Middle East Technical University
Informatics Institute

VALIDATING COST OF QUALITY MODELS IN SUBJECTIVE NON-DETERMINISTIC MICROTASK CROWDSOURCING

Advisor: Prof. Dr. Semih Bilgen
(METU)

Deniz İren
(IS-IS)

December 2013

TECHNICAL REPORT
METU/II-TR-2013-

Orta Doğu Teknik Üniversitesi
Enformatik Enstitüsü

KALİTE MALİYET MODELLERİNİN ÖZNEL GEREKİRCİ OLMAYAN KİTLE KAYNAKLI ÇALIŞMA İLE DOĞRULANMASI

Danışman: Prof. Dr. Semih Bilgen
(ODTÜ)

Deniz İren
(IS-IS)

Aralık 2013

TEKNİK RAPOR
ODTÜ/EE-TR-2013-

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY (Internal Use)	**2. REPORT DATE** 31.12.2013

3. TITLE AND SUBTITLE

VALIDATING COST OF QUALITY MODELS IN SUBJECTIVE NON-DETERMINISTIC MICROTASK CROWDSOURCING

4. AUTHOR (S)	**5. REPORT NUMBER (Internal Use)**
Deniz İren	**METU/II-TR-2013-**

6. SPONSORING/ MONITORING AGENCY NAME(S) AND SIGNATURE(S)

Information Systems Programme, Department of Information Systems, Informatics Institute, METU
Advisor: Semih Bilgen Signature:

7. SUPPLEMENTARY NOTES

8. ABSTRACT (MAXIMUM 200 WORDS)

This technical report provides an overview of cost models of major quality assurance mechanisms which are used in crowdsourcing along with a brief description of cost of quality approach to cost analysis. An experiment was conducted aiming at verifying cost of quality models and comparing quality levels offered by various quality assurance mechanisms. These mechanisms were used to detect poor quality contributions made for subjective microtasks.

9. SUBJECT TERMS	**10. NUMBER OF PAGES**
Crowdsourcing, Cost of Quality, Project Management	**36**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Validating Cost of Quality Models in Subjective Non-Deterministic Microtask Crowdsourcing

Deniz İren, Semih Bilgen

Middle East Technical University

Informatics Institute

diren@metu.edu.tr | bilgen@metu.edu.tr

Abstract

This technical report provides an overview of cost models of major quality assurance mechanisms which are used in crowdsourcing along with a brief description of cost of quality approach to cost analysis. An experiment was conducted aiming at verifying cost of quality models and comparing quality levels offered by various quality assurance mechanisms. These mechanisms were used to detect poor quality contributions made for subjective microtasks.

Keywords: Crowdsourcing, Cost of Quality, Project Management;

1. INTRODUCTION

Crowdsourcing is a process of value creation as a result of an outsourcing initiative, in which the interactive features of the Internet are utilized, by generally an anonymous mass consisting of individuals voluntarily choosing the task to work on. There is neither a well-defined pact nor a service level agreement between the individuals and the employer. Anonymity and the diverse skill set of the crowd combined with the gain-seeking nature of the individuals provide the tendency to low quality results and lack of control in production process. Various quality assurance mechanisms have been in use to deal with this problem. However using quality assurance mechanisms increases project costs and in certain cases causes inefficiencies.

This technical report briefly introduces major quality assurance mechanisms and cost of quality models associated with them. The cost models are verified in an experimental setting in which users are asked to draw and submit images of lizards. Image validity is controlled by separate groups of contributors. Both drawing and control tasks are subject to quality assurance.

Section 1 briefly defines the problem of quality assurance in crowdsourcing. Section 2 provides an overview of quality assurance mechanisms and cost of quality in crowdsourcing. Section 3 describes the experiment setting along with the goals, the procedure and the data to be collected. Section 4 presents the results and observations. Finally Section 5 discusses the results and concludes the report.

2. QUALITY ASSURANCE MECHANISMS AND COST OF QUALITY

2.1 Crowdsourcing Quality Assurance Mechanisms

Work products produced by crowds may fail to comply with the criteria of acceptable quality. This is either because of the erroneous submissions made by individuals or because of a willing act to cheat the system. In any case, crowdsourcing systems should be designed to incorporate certain quality assurance mechanisms to achieve the desired end product quality.

There are many different types of quality assurance techniques utilized in crowdsourcing. A broad categorization which was made according to the characteristics of these mechanisms is provided in Figure 1 (Iren, 2013). These mechanisms offer varying range of effectiveness, applicability and have different costs.

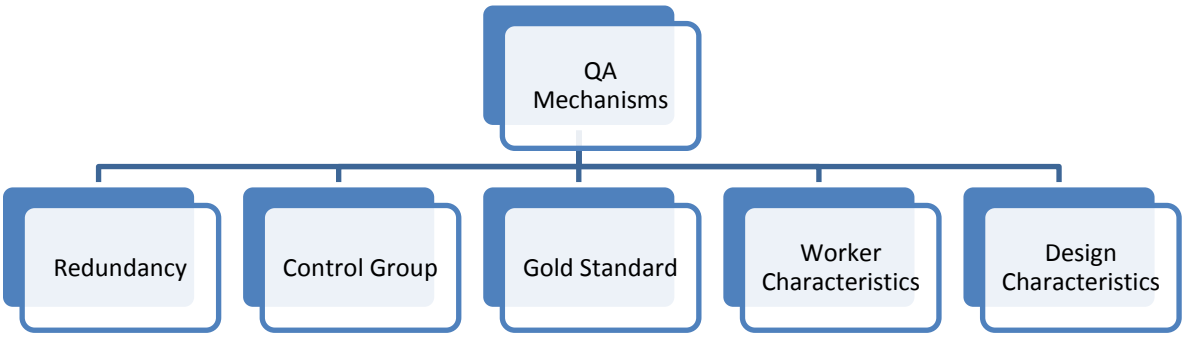


Figure 1: Categories of crowdsourcing quality assurance mechanisms

Redundancy quality assurance mechanisms involve assigning multiple instances of the same microtask to workers and aggregating the results. In **control group** quality assurance process, there exists a different group of workers which controls the outputs of the primary tasks. In **gold standard** quality assurance mechanisms, a set of tasks with predefined results are inserted to the system and the some of the contributions of workers are evaluated by comparing against these expected results. **Worker characteristics** type of quality assurance mechanisms depend on workers' history of submissions, the skills they possess or characteristics they have. Certain mechanisms in which poor quality is prevented or detected via design characteristics of the crowdsourcing system, such as better usability, robust design or statistically correcting submission biases, are categorized as **design characteristics**.

2.2 Crowdsourcing Cost of Quality

Cost does not only mean monetary costs, but also the effort spent. It should be noted that even if the work involves no monetary payment, and a crowd is performing tasks for some other reason, workforce remains a scarce resource. Deciding to utilize effort for quality assurance purposes rather than performing new tasks introduces an opportunity cost.

Cost of Quality (CoQ) is defined as the total cost of all quality related activities. CoQ is expressed as the sum of conformance and non-conformance costs. Conformance costs are associated with prevention of poor quality, non-conformance costs occur due to poor quality (Crosby, 1979).

2.3 Cost Models

Cost models are developed for major crowdsourcing quality assurance mechanisms by using a cost of quality analysis approach (Iren, 2013). Models include probabilistic parameters such as cheat probabilities and impact values such as rework costs. In order to use the models effectively, practitioners should know, or at least estimate these probabilities. These probabilistic values depend on various parameters including crowd characteristics, work type and motivational tools utilized. Currently the literature lacks information which may be used for benchmarking.

However, applying a procedure similar to the one that is explained in this technical report and conducting a pilot project is a good practice. Such a pilot project may yield measurement results guiding the practitioner to derive probabilistic parameters in better accuracy.

Cost of an undetected error emerging in the final product (C_{err}) depends on the criticality of the end product. While deciding on this impact value practitioner shall answer questions such as;

- What would happen if one error is injected to the final product?
- Would the final product be useless if it contains a few errors?
- How much damage unsatisfied end product user can cause?

Accepting poor quality submissions of workers attract cheaters. Similarly, denying good quality work discourages honest workers. Both situations have negative effect on worker community and trust mechanisms, which is represented by C_{dmg}. It is difficult to assign a cost value to this negative effect. This value can be used for risk management purposes, and depends on the risk appetite of the practitioner.

Redundancy

[1] $CoQ_{Red} = N \,.\, (((m-1)\,.\,C_0 + C_{agg}) + P_{IC}\,.\,(m\,.\,C_0 + C_{agg}) + P_{FP}\,.\,(C_{err} + C_{dmg}))$

Control Group

[2] $CoQ_{CG} = N\,.\,((C_1) + (P_{FN} + P_{TN})\,.\,(C_0 + C_1) + (P_{FP} + P_{FN})\,.\,C_{dmg} + P_{FP}\,.\,C_{err})$

Gold Standard

[3] $CoQ_{GS} = X\,.\,C_{exp} + N\,.\left(\frac{k}{t-k}\right).(C_0 + (1-(P_P)^k)\,.\,t\,.\,C_0 + (P_P)^k.\ P_W\,.\,(t-k).\,(C_{err} + C_{dmg}))$

3 METHOD

In this research an experiment is conducted which consists of two phases. In the first phase workers are asked to perform a subjective non-deterministic microtask which is to draw an illustration of a lizard. The expected result of this primary task is a set of images which contains both poor quality and high quality drawings (Figure 8). In the second phase of the experiment, three different crowdsourcing designs are set up for separate groups of workers to evaluate the quality of the same image set produced as a result of the primary tasks. Each design consists of using certain quality assurance mechanisms to make sure the control tasks (secondary tasks) are performed satisfying the quality criteria.

The goals of this experiment are:

- to derive metrics for comparing the quality level provided by various mechanisms,
- to verify the cost models which were formerly defined,
- to observe the effect of utilized evaluation scale (Likert vs. binary) on control decisions,
- to assess the operation of various quality assurance mechanisms in a practical setting,
- to exemplify practical usage of observed cheat probability parameters with the models.

The independent variable is the *quality assurance mechanism* while the dependent variables to be monitored are; *effectiveness* (quality levels achieved), *costs* and *error aptitude* (cheat probability) of respective quality assurance mechanism.

3.1 Phase 1. Making a simple drawing of a lizard

The primary task is to draw a simple illustration of a lizard. An open call is made on Amazon Mechanical Turk (AMT), a crowdsourcing platform, for potential crowd workers to contribute. The contributors are provided with a link referring to a web application which contains an open source canvas drawing tool (litvancas) and the textual description of the task (Figure 2). The canvas tool is a simplified version of the open source LitCanvas tool. The text on the web page simply asks the contributors to draw a lizard figure using the provided tool. It is also stated that

the drawings are to be checked for quality by other AMT workers, and they will be paid only if the other users decide the drawing resembles a lizard. No qualifications are required to participate in lizard drawing.

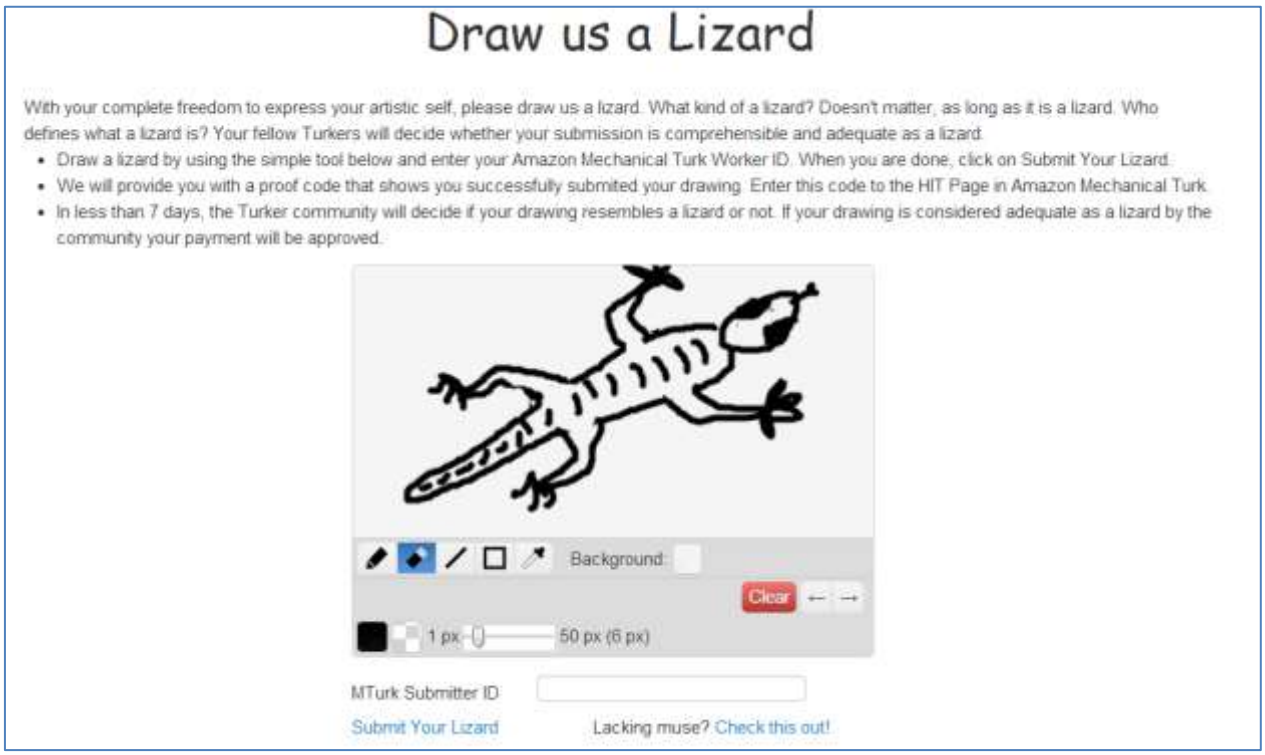


Figure 2: User interface of the primary task

The web application is developed with PHP. All submissions to both primary and secondary tasks are logged on a MySql database for data collection and analysis.

Both the primary and secondary tasks are *subjective* and *non-deterministic*. People can disagree whether a drawing looks like a lizard or not. Drawing a lizard figure is not a straightforward and

easy task. The lizard figure concept is intentionally picked so that the number of poor quality contributions may be high. It is desired to have a data set containing poor quality data so that the effectiveness of the quality assurance mechanisms can be checked.

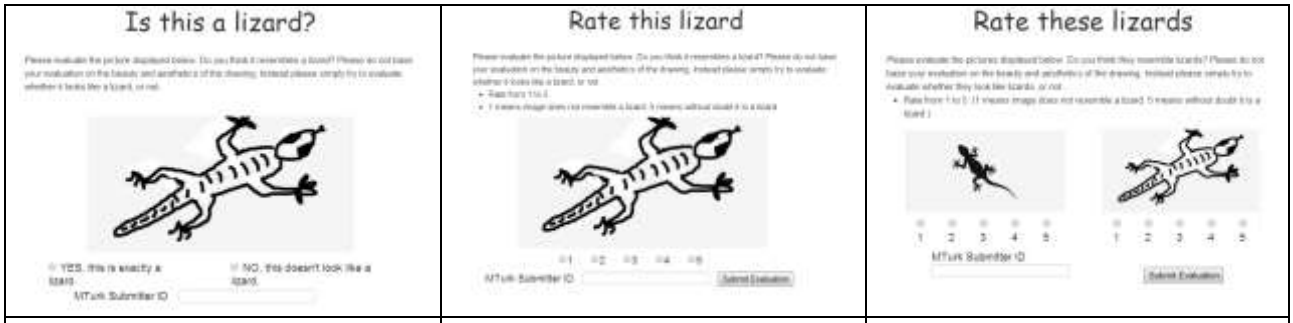
The lizard drawing task was inspired by The Sheep Market (Koblin, 2009), in which AMT users were asked to draw sheep.

Main motivation to contribute in drawing is the payment. The instructions specifically state that the payment is done if the drawing would be accepted by other workers. If the drawing is denied, no payment will be made. Contributors are asked to provide their AMT Worker ID before submitting their drawings and upon completion the system generates a unique proof code for the contributors, for them to enter to AMT task interface. Thus the system records all unique proof codes generated, the AMT Worker ID, the timestamp of the submission and the image submitted. Successful primary submissions receive a payment of 0.15$. The completion criterion of the experiment is to have a set of 500 hand drawn lizard images and have the same image data set quality checked.

All web pages which the users have access to contain links referring to the ethical notice, which states that the task they perform is a part of a university research project. The ethical notice also provides basic information regarding what the experiment is about and specifies which data is collected.

3.2 Phase 2. Evaluating the hand drawn lizard images

The secondary tasks in the experiment are the control tasks in which the contributors are asked to evaluate the hand drawn lizard images. In this phase, 3 separate groups of contributors are provided on AMT with links to 3 different external web applications according to the group they belong (Figure 3). All contributors within the same group are redirected to the same page. Web pages display the image to be evaluated, the input fields for contributors' evaluation and AMT IDs. Participating in any of the groups is voluntary. The participants can submit only one contribution. Instructions indicate that correct submissions are to be paid 0.01$ and incorrect submissions are to be rejected.



Control Group Voting	Control Group Rating	Gold Standard Rating

Figure 3: User interfaces of control (secondary) tasks

Participating in more than one group is not explicitly restricted but since the call for participating the tasks are published in different times, the likelihood of same person participating in more than one group is low. Even if the same worker participates in more than one group, it is highly unlikely for that person to evaluate the same image in different groups because the images are selected randomly from a large data set. Thus it is assumed that the effects of inter-group participation can be ignored.

3.2.1 Control Group Voting (CG Voting)

A group of people are assigned with the task to evaluate the resemblance of an image to a lizard. The participants are shown a random image from the hand drawn lizard image data set and asked to vote if the image resembles a lizard or not. The submission is made in a binary scale; YES or NO.

This task continues until each and every image in the data set is evaluated 3 times to make application of majority decision possible.

CG Voting task is used as a quality assurance mechanism for the primary task. The output of CG Voting directly affects the quality of the primary task.

In this experiment CG Voting is used both with and without quality measures applied, so that the effects can be observed. Each vote for an image is considered a separate quality control for the primary task. The ratio of false quality evaluations (*false positive* and *false negative*) is expected to be higher than the case in which quality measures are used. Thus, applying CG Voting without quality assurance mechanisms is expected to result in lower cost of conformance but higher cost of non-conformance.

Three separate submissions for CG Voting are used to constitute majority decision on each image. Each triple vote casted to evaluate a single image is considered as one quality control by aggregating based on a majority decision (most frequent vote in a total of 3 votes is selected).

3.2.2 Control Group Rating (CG Rating)

A separate group of people is assigned with a quality control task which is similar to the CG Voting but in this case the evaluation is submitted in Likert scale rather than a binary scale. The Likert scale consisted numbers from 1 to 5 representing the sentiment of the contributor against the resemblance of the hand drawn image to a lizard.

CG Rating is also applied with or without quality assurance and the expected outcome is the same which is described earlier in the CG Voting procedure.

3.2.3 Gold Standard Rating (GS Rating)

In order to raise the probability that a control group contributor submits a valid rating, two images are shown at the same time. One of the images is selected among the hand drawn lizard images data set and the other one is selected among a gold standard image database. The gold standard image database includes 40 images. 20 of those images are good examples of lizard drawings and the other 20 are not images of lizards at all.

In GS Rating, the participants are asked to rate both images in a Likert scale (1 to 5). The submitted rating for the gold standard image is checked. If it correctly evaluates the gold standard image the contributor's rating for the hand-drawn lizard image is accepted. If the contributor fails to provide a valid rating for the gold standard image then the rating submitted for the hand drawn image is assumed to be invalid.

In GS Rating, each image in the lizard image set is displayed for evaluation for 3 times so that the results of a majority decision can also be observed.

GS Rating is actually the application of gold standard quality assurance on secondary tasks. Using GS Rating with redundancy is expected to increase the cost of conformance and decrease the cost of non-conformance.

3.2.4 Expert Judgment

The quality of the submissions is decided by comparing the submissions against an expert review. The expert review is done by researchers, completely ignoring the aesthetics of the image submitted. This constitutes the baseline for checking the image quality for images submitted by contributors.

3.2.5 Utilization of quality assurance mechanisms in this experiment

In order to prevent a potential ambiguity a mapping between the quality assurance mechanisms and experiment tasks is provided in Table 1.

Table 1: Mapping between quality assurance mechanisms and experiment tasks

Quality Assurance Mechanism	**Task Design**	**Quality Assurance Applies to:**
Control Group	CG Voting	Primary task
Control Group	CG Rating	Primary task
Redundancy	CG Voting	Secondary task
Redundancy	CG Rating	Secondary task
Redundancy	GS Rating	Secondary task
Gold Standard	GS Rating	Secondary task

Control Group quality assurance mechanism is used only on the primary task, in CG Voting design. All voting and rating tasks (secondary tasks) are performed redundantly enabling

Redundancy quality assurance mechanism to be used on secondary tasks. Gold Standard quality assurance mechanism is applied to secondary tasks in GS Rating design.

3.3 Findings

504 images are drawn and submitted by 283 distinct workers. 27 obvious cheat attempts are detected by expert review in primary task. These attempts are made by 17 distinct workers. Obvious cheat attempts are blank images, scribbles or drawings of other objects such as a house or a car (Figure 4). Even though the nature of the task is subjective, the obvious cheats are easily detectable without causing any judgment conflicts which may cause from subjective opinions of the workers. Thus, it is meaningful to compare *error detection effectiveness* of different quality assurance mechanisms over this obvious faulty contribution set.

Figure 4: Sample results of obvious cheat attempts

A total of 5.183 secondary tasks were performed by 1.230 different workers. 504 of these control tasks were completed by single expert, constituting a baseline for quality control. 1.512 control group voting, 1.512 control group rating and 1.655 gold standard tasks were submitted. Each gold standard task consisted of one gold standard microtask and one ordinary rating microtask. 143 of gold standard submissions were failed.

3.3.1 Error detection effectiveness (EDE)

EDE is the quality assurance mechanisms' effectiveness of detecting errors. EDE is calculated by formula [4]. Using this formula requires knowing the number of errors in the data set. This can be done by deliberately injecting errors, or identifying the errors manually by expert review. The second option is not feasible especially when the data set is large. Thus, selecting a small sample for expert review may solve this issue. In this experiment all 504 images were reviewed by an expert, identifying 27 obvious defects. As stated before, obvious defects are not considered subjective. Some examples of obvious defects are given in Figure 4.

[4] $$EDE = \frac{\text{\# errors detected}}{\text{\# total errors}}$$

Table 2: EDE results

		$\frac{\text{\# errors detected}}{\text{\# total errors}}$	EDE
CG Voting	single	71 / 81	0.88
	with Redundancy	25 / 27	0.93
CG Rating	single	72 / 81	0.89
	with Redundancy	26 / 27	0.96
GS Rating	single	78 / 81	0.96
	with Redundancy	26 / 27	0.96
Expert Review		27 / 27	1.00

Table 2 shows the number of errors detected by quality assurance mechanisms. 3 valid votes (or ratings) were redundantly collected for each of 504 images, in which 27 defects were identified. This results with 27 majority decisions and 81 single decisions for these defects. Majority decisions were made by automatically selecting the most frequent submission in three control votes (or ratings).

Single CG Voting and single CG Rating are basic secondary tasks with no additional quality assurance mechanisms applied upon them. Significant increases in EDE are observed as Redundancy is applied on both CG Voting and CG Rating. When Redundancy is applied, CG Voting EDE increased 0.05 and CG Rating EDE increased 0.07.

When Gold Standard quality assurance mechanism is applied upon CG Rating, EDE increased from 0.89 to 0.96. However applying both redundancy and gold standard together on CG Rating did not increase EDE of CG Rating any further than 0.96.

3.3.2 Decision fitness (subjective accuracy)

A perfect quality assurance mechanism is expected to detect or prevent all poor quality submissions and support the production process to yield exactly the desired products. However quality assurance mechanisms are generally not perfect. Accuracy of a quality assurance mechanism can be measured by comparing the actual outcomes of quality assurance processes and the desired results.

In this experiment both the primary and secondary tasks are in subjective nature. It is not meaningful to measure the accuracy of quality assurance mechanisms over subjective tasks because the outcome of production process depends on many parameters and is not deterministic. Observations in non-deterministic processes are not repeatable thus cannot be generalized.

Therefore, in this experiment the term decision fitness is used to express accuracy of quality assurance mechanisms in a subjective environment. Decision fitness is the ratio of agreement on the quality of microtask outputs, between the quality assurance mechanism and the expert review.

Decision fitness should not be used to evaluate or compare various quality assurance mechanisms. In this experiment decision fitness is measured only to provide a sample case for cost of quality analysis and cost model verification.

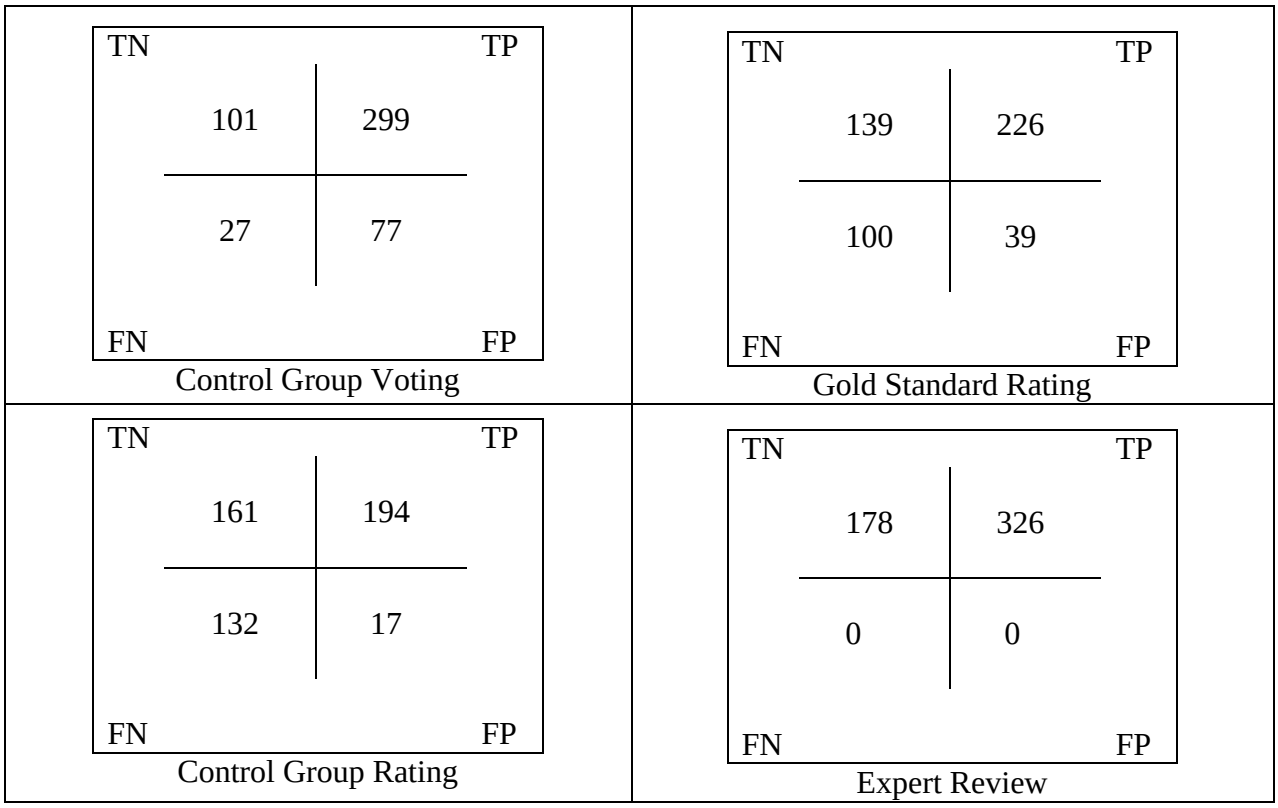


Figure 5: Observed outcomes of various quality assurance mechanisms

Figure 5 displays observed outcomes in experiment tasks. TP (true positive) is the case which the quality assurance mechanism decides a valid submission positively fits the quality criteria. TN (true negative) refers to the situation that the quality assurance mechanism correctly detects a poor quality submission as invalid. FP (false positive) and FN (false negative) are the cases in which the quality assurance mechanism falsely identifies a poor quality submission as valid, or high quality submission as invalid.

When measured on objective tasks quality assurance mechanisms are expected to display decision fitness ratios which are similar to EDE. However when applied on subjective tasks such as in this case, fitness is observed to be significantly lower that EDE, which is expected (Table 3).

The decision fitness of quality assurance mechanisms against the expert review is calculated by the formula [5].

[5] $$FIT = \frac{\text{TN} + \text{TP}}{\text{\# total submissions}}$$

Table 3: Decision fitness results

		$\frac{\text{TN} + \text{TP}}{\text{\# total submissions}}$	FIT
CG Voting	single		0.75
	with Redundancy		0.79
CG Rating	single		0.66
	with Redundancy		0.70
GS Rating	single		0.69
	with Redundancy		0.72
Expert Review			1.00

3.3.3 Cheat probability

Cheating is the act of a contributor to make poor quality submissions whether because of malevolent intentions or simply an attempt of maximizing personal gain. Cheating in crowdsourcing is often in the form of making random submissions. Cheat probabilities are induced by the actual submissions of workers, comparing the submissions with expert review results. P_W is defined as the probability of a worker to make an invalid submission. It depends on many parameters including the task design and crowd characteristics. In this experiment P_W is calculated for both primary and secondary tasks, including lizard drawing, CG Voting, CG Rating, GS Rating.

Poor quality submissions detected and undetected by quality assurance mechanisms cause internal and external failures. Error aptitude of a quality assurance mechanism is defined as the ability of a mechanism to detect or miss errors. Detected errors need to be fixed, thus result in rework. Non detected errors lead to external failures. Making this distinction helps the practitioners to calculate cost of quality, by using different error aptitude values in cost models.

Observations regarding the probabilistic parameters of the cost models are described below.

P_W

Primary task

As a result of expert review 178 images out of 504 contributions are identified as poor quality. Thus PW of the primary task is:

Primary $P_W = 178 / 504 = 0.34$

Secondary task

Poor quality contributions of control task workers actually consist of the cases which control group worker fails to identify invalid and valid submissions. These cases are FN and FP decisions of respective task design. Observed result of secondary task PW values are:

CG Voting $P_W = 383 / 1512 = 0.25$

CG Rating $P_W = 510 / 1512 = 0.34$

GS Rating $P_W = 468 / 1512 = 0.31$

3.3.4 Aptitude values

Redundancy aptitude values:

P_{IC}

Probability of Redundancy process to reach inconsistent (IC) state:

Redundancy quality assurance process reaches to an inconclusive state only if the number of elements in the result set is not less than the number of redundant submissions or number of redundant submissions is even.

In this setting the number of redundant control tasks is odd. Thus reaching an inconclusive state is not possible. However this state can be achieved by disregarding every third submission made for each control tasks for the sake of exemplification. The results are provided in Table 4.

Table 4: Observed P_{IC} values

	# redundant submissions (m)	P_{IC}
CG Voting with redundancy	3	0.00
CG Rating with redundancy	3	0.00
GS Rating with redundancy	3	0.00
CG Voting with redundancy	2	0.28
CG Rating with redundancy	2	0.35
GS Rating with redundancy	2	0.37

P_{FP}

Probability of Redundancy process to reach positive outcome when the quality of submission is actually not acceptable (P_{FP}):

P_{FP} is calculated for CG Voting with redundancy, CG Rating with redundancy and GS Rating with redundancy, by comparing the fitness of majority decision with expert judgment (Table 5).

Table 5: P_{FP} values for different cases

	FP / # total submissions	P_{FP}
CG Voting with redundancy	77 / 504	0.15
CG Rating with redundancy	17 / 504	0.03
GS Rating with redundancy	39 / 504	0.08

Control group error aptitude values:

Control Group quality assurance mechanism is applied on the primary task. Error aptitude values are calculated on CG Voting without redundancy and CG Rating without redundancy.

P_{FN}

P_{FN} is the probability of the worker to submit a valid result but the control group incorrectly decides it is invalid. In this experiment FN is the case in which the primary task worker submits a valid lizard drawing but the control group flags the submission as not a lizard. Observed P_{FN} results are given in Table 6.

Table 6: Observed P_{FN} values for CG Voting and CG Rating tasks

	FN / # total submissions	P_{FN}
CG Voting 1	57 / 504	0.11
CG Voting 2	45 / 504	0.09
CG Voting 3	38 / 504	0.08
CG Rating 1	131 / 504	0.26
CG Rating 2	139 / 504	0.28
CG Rating 3	145 / 504	0.29

P_{TN}

P_{TN} is the probability of the worker to submit an invalid result and the control group correctly identifies it as invalid. In this experiment TN is the case when the primary task worker submits an invalid lizard drawing and the control group detects the invalid submission correctly. Observed P_{TN} results are given in Table 7.

Table 7: Observed P_{TN} values for CG Voting and CG Rating tasks

	TN / # total submissions	P_{TN}
CG Voting 1	107 / 504	0.21
CG Voting 2	94 / 504	0.19
CG Voting 3	90 / 504	0.18
CG Rating 1	150 / 504	0.30
CG Rating 2	146 / 504	0.29
CG Rating 3	143 / 504	0.28

P_{FP}

P_{FP} is the probability of the worker to submit an invalid result but the control group fails to detect the poor quality and identifies it as valid. In this experiment FP is the case when the primary task worker submits an invalid lizard drawing and the control group accepts it as a valid submission. . Observed P_{FP} results are given in Table 8.

Table 8: Observed P_{FP} values for CG Voting and CG Rating tasks

	FP / # total submissions	P_{FP}
CG Voting 1	71 / 504	0.14
CG Voting 2	84 / 504	0.17
CG Voting 3	88 / 504	0.18
CG Rating 1	28 / 504	0.06
CG Rating 2	32 / 504	0.06
CG Rating 3	35 / 504	0.07

Gold standard error aptitude values:

Gold Standard quality assurance mechanism is applied on the secondary task. Error aptitude values are calculated on combined submissions made for GS Rating tasks without redundancy. P_N is the probability of a worker to submit a negative result to a gold standard task. P_P is the probability of a worker to make a valid submission for a gold standard task. The observed P_N and P_P aptitude values are shown in Table 9.

Table 9: Observed P_N and P_P values for GS Rating tasks

	N / # total submissions	P / # total submissions	P_N	P_P
GS Rating	143 / 1655		0.09	0.91

3.4 Practical application of cost of quality models

This section exemplifies utilization of cost models described in section 2.3. Mapping between quality assurance mechanisms and related experimental design is provided in Table 1. Model parameters and their definitions are shown in Table 10.

Table 10: displays descriptions of variables used in the cost models.

Variable	Description
N	The total number of microtasks. For the primary task, N = 504. For the secondary task N = 1512.
C_0	The cost of 1 microtask. Primary task in this experiment is assigned with the monetary value of $0.15.
C_1	The cost of 1 control task. Secondary task in this experiment is assigned with the monetary value of $0.01.
C_{err}	Cost of 1 undetected error emerging in the end result.
C_{dmg}	Costs which occur due to the damage done to trust mechanisms and the worker community when an evaluation fails.
P_{FN}	Probability of the quality assurance mechanism to reach a false negative state. This parameter is used in formula [2]. Measured values for this parameter are given in Table 6.
P_{FP}	Probability of the quality assurance mechanism to reach a false positive state. This parameter is used in formulas [1] and [2]. Measured values for this parameter are given in Table 5.
P_{IC}	Probability of the quality assurance mechanism to reach an inconclusive state. This parameter is used in formula [1]. Measured values for this parameter are given in Table 4.
P_{TN}	Probability of the quality assurance mechanism to reach a true negative state. This parameter is used in formula [2]. Measured values for this parameter are

	given in Table 7.
P_W	Probability of a worker to make a poor quality submission. This parameter is used in formula [3].
P_N	Probability of the gold standard quality assurance mechanism to reach a negative state. This parameter is used in formula [3]. Measured values for this parameter are given in Table 9.
P_P	Probability of the gold standard quality assurance mechanism to reach a positive state. This parameter is used in formula [3]. Measured values for this parameter are given in Table 9.
k	Number of gold standard tasks in a set of task batch. This parameter is used in formula [3].
t	Total number of tasks in a set of task batch. This parameter is used in formula [3].
m	Number of repetitions in redundancy quality assurance mechanism. This parameter is used in formula [1]. m = 3.
X	Number of elements in the gold standard task set. In this case X = 40.
C_{exp}	Cost of introducing 1 gold standard task in the gold standard task set.

3.4.1 CG Voting

In CG Voting setting, the primary task is lizard drawing and the control group quality assurance technique is used on the primary task. Cost of quality for control group quality assurance mechanism is calculated by substituting measured values into the formula [2].

[2] $CoQ_{CG} = N \,.\, ((C_1) + (P_{FN} + P_{TN}) \,.\, (C_0 + C_1) + (P_{FP} + P_{FN}) \,.\, C_{dmg} + P_{FP} \,.\, C_{err})$

$CoQ_{CG} = 504 \,.\, ((0.01) + (0.09 + 0.19) \,.\, (0.15 + 0.01) + (0.16 + 0.09) \,.C_{dmg} + 0.16 \,.\, C_{err})$

$CoQ_{CG} = 504 \,.\, 0.01 + 504 \,.\, 0.28 \,.\, 0.16 + 504 \,.\, 0.25 \,.\, C_{dmg} + 504 \,.\, 0.16 \,.\, C_{err}$

$CoQ_{CG} = 27.62 + 126 \,.\, C_{dmg} + 80{,}64 \,.\, C_{err}$

3.4.2 CG Voting with Redundancy

In CG Voting with Redundancy setting, redundancy technique is used on the secondary task. Cost of quality for redundancy quality assurance mechanism is calculated by substituting measured values into the formula [1].

[1] $CoQ_{Red} = N \,.\, (((m - 1) \,.\, C_0 + C_{agg}) + P_{IC} \,.\, (m \,.\, C_0 + C_{agg}) + P_{FP} \,.\, (C_{err} + C_{dmg}))$

$CoQ_{Red} = 504 \,.\, (((3\text{-}1) \,.\, 0.01 + 0) + 0 \,.\, (3 \,.\, 0.01 + 0) + 0.15 \,(C_{err} + C_{dmg}))$

$CoQ_{Red} = 504 \,.\, 0.02 + 504 \,.\, 0.15 \,C_{dmg} + 504 \,.\, 0.15 \,C_{err}$

$CoQ_{Red} = 10.08 + 75.6 \,C_{dmg} + 75.6 \,C_{err}$

3.4.3 CG Rating

In CG Rating setting, the primary task is lizard drawing and control group quality assurance technique is used on the primary task. Cost of quality for control group mechanism is calculated by substituting measured values into the formula [2].

$CoQ_{CG} = N \,.\, ((C_1) + (P_{FN} + P_{TN}) \,.\, (C_0 + C_1) + (P_{FP} + P_{FN}) \,.\, C_{dmg} + P_{FP} \,.\, C_{err})$

$CoQ_{CG} = 504 \,.\, ((0.01) + (0.28 + 0.29) \,.\, (0.15 + 0.01) + (0.06 + 0.28) \,.\, C_{dmg} + 0.06 \,.\, C_{err})$

$CoQ_{CG} = 504 \,.\, 0.01 + 504 \,.\, 0.0912 + 504 \,.\, 0.34 \,.\, C_{dmg} + 504 \,.\, 0.06 \,C_{err}$

$CoQ_{CG} = 51.01 + 171.36 \,.\, C_{dmg} + 30.24 \,C_{err}$

3.4.4 CG Rating with Redundancy

In CG Rating with Redundancy setting, redundancy technique is used on the secondary task. Cost of quality for redundancy quality assurance mechanism is calculated by substituting measured values into the formula [1].

[1] $\mathrm{CoQ_{Red}} = \mathrm{N} \,.\, (((\mathrm{m}-1) \,.\, \mathrm{C_0} + C_{agg}) + P_{IC} \,.\, (m \,.\, C_0 + C_{agg}) + P_{FP} \,.\, (C_{err} + C_{dmg}))$

$\mathrm{CoQ_{Red}} = 504 \,.\, (((3\text{-}1) \,.\, 0.01 + 0) + 0 \,.\, (3 \,.\, 0.01 + 0) + 0.03\ (\mathrm{C_{err}} + \mathrm{C_{dmg}}))$

$\mathrm{CoQ_{Red}} = 504 \,.\, 0.02 + 504 \,.\, 0.03\ \mathrm{C_{dmg}} + 504 \,.\, 0.03\ \mathrm{C_{err}}$

$\mathrm{CoQ_{Red}} = 10.08 + 15.12\ \mathrm{C_{dmg}} + 15.12\ \mathrm{C_{err}}$

3.4.5 GS Rating

In GS Rating setting, gold standard mechanism is used on the secondary task. Cost of quality for gold standard quality assurance mechanism is calculated by substituting measured values into the formula [3].

[3] $\mathrm{CoQ_{GS}} = X \,.\, \mathrm{C_{exp}} + N \,.\, \left(\frac{k}{t-k}\right) \,.\, (C_0 + \left(1 - (P_P)^k\right) \,.\, \mathrm{t} \,.\, C_0 + (P_P)^k \,.\, P_W \,.\, (\mathrm{t}-\mathrm{k}) \,.\, \left(C_{err} + C_{dmg}\right))$

$\mathrm{CoQ_{GS}} = 40 \,.\, \mathrm{C_{exp}} + 504 \,.\, (1 / 1) \,.\, (0.01 + 0.09 \,.\, 2 \,.\, 0.01 + 0.91 \,.\, 0.31 \,.\, (2 - 1) \,.\, (\mathrm{C_{err}} + \mathrm{C_{dmg}}))$

$\mathrm{CoQ_{GS}} = 40 \,.\, \mathrm{C_{exp}} + 5.04 + 0.9072 + 142.1784 \,.\, (\mathrm{C_{err}} + \mathrm{C_{dmg}}))$

$\mathrm{CoQ_{GS}} = 40 \,.\, \mathrm{C_{exp}} + 5.9472 + 142.1784 \,.\, (\mathrm{C_{err}} + \mathrm{C_{dmg}}))$

3.4.6 GS Rating with Redundancy

In GS Rating with redundancy setting, both gold standard and redundancy techniques are used on the secondary task. This is a special case in which two different cost models [1], [2] need to be combined in order to derive the correct cost model.

[1] $\mathrm{CoQ_{Red}} = \mathrm{N} \,.\, (((\mathrm{m}-1) \,.\, \mathrm{C_0} + C_{agg}) + P_{IC} \,.\, (m \,.\, C_0 + C_{agg}) + P_{FP} \,.\, (C_{err} + C_{dmg}))$

[3] $\mathrm{CoQ_{GS}} = X \,.\, \mathrm{C_{exp}} + N \,.\, \left(\frac{k}{t-k}\right) .\, (\, C_0 + \left(1 - (P_P)^k\right) .\, \mathrm{t} \,.\, C_0 + (\, P_P)^k .\; P_W \,.\, (\mathrm{t} - \mathrm{k}) .\, \left(C_{err} + C_{dmg}\right))$

By design, first gold standard is applied. Both gold standard task and the ordinary task are performed redundantly.

[4] $\mathrm{CoC_{mix}} = X \,.\, \mathrm{C_{exp}} + N \,.\, C_0 \,.\, (\left(\frac{k}{t-k}\right) + (m-1) \,.\, \left(\frac{t}{t-k}\right))$

Cost of conformance for this mix model is shown in formula [4]. Introducing X gold standard tasks into the system has the cost of $X \,.\, C_{exp}$. Gold standard tasks are only performed for quality assurance purposes so cost of completing any gold standard task is considered a cost of conformance. $N \,.\, C_0 \,.\, \left(\frac{k}{t-k}\right)$ is the cost of completing all gold standards. Both gold standard and ordinary tasks are done m times. So $N \,.\, (m-1) \,.\, \left(\frac{t}{t-k}\right)$ tasks are performed redundantly.

[5] $\mathrm{IF_{mix}} = N \,.\, m \,.\, \left(\frac{k}{t-k}\right) .\, \left(1 - (P_P)^k\right) .\, \mathrm{t} \,.\, C_0 + P_{IC} \,.\, \left(m \,.\, t \,.\, C_0 + C_{agg}\right)$

Internal failure costs are shown in formula [5]. In this hybrid quality assurance mechanism internal failure can occur if a gold standard task is answered incorrectly or the redundancy process results at an inconclusive state. $N \,.\, m \,.\, \left(\frac{k}{t-k}\right) .\, \left(1 - (P_P)^k\right)$ describes the probability of incorrect performance of gold standard tasks out of all tasks completed in the process. $\mathrm{t} \,.\, C_0$ is the impact of an internal failure caused by incorrect gold standard task performance. $P_{IC} \,.\, \left(m \,.\, t \,.\, C_0 + C_{agg}\right)$ denotes the probability and the impact of the hybrid quality assurance mechanism to result with an inconclusive state.

[6] $\mathrm{EF_{mix}} = N \,.\, \left(\frac{k}{t-k}\right) .\; (\, P_P)^k .\; P_W \,.\, (\mathrm{t} - \mathrm{k}) \,.\, P_{FP} \,.\, \left(C_{err} + C_{dmg}\right)$

External failure costs are displayed in formula [6]. $N \,.\, \left(\frac{k}{t-k}\right) .\; (\, P_P)^k$ represents the number of task batches in which gold standard tasks are answered correctly. Along with $P_W \,.\, (\mathrm{t} - \mathrm{k})$ the number of incorrectly performed ordinary tasks in task batches containing no incorrectly performed gold standard tasks. The probability of false positive selection of redundancy process should also be included in the formula. $\mathrm{P_{FP}}$ for GS Rating with Redundancy is measured and can be used directly in the formula. The impact of an external failure is represented with $\left(C_{err} + C_{dmg}\right)$.

[7] $\mathrm{CoQ_{mix}} = X \,.\, \mathrm{C_{exp}} + N \,.\, C_0 \,.\, \left(\left(\frac{k}{t-k}\right) + (m-1) \,.\, \left(\frac{t}{t-k}\right)\right) + N \,.\, m \,.\, \left(\frac{k}{t-k}\right) .\, \left(1 - (P_P)^k\right) .\, \mathrm{t} \,.\, C_0 + P_{IC} \,.\, \left(m \,.\, t \,.\, C_0 + C_{agg}\right) + N \,.\, \left(\frac{k}{t-k}\right) .\; (\, P_P)^k .\; P_W \,.\, (\mathrm{t} - \mathrm{k}) \,.\, P_{FP} \,.\, \left(C_{err} + C_{dmg}\right)$

$\mathrm{CoQ_{mix}} = 40 \,.\, \mathrm{C_{exp}} + 504 \,.\, 0.01 \,.\, ((1 / 1) + (3 - 1) \,.\; (2 / 1) + 504.\, 3.\, (1 / 1) \,.\, (1 - 0.91) \,.\, 2 \,.\, 0.01 + 0 \,.\, (3 \,.\, 2 \,.\, 0.01 + \mathrm{C_{agg}}) + 504 \,.\, (1 / 1) \,.\, (0.91)^1 \,.\, 0.31 \,.\, 1 \,.\, 0.08 \,.\, (\mathrm{C_{err}} + \mathrm{C_{dmg}})$

$CoQ_{mix} = 40 \cdot C_{exp} + 25.2 + 2.42 + 0 + 11.37 \cdot C_{err} + 11.37 \cdot C_{dmg}$

$CoQ_{mix} = 40 \cdot C_{exp} + 27.62 + 11.3743 \cdot C_{err} + 11.3743 \cdot C_{dmg}$

3.4.7 Various external failure scenarios

To analyze the effect of external failures on the cost of different quality assurance mechanisms, various values for variables are inserted to the formulas.

The total direct cost of the producing the end product is basically the sum of task performance costs excluding all quality related costs.

Primary task: $C_{prod} = 504 \cdot 0.15 = 75.6$

Secondary task: $C_{prod} = 504 \cdot 0.01 = 5.04$

The cost of an expert introducing 1 gold standard task into the system is assumed to be 10 times higher than the cost of 1 microtask. This assumption is case dependent, thus should be adjusted by crowdsourcing practitioners for better accuracy estimations.

$C_{exp} = 10 \cdot C_0$

It is difficult to assign a cost value for C_{dmg}, the damage done to the trust mechanism and crowd community when a wrong evaluation regarding a worker's submission is made. This value should be determined according to the project environment, crowd characteristics and risk appetite of the practitioner. In this experiment we used two separate values for C_{dmg} in order to cover different cases. In the first case we assume that damage done to trust mechanism is insignificant therefore C_{dmg} is equal to 0. In the second case we assume that one failure in evaluating one submission has an impact of the cost of one microtask, therefore $C_{dmg} = C_0$.

$C_{dmg} = 0$ OR $C_{dmg} = C_0$.

We provide a value interval for C_{err}. Minimum value for C_{err} is assumed to be equal to C_0. The maximum value is assumed to be one tenth of the total direct cost of the product. When this value interval is fed into the cost models we can observe the cost outcomes for different designs changing according to the C_{err}. C_{err} should be determined by the crowdsourcing practitioner by carefully evaluating the impacts of external failures.

$$Min(C_{err}) = C_0, \quad Max(C_{err}) = 1/10 \; . \; C_{prod}$$

Figure 6 shows the cost of quality of CG Voting and CG Rating designs for varying C_{err} values. According to the observed values CG Rating has a lower C_{err} coefficient than CG Voting which makes it a more robust design against the impacts of undetected errors. Only when the impact of C_{err} is lower than 0.4 – 0.6, cost of quality of CG Voting is lower than cost of quality of CG Rating.

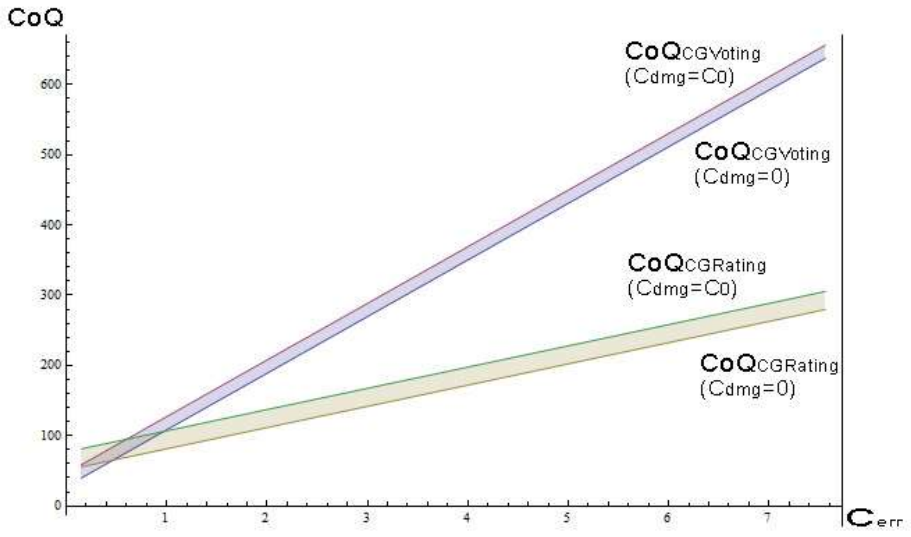


Figure 6: Effect of various external failure values on cost of quality for CG Voting and CG Rating designs

Figure 7 shows cost of quality of GS Rating, GS Rating with Redundancy, CG Voting with Redundancy and CG Rating with Redundancy designs.

Both GS Rating and GS Rating with Redundancy have higher initial costs than the other designs due to the cost of introducing gold standard tasks into the system by experts. Since the cost of building a gold standard task pool is directly proportionate with the pool size, the number of tasks should be selected carefully.

Cost of quality of using gold standard along with redundancy is higher than other quality assurance mechanisms as expected. However it is observed that the cost of quality of GS Rating with Redundancy becomes lower than cost of quality of GS Rating when C_{err} is higher than 0.45.

The lowest cost of quality values are observed in CG Rating with Redundancy. Furthermore, CG Rating is observed to be more robust for large C_{err} values than other designs.

It is also observed that rating designs (CG Rating with Redundancy and GS Rating with Redundancy) show a more robust pattern than the voting design.

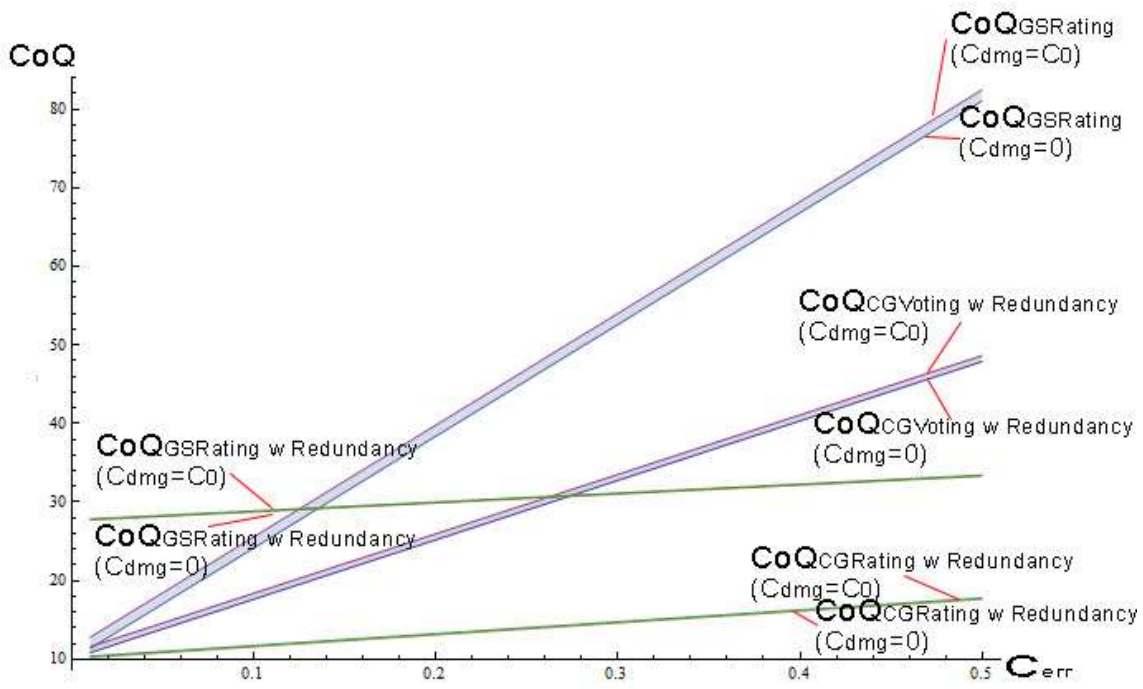


Figure 7: Effect of various external failure values on cost of quality for CG Voting with redundancy, CG Rating with redundancy, GS Rating and GS Rating with redundancy designs

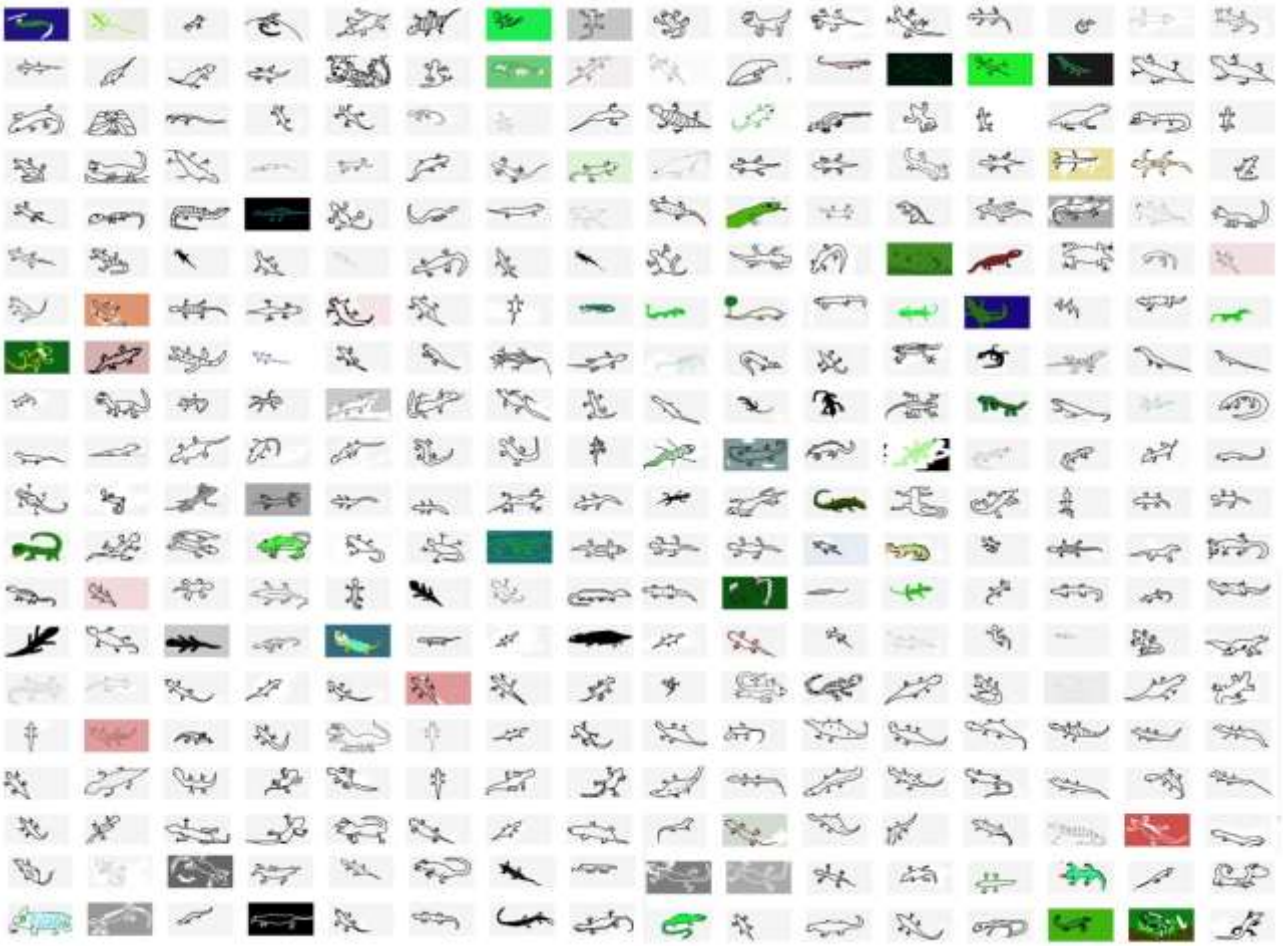


Figure 8: 320 lizards accepted by the quality assurance mechanisms

4 DISCUSSION

Rating vs. Voting

CG Voting and CG Rating designs only differ in the input type. CG Voting uses a binary voting scheme whereas CG Rating uses Likert scale, asking workers to grade images from 1 to 5. We observed significant differences in both decision fitness and error aptitude values of CG Voting and CG Rating.

In order to compare the results with each other, we mapped 1, 2 and 3 to NO; and 4, 5 to YES. The reason we acknowledge 3 as negative is because of the semantics of the task description. The question asks: "*is this a lizard?*" As the practitioners of this crowdsourcing project we want to identify positive and non-positive results. Non-positive results are either negative or cases in which making a decision is not possible. Even though it was clearly stated in the instructions some workers chose to rate the lizards according to the images' aesthetics and submitted a grade of 3 for images which resemble lizards.

We experienced that Likert and binary answer formats are not comparable. This conclusion fits the results reported by Dolnicar (Dolnicar, 2006). The reason for this incomparability is that there is no way of knowing respondents' response style. When such comparisons are made, quality of interpretation of data may be hindered.

Dolnicar reports that binary answer format is significantly faster than other formats, but not perceived easier by respondents (Dolnicar, 2006). In this experiment average task completion time for CG Voting is 42.44 seconds and 42.39 seconds for CG Rating, which indicates there is almost no difference in task completion time of binary and Likert answer formats.

Figure 6 shows how the cost of quality of CG Voting and CG Rating designs change for varying C_{err} values. According to this observation CG Rating makes a more robust design against the impacts of undetected errors. The quality assurance mechanism used in both designs and the cost of conformance of both designs is the same. The difference in cost of quality originates only from the cost of non-conformance parameters. When IF and EF values of both designs are compared it can be seen that $IF_{CG\ Rating}$ is greater than $IF_{CG\ Voting}$. IF only emerges when the

quality assurance mechanism detects a failure (TN and FN outcomes), and the observations in this experiment indicates that CG Rating design is more likely to detect a submission as negative compared to CG Voting ($P_{(TN+FN)\ CG\ Rating} = 0.57$ and $P_{(TN+FN)CG\ Voting} = 0.28$). This makes CG Rating a more strict (or pessimistic) method of controlling than CG Voting, resulting in less undetected errors to emerge in the end product. According to these findings it is concluded that a rating scheme is a better method than voting when tolerance for external failure is low but internal failure is more acceptable.

Subjectivity

The most critical risk of this research is that the chosen tasks are subjective. It is probable that if this experiment is conducted once again different results can be observed. However, subjectivity is a fact of crowdsourcing tasks. Crowdsourcing usage for subjective tasks is increasing each day and ways to derive statistically significant results are reported in the literature (Ribeiro, 2010). In certain cases practitioners only pay to know what the crowd thinks. In other words, to access the wisdom of crowds… Thus, the original research problems are still valid. The probability values we report may not be generalizable but the approach we present for measuring these parameters is an important contribution and can be used by practitioners to conduct pilot projects. Results of these pilot projects may be used to make cost models more accurate in practitioners' specific cases.

Better gold standard design

Poor quality submission or cheating is mostly caused by the desire of the workers to maximize their income while minimizing the effort spent on the task. Cheating is often done in the form of making random submissions. Thus, quality assurance mechanisms should be able to deal with this type of worker behavior.

Using a gold standard task with 2 possible options enables the practitioner to detect a random submission with the probability of 50%. When applied without the support of other quality assurance mechanisms this effectiveness is very low. Thus, it is imperative to design the gold

standard tasks so that the probability of a random submission to reach a positive result is minimal.

Optimizing cost of conformance and cost of non-conformance

One of the important goals of performing a cost of quality analysis is to optimize the conformance costs and non-conformance costs. In certain cases which external failures have significant impact, investing in introducing additional quality assurance mechanisms will result in decreased cost of non-conformance. However, in crowdsourcing introducing additional quality assurance mechanisms heavily impacts the cost of conformance.

As shown in Table 2, EDE increases by introducing additional quality assurance mechanisms. When a control group quality assurance mechanism (CG Rating) is used in solitary it provides 89% EDE. Introducing an additional redundancy or gold standard mechanism increases the EDE to 96%. However introducing both gold standard and redundancy together cannot increase EDE any further while significantly increasing the cost of conformance.

Comparison of designs

Figure 7 shows the change of cost of quality values of crowdsourcing designs for varying C_{err} values. The slopes of the graph lines indicate the cost effectiveness of respective quality assurance mechanisms. The lower the slope, the more effective the quality assurance mechanism is. Cost effectiveness is related with non-conformance costs. Thus the most cost effective quality assurance mechanism is not necessarily the one with the lowest cost of quality.

Both CG Rating with Redundancy and GS Rating with Redundancy display a robust profile against increasing C_{err}. Even though both designs are similar in robustness, CG Rating with Redundancy has a lower cost of quality, due to high initial quality costs of GS Rating with Redundancy design.

Using redundancy in GS Rating (GS Rating with Redundancy) leads to a higher cost of quality when C_{err} is small ($C_{err} < 0.13$). However when C_{err} increases redundancy provides cost savings by eliminating errors more effectively and causing less error to remain undetected.

5 REFERENCES

Crosby, B. P. (1979). Quality is free: The Art of Making Quality Certain McGraw-Hill. New York.

Dolnicar, S., & Grün, B. (2007). How constrained a response: a comparison of binary, ordinal and metric answer formats. Journal of Retailing and Consumer Services, 14(2), 108-122.

Iren, D. & Bilgen, S. (2013). Cost models of crowdsourcing quality assurance mechanisms. Ankara. METU. Retrieved from: http://www.expertjudgment.com/academics/TR31122013.pdf

Koblin, A. M. (2009, October). The sheep market. In Proceedings of the seventh ACM conference on Creativity and cognition (pp. 451-452). ACM.

Ribeiro, F., Florencio, D., & Nascimento, V. (2011, September). Crowdsourcing subjective image quality evaluation. In Image Processing (ICIP), 2011 18th IEEE International Conference on (pp. 3097-3100). IEEE.