Assessing Crowdsourcing Quality through Objective Tasks

Ahmet Aker[◊], Mahmoud El-Haj[*],
Dyaa Albakour[*], Udo Kruschwitz[*]

[◊]Department of Computer Science, University of Sheffield
[*]School of Computer Science and Electronic Engineering, University of Essex
Email: a.aker@dcs.shef.ac.uk, melhaj@essex.ac.uk, malbak@essex.ac.uk,
udo@essex.ac.uk

Overview

CrowedSourcing:

- Collect vast quantities of human assessments.
- Collect annotations rapidly without the need of an expert.
- Crowdsourcing: an alternative in creating resources for NLP.
- Obtaining reliable results from the crowd remains a challenge.

Objectives

- Investigate factors which can influence the quality of results from the crowd.

Objectives

- Investigate factors which can influence the quality of results from the crowd.
- Investigate the impact of different factors and presentation methods.

Objectives

- Investigate factors which can influence the quality of results from the crowd.
- Investigate the impact of different factors and presentation methods.
- We run two different experiments using objective tasks: maths and general text questions

Objectives

- Investigate factors which can influence the quality of results from the crowd.
- Investigate the impact of different factors and presentation methods.
- We run two different experiments using objective tasks: maths and general text questions
- We present our results comparing the influence of the different factors used.

Experimental Setup

Objective Tasks:

Experimental Setup

Objective Tasks:

- In our experiments we use maths questions and general text questions (travel and history categories).

Experimental Setup

Objective Tasks:

- In our experiments we use maths questions and general text questions (travel and history categories).
- In both tasks the answers are unique, which eliminates the uncertainty.

Experimental Setup

Objective Tasks:

- In our experiments we use maths questions and general text questions (travel and history categories).
- In both tasks the answers are unique, which eliminates the uncertainty.
- We investigated the impact of the following variables on the quality of the results:

Experimental Setup

Objective Tasks:

- In our experiments we use maths questions and general text questions (travel and history categories).
- In both tasks the answers are unique, which eliminates the uncertainty.
- We investigated the impact of the following variables on the quality of the results:
  1. Presentation method: free text vs radio buttons.

Experimental Setup

Objective Tasks:

- In our experiments we use maths questions and general text questions (travel and history categories).
- In both tasks the answers are unique, which eliminates the uncertainty.
- We investigated the impact of the following variables on the quality of the results:
  1. Presentation method: free text vs radio buttons.
  2. Workers' base: US or India.

Experimental Setup

Objective Tasks:

- In our experiments we use maths questions and general text questions (travel and history categories).
- In both tasks the answers are unique, which eliminates the uncertainty.
- We investigated the impact of the following variables on the quality of the results:
  1. Presentation method: free text vs radio buttons.
  2. Workers' base: US or India.
  3. Payment scale: an estimated $4, $8 and $10 per hour.

Experimental Setup

- For each run we assessed the results provided by 25 workers on a set of 10 tasks.

Experimental Setup

- For each run we assessed the results provided by 25 workers on a set of 10 tasks.
- Limitations on the workers origins: we include only the two selected countries, i.e. US and India.

Experimental Setup

- For each run we assessed the results provided by 25 workers on a set of 10 tasks.
- Limitations on the workers origins: we include only the two selected countries, i.e. US and India.
- No limitation on the confidence rate (real workers and spammers).

Maths Questions

- Word problems, collected from different online learning sites.

Maths Questions

- Word problems, collected from different online learning sites.
- The level of the questions vary from primary school up to the 6th grade (relatively easy).

Maths Questions

- Word problems, collected from different online learning sites.
- The level of the questions vary from primary school up to the 6th grade (relatively easy).
- In total we have 10 such questions. The questions vary in text length (min: 4, max: 75 and ave: 40 words).

Maths Question Example

Table: Short and an average example maths question.

What is double 80?
There was a fire in the building down the street. It was so large that our city had to call in 6 fire trucks. Each truck had 9 firemen riding on it. How many firemen arrived to fight the fire?

General Text Questions

- A number of 10 multiple choice questions have been selected for this experiment.

General Text Questions

- A number of 10 multiple choice questions have been selected for this experiment.
- The questions fall in the travel and history categories (one correct answer).

General Text Questions

- A number of 10 multiple choice questions have been selected for this experiment.
- The questions fall in the travel and history categories (one correct answer).
- Not straightforward, the answer is not easily derivable from the text itself.

General Text Questions

- A number of 10 multiple choice questions have been selected for this experiment.
- The questions fall in the travel and history categories (one correct answer).
- Not straightforward, the answer is not easily derivable from the text itself.
- Requires some general knowledge or the willingness to search the web.

General Text Questions

- A number of 10 multiple choice questions have been selected for this experiment.
- The questions fall in the travel and history categories (one correct answer).
- Not straightforward, the answer is not easily derivable from the text itself.
- Requires some general knowledge or the willingness to search the web.
- The length of the questions in average is 13 words with a maximum of 29 and a minimum of 3 words.

General Text Question Example

Table: Example of History and Travel questions

Genre	Questions	Answers
Travel	Which country is also called the Hellenic Republic?	(A)Sweden, (B)Denmark, (C)Greece, (D)Finland.
History	What U.S. president was born William Jefferson Blythe IV?	(A)Richard-Nixon, (B)Bill-Clinton, (C)Andrew-Johnson, (D)Grover_Cleveland.

Experimental Design

- For each question type we use two different designs:

Experimental Design

- For each question type we use two different designs:
  - Radio Buttons.
  - Free text input.

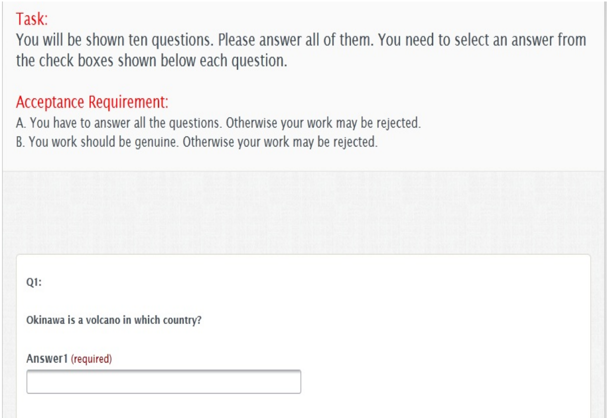
Experimental Design

- For each question type we use two different designs:
  - Radio Buttons.
  - Free text input.
- In each HIT we show 10 questions and asked for 25 workers.

Experimental Design

- For each question type we use two different designs:
  - Radio Buttons.
  - Free text input.
- In each HIT we show 10 questions and asked for 25 workers.
- Workers are supposed to write the answer or select one of the provided choices.

General Text Questions Design



Task:
You will be shown ten questions. Please answer all of them. You need to select an answer from the check boxes shown below each question.

Acceptance Requirement:
A. You have to answer all the questions. Otherwise your work may be rejected.
B. You work should be genuine. Otherwise your work may be rejected.

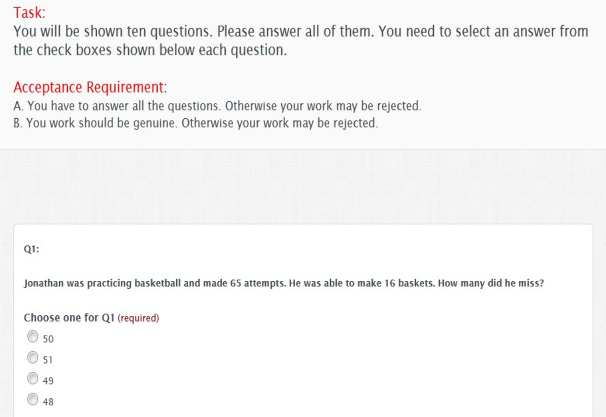
Q1:

Okinawa is a volcano in which country?

Answer1 (required)

Figure: Text questions with free text design.

Math Questions Design



Task:

You will be shown ten questions. Please answer all of them. You need to select an answer from the check boxes shown below each question.

Acceptance Requirement:

A. You have to answer all the questions. Otherwise your work may be rejected.

B. You work should be genuine. Otherwise your work may be rejected.

Q1:

Jonathan was practicing basketball and made 65 attempts. He was able to make 16 baskets. How many did he miss?

Choose one for Q1 (required)

○ 50

○ 51

○ 49

○ 48

Figure: Maths questions with radio button design.

Results

- In each experiment we count the number of correct answers.
- This means that every experiment has 25 such fields (different worker).

Math Question Results

USA	Average Score	India	Average Score
$4_RB	9.88	$4_RB	8.30
$8_RB	9.64	$8_RB	9.16
$10_RB	9.44	$10_RB	9.80
$4_TF	9.28	$4_TF	8.24
$8_TF	9.28	$8_TF	8.52
$10_TF	9.40	$10_TF	9.44

Table: Average scores of the math questions.

General Text Question Results

USA	Average Score	India	Average Score
$4_RB	9.25	$4_RB	8.23
$8_RB	9.30	$8_RB	7.69
$10_RB	9.07	$10_RB	9.42
$4_TF	9.32	$4_TF	8.12
$8_TF	8.88	$8_TF	9.40
$10_TF	9.00	$10_TF	9.29

Table: Average scores of the general text questions.

Findings (Math Questions)

- The results tend to be generally better with radio button design.

Findings (Math Questions)

- The results tend to be generally better with radio button design.
- For the workers from India we can see that the higher the payment the better results.

Findings (Math Questions)

- The results tend to be generally better with radio button design.
- For the workers from India we can see that the higher the payment the better results.
- There is no statistically measurable impact of the country of workers' origin on the quality of the results.

Findings (Math Questions)

- The results tend to be generally better with radio button design.
- For the workers from India we can see that the higher the payment the better results.
- There is no statistically measurable impact of the country of workers' origin on the quality of the results.
- The design and the payment do in some cases have a significant impact on the quality of the results.

Findings (Math Questions)

- The results tend to be generally better with radio button design.
- For the workers from India we can see that the higher the payment the better results.
- There is no statistically measurable impact of the country of workers' origin on the quality of the results.
- The design and the payment do in some cases have a significant impact on the quality of the results.
- When the radio button design is used the results can be significantly better.

Findings (Math Questions)

- The results tend to be generally better with radio button design.
- For the workers from India we can see that the higher the payment the better results.
- There is no statistically measurable impact of the country of workers' origin on the quality of the results.
- The design and the payment do in some cases have a significant impact on the quality of the results.
- When the radio button design is used the results can be significantly better.
- The payment incentives seem to have also a significant positive impact on the results.

Findings (Math Questions)

- The results tend to be generally better with radio button design.
- For the workers from India we can see that the higher the payment the better results.
- There is no statistically measurable impact of the country of workers' origin on the quality of the results.
- The design and the payment do in some cases have a significant impact on the quality of the results.
- When the radio button design is used the results can be significantly better.
- The payment incentives seem to have also a significant positive impact on the results.
- This does not confirm Mason and Duncan [2] (improved the quantity, but not the quality).

Findings (General Text Questions)

- Only some significant impact of the country of origin.

Findings (General Text Questions)

- Only some significant impact of the country of origin.
- The payment tends to be a significant factor in the quality of the results.

Findings (General Text Questions)

- Only some significant impact of the country of origin.
- The payment tends to be a significant factor in the quality of the results.
- Participants tend to make more effort in solving the questions when higher payments are made

Discussions

- Small Sample (questions, users, difficult to generalise).

Discussions

- Small Sample (questions, users, difficult to generalise).
- Average Score is high (very difficult questions?).

Discussions

- Small Sample (questions, users, difficult to generalise).
- Average Score is high (very difficult questions?).
- Focusing at individual performance (get the best out of individuals).

Discussions

- Small Sample (questions, users, difficult to generalise).
- Average Score is high (very difficult questions?).
- Focusing at individual performance (get the best out of individuals).
- Serve as a baseline (comparison for future experiments).

Questions

Questions?

Experimental Pair
Impact of country of origin
nil
Impact of design
USA_4_TF – USA_4_RB
USA_8_TF – USA_8_RB
India_10_TF – India_10_RB
Impact of payment
India_4_TF – India_10_TF
India_8_RB – India_10_RB

Table: Results of the maths question. The significantly better (at level p < 0.05) results are on the right of "–". "nil" indicates the absence of any significantly different result. TF stands for text field design and RB for the radio button design. 4, 8 and 10 are the payments per hour in US dollars.

Experimental Pair
Impact of country of origin
India_4_TF – USA_4_TF
India_8_RB – USA_8_RB
Impact of design
India_8_RB – India_8_TF
Impact of payment
India_4_TF – India_8_TF
India_4_RB – India_10_RB
India_8_RB – India_10_RB
India_4_TF – India_10_TF

Table: Results of the general text question. The significantly better (at level $p < 0.05$) results are on the right of "–". "nil" indicates the absence of any significantly different result. TF stands for text field design and RB for the radio button design. 4, 8 and 10 are the payments per hour in US dollars.

- Donghui Feng, Sveva Besana, and Remi Zajac.
  Acquiring High Quality Non-expert Knowledge from On-demand Workforce.
  In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, pages 51–56, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

- Winter Mason and Duncan J. Watts.
  Financial Incentives and the "Performance of Crowds".
  SIGKDD Explor. Newsl., 11:100–108, May 2010.