information offered (the “anchor”) when making decisions. During decision making, anchoring occurs when individuals use an initial piece of information to make subsequent judgments. Once an anchor is set, other judgments are made by adjusting away from that anchor, and there is a bias toward interpreting other information around the anchor.

Examples:

- The initial price offered for a used car sets the standard for the rest of the negotiations, so that prices lower than the initial price seem more reasonable even if they are still higher than what the car is really worth.
- workers who were paid a small amount believed the value of their work to be greater, and were less motivated to produce quality than those doing it without compensation.

Risk Attitude

Risk attitude plays an essential role in evaluating incentives under uncertainty.

Loss aversion:

Pay-to-Quit Inducement Amounts

highly-performing workers:

- refuse the offer and continue to work

poorly-performing workers:

- accept the offer and quit the task early

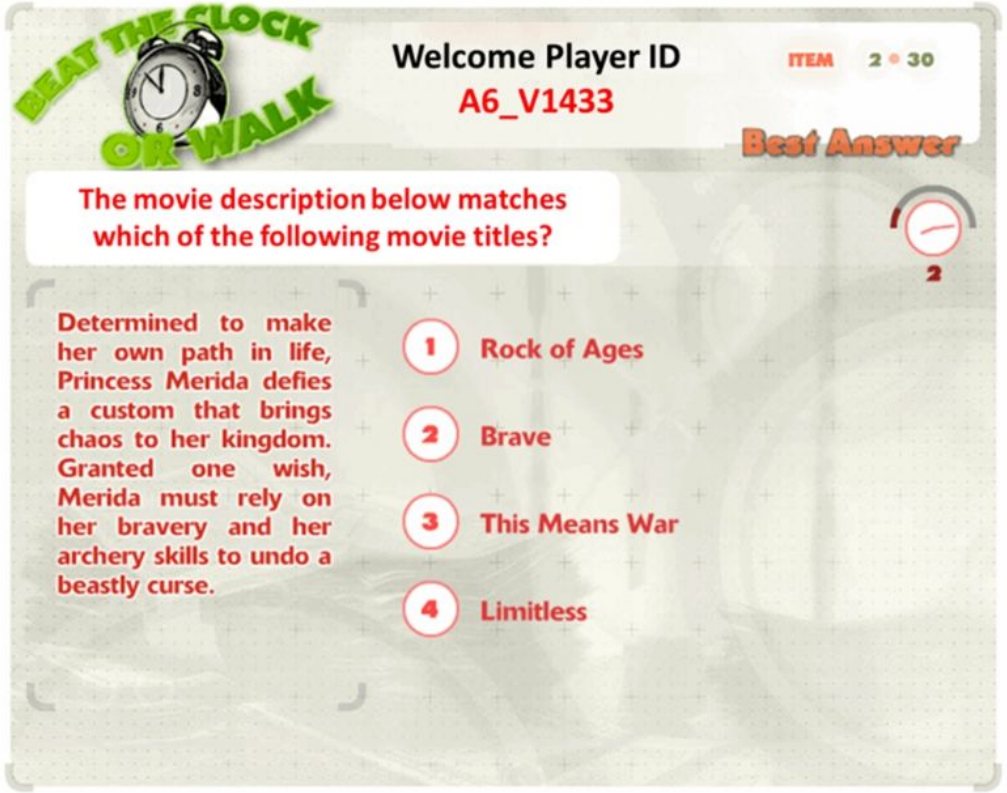
Research Questions

1. Does **explicit window** to accept pay-to-quit improve performance compared to permitting the right to leave at any time?
2. How frequent should **performance feedback** be provided to discriminate the choice to keep playing or the choice to quit?
3. How should **incentive bonus schemes** (additive or subtractive) be structured so that requesters only retain the highest-performing workers?
4. What **amount** should the bonus be to induce the poorly performing participants to quit?
5. Are there certain **demographics** that are more responsive to pay-to-quit incentives than others?

Experimental Design : Beat the Clock or Walk

Games vs Crowdsourcing

- less spam
- lower cost



Experimental Design : Participant Assignments

1200 MTurkers

Experiment 1 - 4:
- control group: 60 participants
- treatment: 60 participants

Experiment 5:
- control group: 120 participants
- treatment: 120 participants

30 questions = 1 Task

Question = Round

each round: 15 seconds

$0.30 for completing the task
$0.01 per question

Experimental Design : Performance Measurements

Mean Accuracy

$$Accuracy_i = \sum_{i=1}^{k} \frac{Correct_i}{Answered_i}$$

Experimental Design : Demographic Survey

Demographic Survey:

- gender
- country of residence
- income
- experience with games
- if crowdsourcing was their primary source of income

Game Experience Survey:

- if the time given was sufficient
- if the task was challenging enough
- (only for those who quit before completion) reasons to leave

Research Questions

1. Does **explicit window** to accept pay-to-quit improve performance compared to permitting the right to leave at any time?
2. How frequent should **performance feedback** be provided to discriminate the choice to keep playing or the choice to quit?
3. How should **incentive bonus schemes** (additive or subtractive) be structured so that requesters only retain the highest-performing workers?
4. What **amount** should the bonus be to induce the poorly performing participants to quit?
5. Are there certain **demographics** that are more responsive to pay-to-quit incentives than others?

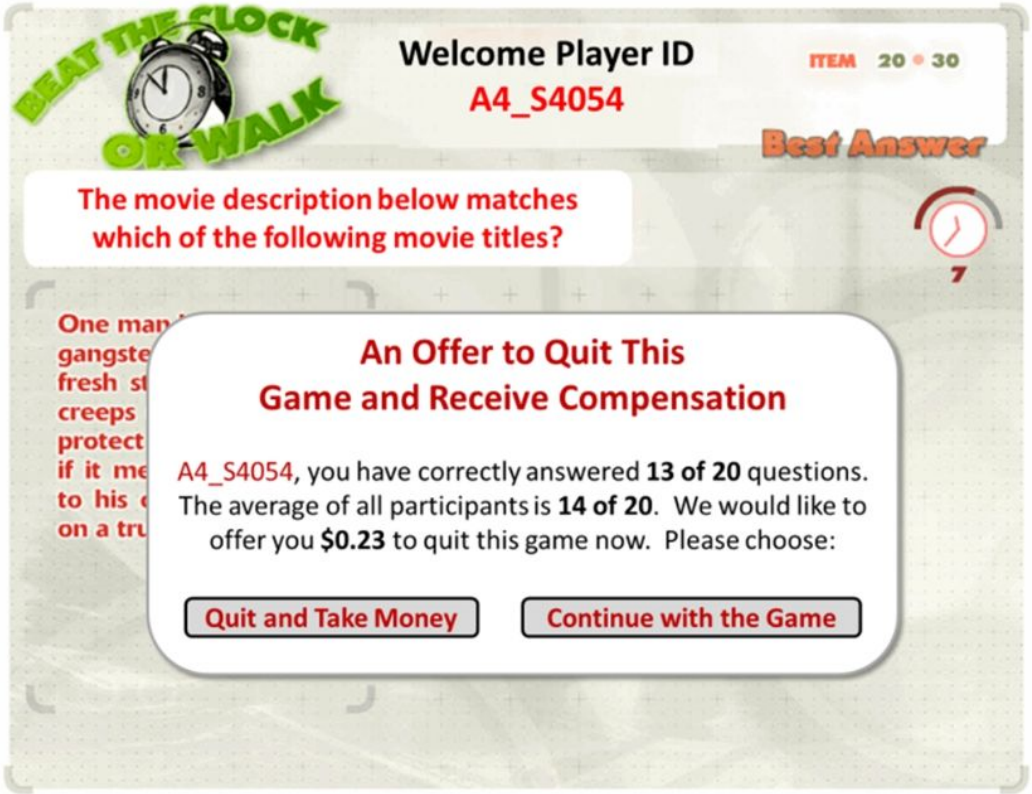
Experiment 1: Explicit Pay-to-Quit Offers

Control group (N=60)
- message before task
- quit at any time
- $0.01 / question

Treatment 1 (N=60)
- pop-up window at 15th question

Treatment 2 (N=60)
- pop-up window at 10th question
- pop-up window at 20th question

Treatment 3 (N=60)
- pop-up window at 7th question
- pop-up window at 14th question
- pop-up window at 21th question



Experiment 1: Explicit Pay-to-Quit Offers

Table 1. Mean accuracy and standard deviations for the control group (N=60) and three treatment groups (each N=60), each offering a different number of inducement prompts.

	Control		1 Prompt		2 Prompts		3 Prompts	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
All	.647	.175	.631	.187	**.708**	.135	.612	.165
Finishers	.653	.171	.639	.193	**.738**	.126	.610	.178
Quitters	.645	.194	.597	.164	**.575**	.111	.619	.095

two-prompt treatment is better for both finishers and quitters

Experiment 1: Explicit Pay-to-Quit Offers

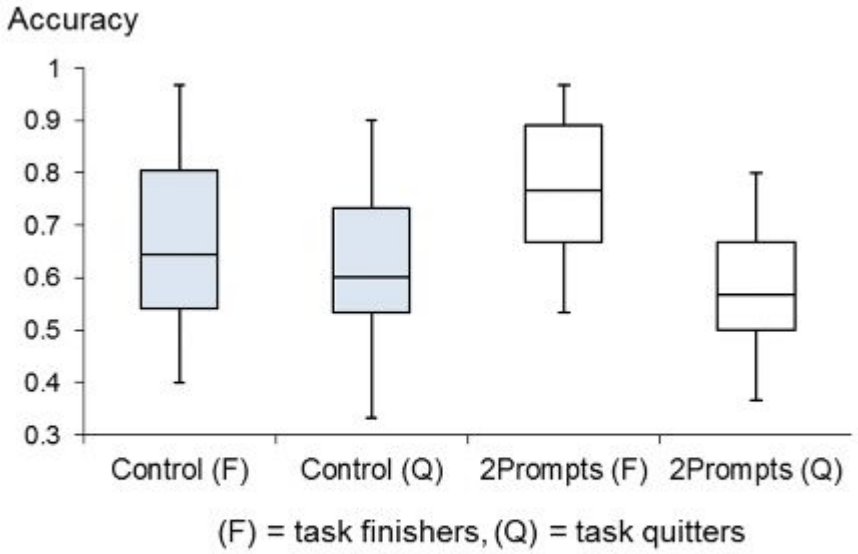


Figure 3. Box-plot comparison of mean accuracy between the control (light blue) and two-prompt group (unshaded)

two-prompt treatment is effective to

- retain the better performer
- eliminate the poorer performers

Research Questions

1. Does **explicit window** to accept pay-to-quit improve performance compared to permitting the right to leave at any time?
2. How frequent should **performance feedback** be provided to discriminate the choice to keep playing or the choice to quit?
3. How should **incentive bonus schemes** (additive or subtractive) be structured so that requesters only retain the highest-performing workers?
4. What **amount** should the bonus be to induce the poorly performing participants to quit?
5. Are there certain **demographics** that are more responsive to pay-to-quit incentives than others?

Experiment 2: Performance Feedback

Control group (N=60)
- feedback at the end of the task

Treatment 1 (N=60)
- feedback after 10th and 20th with quit option

Treatment 2 (N=60)
- feedback after each of the 30 questions

Treatment 3 (N=60)
- quit pop-up at 10th and 20th
- feedback after correct answer
- content: outperform the average

Treatment 4 (N=60)
- quit pop-up at 10th and 20th
- feedback after correct answer
- content: outperform the average

Feedback:
- number of correct questions
- average correct number

Experiment 2: Performance Feedback

Table 2. Mean accuracy and standard deviations for the control group (N=60) and two treatment groups (N=60), each offering a different frequency of feedback to the participant.

	Control		Each Round		Each Prompt	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**
All	.629	.159	.649	.176	**.686**	.153
Finishers	.636	.157	.668	.177	**.712**	.151
Quitters	.603	.167	.583	.171	.559	.159

each-prompt treatment is better for both finishers and quitters

Experiment 2: Performance Feedback

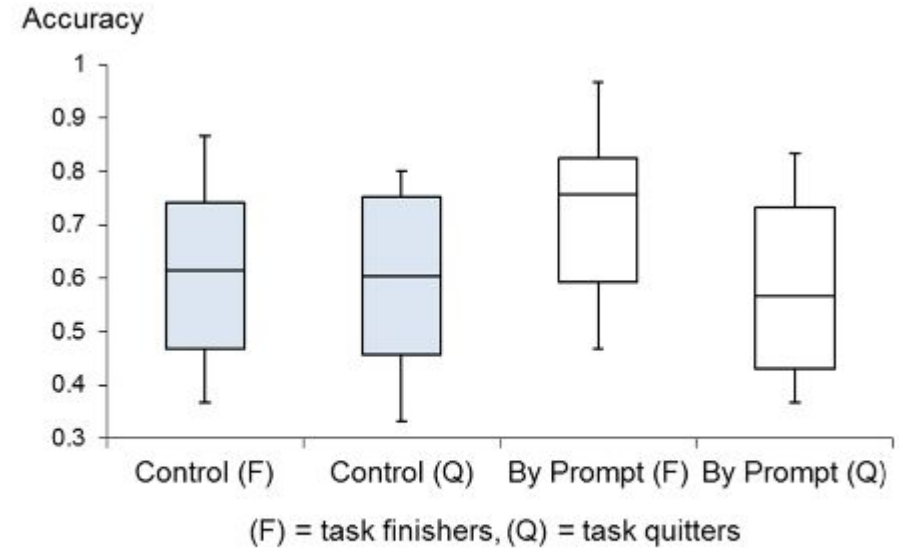


Figure 4. Box-plot comparison of mean accuracy between the control (light blue) and feedback after each prompt (unshaded).

Similar result as the box-plot of previous experiment

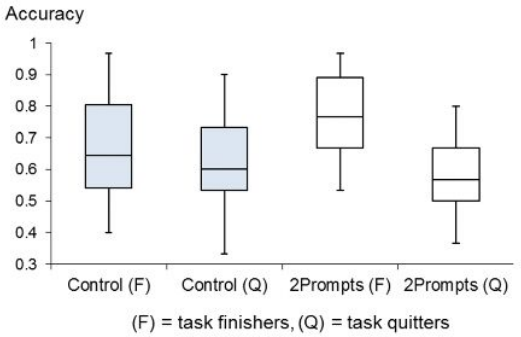


Figure 3. Box-plot comparison of mean accuracy between the control (light blue) and two-prompt group (unshaded)

Experiment 2: Performance Feedback

Table 3. Mean accuracy and standard deviations for three treatment groups (each N=60), each providing a different feedback message during the pay-to-quit incentive prompt.

	Accurate		Outperform		Underperform	
Group	M	SD	M	SD	M	SD
All	.658	.153	**.717**	.157	.627	.151
Finishers	.712	.151	.727	.155	.695	.155
Quitters	.559	.159	**.698**	.167	.502	.143

0.727 - 0.717 = 0.01

- over-encouraging feedback does not persuade the best performers to improve their performance any further

0.695 - 0.627 = 0.07

- underperforming feedback weeds out poorly-performing participants

outperform feedback treatment is better for both finishers and quitters

Experiment 2: Performance Feedback

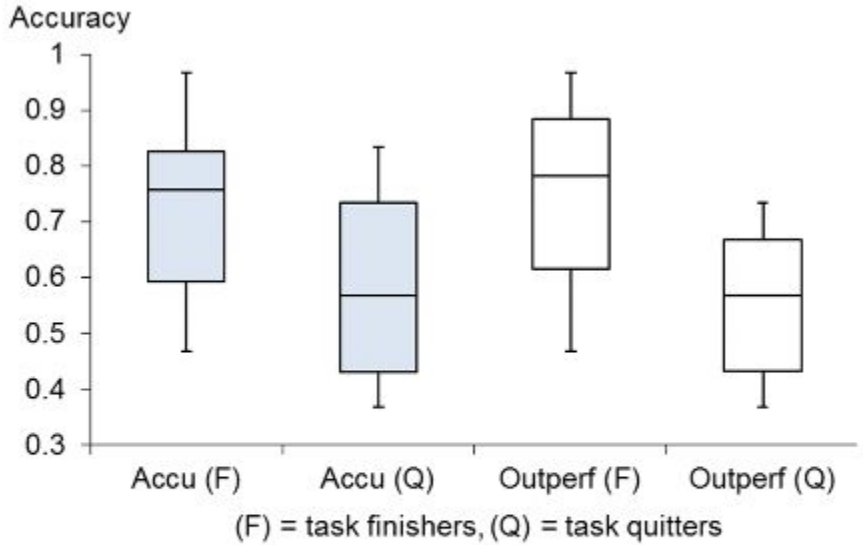


Figure 5. Box-plot comparison of mean accuracy between those accurately reporting task performance (light blue) and those told they are outperforming the group average (unshaded).

Research Questions

1. Does **explicit window** to accept pay-to-quit improve performance compared to permitting the right to leave at any time?
2. How frequent should **performance feedback** be provided to discriminate the choice to keep playing or the choice to quit?
3. How should **incentive bonus schemes** (additive or subtractive) be structured so that requesters only retain the highest-performing workers?
4. What **amount** should the bonus be to induce the poorly performing participants to quit?
5. Are there certain **demographics** that are more responsive to pay-to-quit incentives than others?

Experiment 3: Incentive Type

Control group (N=60)
- two pay-to-quit windows
- $0.01 for every question answered

Treatment 1 (N=60) **Increasing Incentive**
- two pay-to-quit windows
- $0.01 for every question answered
- $0.01 per correct answer

Treatment 2 (N=60) **Decreasing Incentive**
- two pay-to-quit windows
- **amount = 0.5 * (30 + #incorrect answers - #answered)**
- induce participants to quit early if they initially perform poorly

Experiment 3: Incentive Type

Table 4. Mean accuracy and standard deviations for a control group (N=60) and two treatment groups (each N=60), each offering a different incentive type.

	Control		Increasing		Decreasing	
Group	M	SD	M	SD	M	SD
All	.648	.156	.682	.171	.653	.151
Finishers	.656	.152	.707	.172	.699	.149
Quitters	.618	.170	.573	.164	.525	.158

Decreasing incentive is the treatment with the largest discrimination power

Research Questions

1. Does **explicit window** to accept pay-to-quit improve performance compared to permitting the right to leave at any time?
2. How frequent should **performance feedback** be provided to discriminate the choice to keep playing or the choice to quit?
3. How should **incentive bonus schemes** (additive or subtractive) be structured so that requesters only retain the highest-performing workers?
4. What **amount** should the bonus be to induce the poorly performing participants to quit?
5. Are there certain **demographics** that are more responsive to pay-to-quit incentives than others?

Experiment 4: Incentive Size

Control group (N=60)
- two pay-to-quit windows
- $0.01 for every question answered

Treatment 1 (N=60) **Smaller Decreasing Incentive**
- two pay-to-quit windows
- amount = **0.5** * (30 + #incorrect answers - #answered)

Treatment 2 (N=60) **Moderate Decreasing Incentive**
- two pay-to-quit windows
- amount = **0.75** * (30 + #incorrect answers - #answered)

Treatment 2 (N=60) **Large Decreasing Incentive**
- two pay-to-quit windows
- amount = **1** * (30 + #incorrect answers - #answered)

Experiment 4: Incentive Size

Table 5. Mean accuracy and standard deviations for the control group (N=60) and two treatment groups, (each N=60), each offering a different incentive size.

	Control		Small (0.5)		Med (0.75)		Large (1.0)	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
All	.649	.154	.673	.159	.679	.141	.681	.139
Finishers	.651	.156	.691	.161	**.719**	.144	**.722**	.138
Quitters	.621	.146	.599	.149	.568	.131	.569	.142

medium-sized and large incentives have the best discriminatory power between finishers and quitters

large incentive treatment is slightly better

Experiment 4: Incentive Size

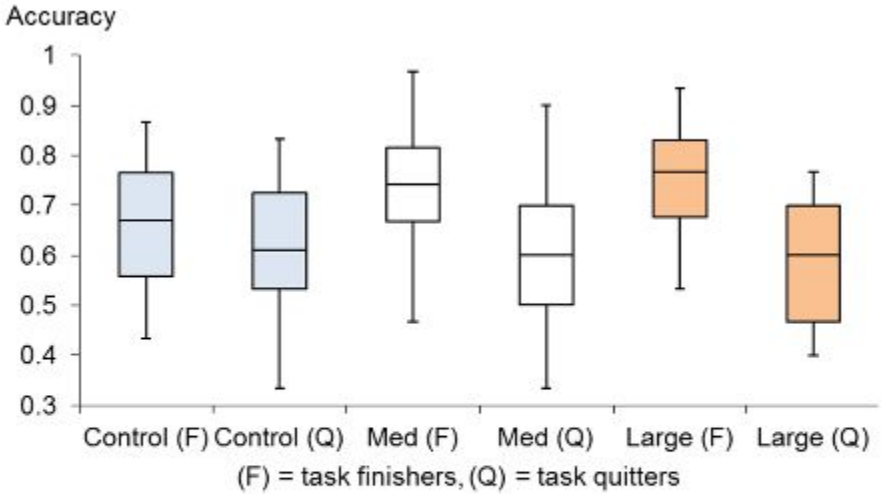


Figure 6. Box-plot comparison of mean accuracy between our control group (shaded light blue, on left), those offered a medium-sized incentive (unshaded, center), and those offered a large incentive (shaded orange, right).

Research Questions

1. Does **explicit window** to accept pay-to-quit improve performance compared to permitting the right to leave at any time? Yes, prompt after 10th and 20th questions
2. How frequent should **performance feedback** be provided to discriminate the choice to keep playing or the choice to quit? During the prompt
3. How should **incentive bonus schemes** (additive or subtractive) be structured so that requesters only retain the highest-performing workers? Decreasing incentive
4. What **amount** should the bonus be to induce the poorly performing participants to quit? Moderate decreasing incentive
5. Are there certain **demographics** that are more responsive to pay-to-quit incentives than others?

Experiment 5: Putting it All Together

Best model:

- two pop-up windows after 10th and 20th questions
- accurate feedback
- decreasing incentive
- moderate amount = **0.75** * (30 + #incorrect answers - #answered)

Participants:

- 120 for control group
- 120 for best strategy group

Experiment 5: Putting it All Together

Table 6. Mean accuracy and standard deviations comparing a control group (N=120) and a treatment group (N=120) comprised of the best strategies from Experiments 1 through 4

	Control		Single Best Strategy	
Group	**M**	**SD**	**M**	**SD**
All	.637	.146	**.687**	.133
Finishers	.642	.145	**.717**	.133
Quitters	.618	.151	.578	.131

mean accuracy gain from the finishers:
from 0.642 - 0.637 = 0.05 of control group
to 0.717 - 0.687 = 0.3 of best strategy group

Experiment 5: Putting it All Together

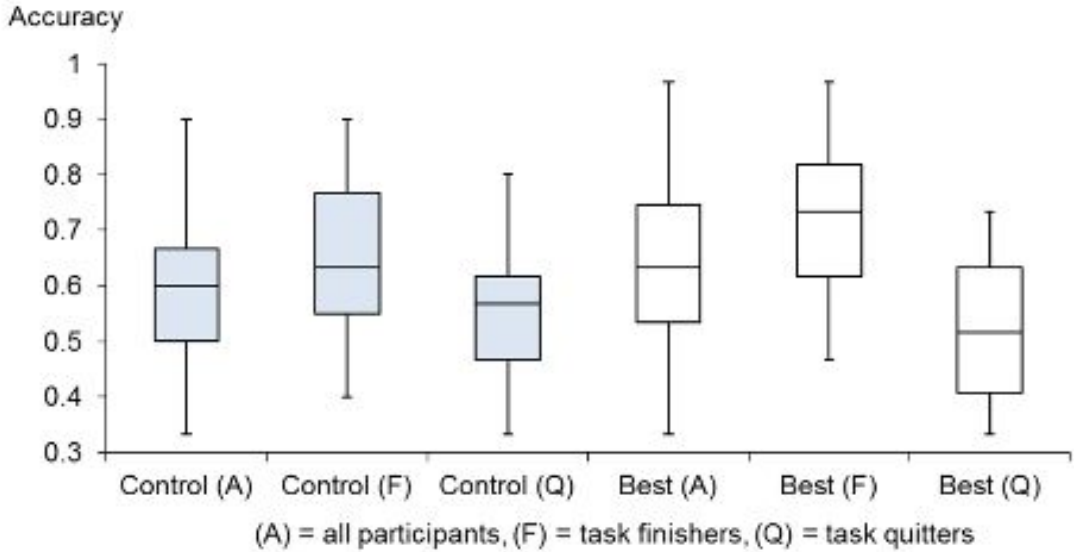


Figure 7. Box-plot comparison of mean accuracy between our baseline (Control) and the combination of treatments from Experiments 1 to 4 that provided the best performance (Best).

The distinction between the finishers and quitters is greatest

Research Questions

1. Does **explicit window** to accept pay-to-quit improve performance compared to permitting the right to leave at any time?
2. How frequent should **performance feedback** be provided to discriminate the choice to keep playing or the choice to quit?
3. How should **incentive bonus schemes** (additive or subtractive) be structured so that requesters only retain the highest-performing workers?
4. What **amount** should the bonus be to induce the poorly performing participants to quit?
5. Are there certain **demographics** that are more responsive to pay-to-quit incentives than others?

Demographics

- Who are **more likely** to accept pay-to-quit offer?
  - females
  - South Asian residents
  - people earning higher wages
- Greater risk aversion
- Greater opportunity cost for their time

- Who are **less likely** to accept pay-to-quit offer?
  - play games regularly
  - depend on income from crowdsourcing
  - workers in the lowest income band

Discussion

1）What is the underlying assumption of pay-to-quit incentive scheme to obtain increases in productivity?

2）What could the possible limitation to this study be?