Human Computation (2014) 1:2:283-314

ISSN: 2330-8001, DOI: 10.15346/hc.v1i2.14

Cost of Quality in Crowdsourcing

DENİZ İREN, Middle East Technical University, Ankara, Turkey

SEMİH BİLGEN, Electrical and Electronics Engineering, Yeditepe University, İstanbul, Turkey

ABSTRACT

Crowdsourcing is a model which allows practitioners to access a relatively inexpensive and scalable workforce. However, due to loose worker-employer relationships, skill diversity of the crowd and anonymity of participants, it tends to result in lower quality compared to traditional ways of doing work. Thus crowdsourcing practitioners have to rely heavily on certain quality assurance techniques to make sure that the end product complies with the quality requirements. Quality assurance techniques increase project cost and time. A well-defined methodology is needed to estimate these impacts in order to manage the crowdsourcing process effectively and efficiently. This paper introduces cost models of common quality assurance techniques that may be applied in crowdsourcing and describes a proposed cost of quality approach for analyzing quality related costs.

1. INTRODUCTION

As a frequently used genre of human computation (Law & Ahn, 2011, p. 3), crowdsourcing can be defined as a value creation process in which the interactive features of the Internet are utilized by a generally large group of people with a certain degree of anonymity who voluntarily choose tasks to work on. Citizen science is a type of research in which members of the public collaborate with researchers (Wiggins & Crowston, 2011). Similar to crowdsourcing, citizen science can often engage large groups of volunteers who have different backgrounds and possess a wide variety of skills. Therefore citizen science faces similar challenges regarding the quality assurance of contributions.

The effectiveness and benefits of crowdsourcing as a business model are no longer under debate due to the continuously growing number of crowdsourcing success stories (von Ahn & Dabbish, 2008; "Wikipedia," n.d.). However, managerial concerns such as economics (Grier, 2011) and minimizing costs (Vukovic & Bartolini, 2010) while improving quality to a defined level of error tolerance (Kittur et al., 2013) still need to be satisfactorily addressed.

In contrast to traditional business models, crowdsourcing lacks a clearly defined pact or a binding service level agreement between workers and the employer. In crowdsourcing settings, crowd workers do not have as high a level of accountability as permanent employees. Challenges regarding the control of crowd-based production (Kittur et al., 2013) raise concerns about the quality of the end product (Kern, Zirpins, & Agarwal, 2009). Therefore crowdsourcing practitioners utilize certain quality assurance techniques which increase costs significantly. Crowdsourcing practitioners require certain methods to estimate the impact of different quality assurance mechanisms on overall cost and quality levels they yield to make accurate plans, select more suitable quality assurance mechanisms and optimize resource allocation.

As a part of total quality management, a Cost of Quality (CoQ) approach has been used in various domains frequently and successfully since the 1970's (Schiffauerova & Thomson, 2006). CoQ is defined as the total cost of all quality related activities which can be expressed as the sum of *conformance* and *non-conformance* costs. Conformance costs are costs spent on activities to avoid poor quality whereas non-conformance costs are costs that occur due to poor quality (Crosby, 1979). Generally failure costs decrease as more investment is made on quality assurance activities. Therefore there is a tradeoff between conformance and non-conformance costs, which must be managed to optimize quality costs.

In this study we apply CoQ analysis and use observed process outcomes to derive cost models of common crowdsourcing quality assurance techniques. These cost models are verified through multiple action research consisting of 3 cases, which differ in terms of the nature of the task. This research has impact at two different levels. At an individual level, cost models can be used by practitioners for estimating achievable quality and cost to make crowdsourcing more manageable. At a global level, extensive utilization of cost models can lead to efficient resource (crowd) utilization.

This paper is organized as follows: After this introductory section, Section 2 sets the background and briefly reviews the existing literature on alternative techniques of achieving crowdsourcing quality. In Section 3, CoQ models of common crowdsourcing quality assurance techniques are introduced. Section 4 describes the multiple action research carried out to assess the validity and applicability of the proposed CoQ models, and discusses the findings. Section 5 concludes the paper.

2. BACKGROUND

2.1 Crowdsourcing and Related Concepts

Often used synonymously with other related concepts such as human computation and social systems (Law & Ahn, 2011, p. 3) crowdsourcing can be defined as an umbrella term which refers to a wide variety of value creation processes and business models with the shared characteristic of using a large group of people as a resource. Taxonomies have been constructed to clarify alternative definitions, draw a borderline separating crowdsourcing from related concepts, and categorize essential characteristics of crowdsourcing (Estellés-Arolas & González-Ladrón-de-Guevara, 2012; Geiger & Seedorf, 2011; A. Quinn & Bederson, 2011; Rouse, 2010; Schenk & Guittard, 2011).

Crowdsourcing quality assurance techniques vary in effectiveness and costs for different situations. In the present research we apply a simple categorization (Figure 1), with no claims for comprehensiveness, which covers the dimensions of *nature of task*, *work output type*, *crowd type* and *quality assurance technique,* with the aim of observing the relationship among these characteristics.

The *nature of task* emphasizes objectivity of the task. For example, counting the number of road junctions on a satellite image of a town is an objective task, thus the results can be checked automatically. However subjective tasks are not. In order to make subjective outputs comparable, the task is usually defined in a way which limits the potential result set of the *work output*. For example, evaluating whether a hand drawn picture resembles the figure of a cat or not and submitting a vote for or against it, is a subjective task which has a finite set of potential results. The potential outcome of this task is binary, either positive or negative, thus, the frequency of the votes cast for the same task instance can be calculated and the result can be automatically aggregated by selecting the majority vote. On the other hand, reading a long text block and summarizing it with a couple of sentences is another example of a subjective task, yet with an infinite set of potential results. In this case the results can only be aggregated manually.

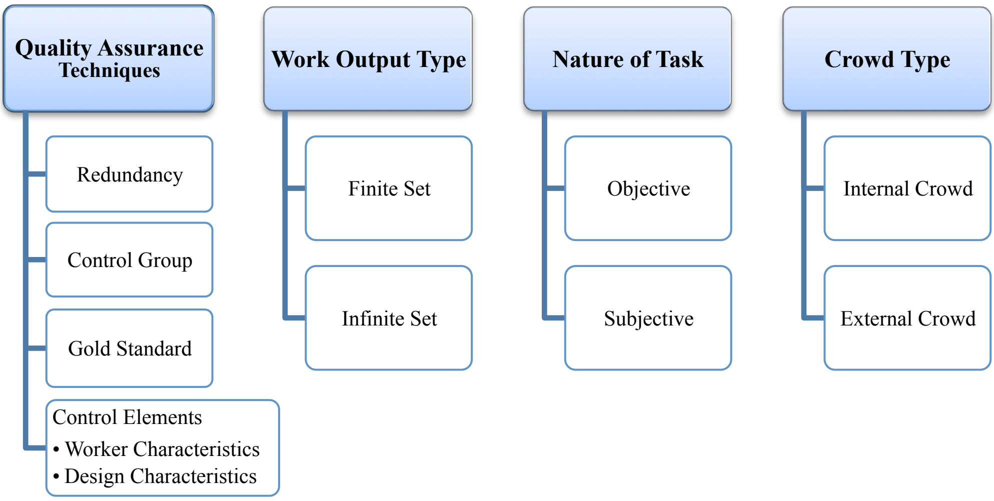


***Figure 1.* Basic crowdsourcing taxonomy dimensions**

Crowd type emphasizes the difference between an *internal crowd* and an *external crowd*. An internal crowd consists of individuals who belong to the same organization such as a company or an association. Therefore these individuals are not anonymous. When an internal crowd is used as a resource, the type of crowdsourcing is generally categorized as enterprise crowdsourcing (Vukovic, 2009). On the other hand, an external crowd refers to online individuals with a certain degree of anonymity. Both crowd types have different uses. For instance, utilizing an internal crowd is more effective for a "wisdom of crowds" type crowdsourcing scenario because of the shared characteristics and the availability of common knowledge among the individuals of the internal crowd.

2.2 Crowdsourcing Quality Assurance

The quality of results is directly influenced by crowd characteristics. In the literature, work quality has been related to crowd demographics (Ross, Irani, & Silberman, 2010; Sheng, Provost, & Ipeirotis, 2008), contributors' gender, profession and age (Downs, Holbrook, Sheng, & Cranor, 2010), and other worker characteristics (Kazai, Kamps, & Milic-Frayling, 2011).

Crowds' failure to produce products that comply with the criteria of acceptable quality is either because of the erroneous submissions made by individuals or because of a willing act to cheat the system. These two different causes of problem can be handled with different methods. For instance, honest mistakes made by workers can be avoided by careful task design, appropriate task granularity (Hossfeld, Hirth, & Tran-Gia, 2011) and the information provided about the task procedure (Downs et al., 2010). On the other hand, identifying cheaters and removing them from the crowd requires tighter quality assurance techniques.

A recent study categorizes quality assurance approaches as *design-time* and *run-time* (Allahbakhsh et al., 2013). Design-time quality assurance consists of good practices of task design, selective worker assignment and data correction methods. Cost of design-time quality assurance basically consists of software development effort to build the crowdsourcing system/tasks or historical data analysis and decision support systems to evaluate worker performance. Design-time quality assurance costs can be estimated by traditional techniques without requiring cost modeling. On the other hand cost of run-time quality assurance techniques depends on the quantity of tasks and probability of erroneous submissions, which require cost modeling.

Below, Table 1 provides a categorization of crowdsourcing quality assurance research according to the techniques applied, and then respective techniques are briefly reviewed. In this study, we take certain design-time characteristics as independent variables of research, and we propose CoQ models for various run-time quality assurance techniques.

Quality assurance mechanisms yield varying cost effectiveness under different conditions. Hirth et al. compare *control group* and *majority decision* techniques in terms of cost effectiveness by examining simulation data. Using probabilistic cost models, they show that both techniques offer the same cheat detection effectiveness but have different costs and applicability. While the *control group* technique is more cost-effective for more complex and expensive tasks, *majority decision* is more cost-effective for simple and inexpensive tasks (Hirth et al., 2013).

***Table 1.* Common quality assurance techniques used in crowdsourcing**

Design-time Quality Assurance		***Run-time Quality Assurance***		
Worker Characteristics	***Design Characteristics***	***Redundancy***	***Control Group***	***Gold Standard***
Reputation (A. J. Quinn & Bederson, 2011)	Defensive task design (A. J. Quinn & Bederson, 2011)	Majority voting (Sheng et al., 2008)	Control group (Hirth, Hoßfeld, & Tran-Gia, 2013)	Gold standard (Oleson, Sorokin, Laughlin, & Hester, 2011)
Selective assignment (Ho & Vaughan, 2012)	Statistical filtering (A. J. Quinn & Bederson, 2011)	Majority decision (Hirth et al., 2013)	Multilevel review (A. J. Quinn & Bederson, 2011)	Injection (Hsueh, Tsai, & Iyer, 1997)
	Bias / error distinction and recovery (Ipeirotis, Provost, & Wang, 2010)	Multiple annotations (Sorokin & Forsyth, 2008)	Grading / voting (Sorokin & Forsyth, 2008)	Ground truth seeding (A. J. Quinn & Bederson, 2011)
	Granularity (Hossfeld et al., 2011)	Repeated labeling (Sheng et al., 2008)	Validation review (Kern, Thies, Bauer, & Satzger, 2010)	
		Redundancy (A. J. Quinn & Bederson, 2011)	Improving review (Kern et al., 2010)	
		Input / output agreement (von Ahn & Dabbish, 2008)		

2.3 Cost of Quality

The aim of any attempt for quality improvement is not limited with achieving quality but also with doing it at the lowest possible cost (Schiffauerova & Thomson, 2006). Numerous studies in the literature address cost optimization of common quality assurance techniques (Hirth et al., 2013; Karger, Oh, & Shah, 2011; Okubo, Kitasuka, & Aritsugi, 2013; Welinder & Perona, 2010).

CoQ is defined as the overall costs undertaken for assuring the quality of a work product. Initial models expressed CoQ as the sum of prevention, appraisal and failure (P-A-F) costs (Feigenbaum, 1956). Simply, these costs are categorized as conformance and non-conformance costs. Conformance costs refer to costs associated with the prevention of poor quality, whereas non-conformance costs are the costs incurred due to poor quality (Crosby, 1979). Quality appraisal and defect prevention costs are considered as conformance costs. Costs of errors surfaced after product delivery, non-detected errors yet to be found, non-conformances detected via quality assurance measures and rework performed to fix detected non-conformances are non-conformance costs.

It should be noted that even if the work involves no monetary payment, and a crowd is performing tasks for another reason, any workforce remains a scarce resource. Deciding to spend effort for quality assurance purposes rather than performing new tasks introduces additional costs. Especially in enterprise crowdsourcing (Vukovic, 2009), significant additional costs exist, since the crowd consists of an organization's personnel whose primary job is not performing the crowdsourced tasks, and effort not spent on primary jobs results in lost revenue for the organization.

Due to difficulties of governing a crowd of workers, the share of CoQ in the overall cost is generally higher compared to traditional production processes. Major CoQ categories and examples of crowdsourcing scenarios are listed in Table 2.

***Table 2.** Major types of CoQ and examples in a crowdsourcing setting*

Type	Description	Example in a crowdsourcing setting
Cost of Conformance		
- Prevention costs	Costs incurred in activities to prevent the end result from failing the quality requirements	Robust design, fitting granularity, easy to use interface
- Appraisal costs	Costs incurred to finding errors	Using a control group to detect faulty submissions
Cost of Non-conformance		
- Internal Failure (rework + retest)	Costs incurred due to non-conformances detected via quality assurance measures	Reassigning a microtask instance because the worker fails to make a submission which complies with the gold standard
- External Failure (errors emerge)	Errors surfaced after product delivery	Majority of the people translating the same work makes a deliberate cheat attempt and the wrong translation is displayed on a user's screen
- External Failure (other)	Harm done to the community or trust mechanisms	Attracting cheaters by continuously failing to detect cheat attempts, or discouraging honest contributors by frequently denying high quality submissions by mistake

Different quality assurance techniques will incur different costs of conformance and non-conformance. Since non-conformance may result in lost reputation and profit to an unknown extent, it is considered as more risky, thus, practitioners often tend to minimize non-conformance. Utilization of additional quality assurance techniques cause the cost of non-conformance to decrease, while expectedly increasing the costs of conformance. Therefore, in order to optimize quality costs, analyzing conformance and non-conformance costs is imperative.

3. CROWDSOURCING COST MODELS

Quality assurance techniques are used to prevent or detect low quality. A generic microtask crowdsourcing process is shown in Figure 2. Microtasks are represented by unlabeled small boxes and labeled boxes depict potential outcomes. The potential outcomes of quality assurance processes constitute a finite set, consisting of the values True Positive (TP), True Negative (TN), False negative (FN) and False Positive (FP), with the probabilities P_{TP}, P_{TN}, P_{FN} and P_{FP}, respectively. In cases when the quality assurance technique fails to reach a decision about the submission, the outcome is Inconsistent (IC). The probability to reach an IC outcome is represented by P_{IC}.

Each outcome results in different costs. For example a TN or IC outcome requires rework thus increasing costs by the cost of 1 task. On the other hand a FN outcome not only results in rework but also introduces additional costs due to failure. The present study introduces crowdsourcing cost models which can be used to estimate costs of common crowdsourcing quality assurance techniques. These models are derived according to probability of quality assurance process outcomes.

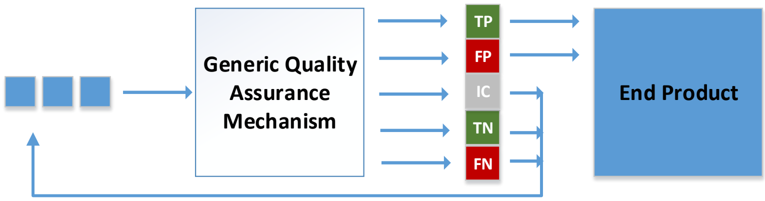


***Figure 2.** Possible outcomes of a generic quality assurance mechanism*

3.1 COST MODELS

Conformance costs vary depending on the quality assurance process design. These do not include direct costs. Direct cost is the total cost of the job when all tasks are performed in perfect quality and no measures are added for quality assurance. Defects detected by the quality assurance techniques are referred to as internal failures (IF). The probability of occurrence of an IF is represented by P_{IF}. Errors which cannot be detected are passed on to the end product, potentially resulting in external failures (EF). The probability of occurrence of an EF is denoted by P_{EF}. Non-conformance costs are equal to the sum of IF and EF costs. Total CoQ is the sum of costs which emerge due to all outcomes of respective quality assurance techniques:

$$CoQ = CoC + C_{IF} + C_{EF} \qquad (1)$$

In order to achieve a complete end product, it is assumed that all outputs which fail to comply with the quality criteria need to be replaced, therefore IF causes rework and retest.

The consequences of EF such as impacts on business continuity, warranties, customer loss or even legal actions, are often difficult to map to monetary costs. In this paper such costs are represented as C_{err}. C_{err} largely depends on the end product and the business domain in which the product is to be used.

Furthermore, when quality assurance techniques fail to distinguish between poor and high quality, long term problems may arise regarding trust mechanisms and crowd behavior. If workers' good quality submissions are being frequently rejected, they may change their behavior and cease to complete tasks in good faith. Similarly, if cheaters observe that their poor quality contributions are often being accepted, they are encouraged to continue cheating. The damage done to the worker community, employer reputation and trust mechanisms are denoted as C_{dmg}. C_{dmg}, by definition, is a common variable for all crowdsourcing initiatives and currently there is no way to estimate or control this type of damage and its long lasting effects. However this does not mean that it should be ignored. A good practice is to use C_{dmg} as a risk / cost adjustment factor within the CoQ calculations.

Table 3 shows the outcomes of a generic quality assurance process and different categories of non-conformance raised by those outcomes.

Table 3. Quality assurance process outcomes and respective non-conformance costs

Non-conformance costs		***Outcomes***	***Cost***
IF	Rework and retest	TN, FN, IC	C_{IF}
EF	Undetected error emerging in the end product	FP	C_{err}
	Damage done to trust system and worker community by falsely rejecting good submissions or approving poor quality contributions.	FP, FN	C_{dmg}

The cost models explained in this section can be used to estimate the cost of utilizing respective quality assurance techniques in a crowdsourcing scenario. In order to use these models, first, the run-time quality assurance techniques in the crowdsourcing scenario need to be identified.

Cost models are derived by multiplying the probability of an outcome with its estimated impact. Thus, probabilities of outcomes need to be known in advance. These values can either be obtained from empirical experiments such as the ones covered in this paper or a pilot project can be initiated to observe outcome probabilities.

3.1.1 Redundancy

To achieve quality assurance via *redundancy* (Figure 3) multiple instances of the same microtask are assigned to different workers who perform the tasks separately. Multiple results are then aggregated to build the final product.

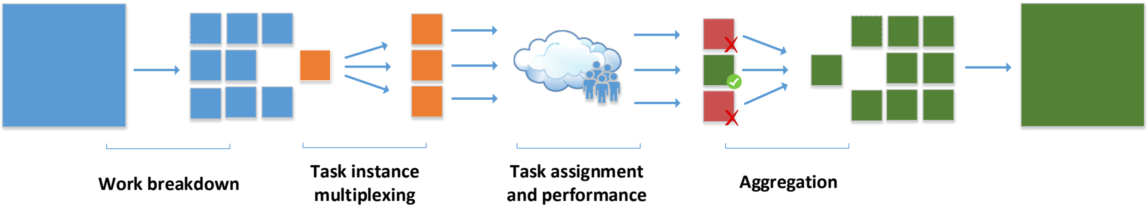


Figure 3. Redundancy quality assurance process

The aggregation step consists of selection of the result with best perceived quality among the set of submissions produced as a result of completing the instances of the same microtask. Selection can be made automatically or manually. Automatic selection is possible when the tasks are *subjective with a finite set of potential results.* This way results can be compared and the frequency of each submission can be determined automatically, and the most frequent submission can be assumed to be the best result. Manual aggregation can be performed by a different set of workers or domain experts. This is basically utilization of *control group.*

Redundancy will lead to decreased resource efficiency. Thus, using cost models when designing crowdsourcing tasks is vital for optimizing resource utilization.

Direct cost of any microtask is assumed to be C_0. The end product consists of outputs produced as a result of N microtasks. The conformance cost of *redundancy* (CoC_{Red}) is caused by the repeated work and output aggregation. Completing m multiple instances of a single microtask as a means of assuring quality increases the costs *(m-1)* times C_0 plus the costs of aggregation: C_{agg} :

$$CoC_{\mathrm{Re}\,d} = N \cdot ((m-1) \cdot C_0 + C_{agg}) \qquad (2)$$

In contrast to other quality assurance techniques, in *redundancy*, rework occurs only when the outcome is IC, with probability P_{IC}. The cost of rework and retest of one submission is m . C_0 + C_{agg}:

$$C_{IF} = N \cdot P_{IC} \cdot (m \cdot C_0 + C_{agg}) \qquad (3)$$

With probability P_{FP}, an EF occurs and leads to potential error in the end product (C_{err}) and damages the reputation and trust mechanisms (C_{dmg}) :

$$C_{EF} = N \cdot P_{FP} \cdot (C_{err} + C_{dmg}) \qquad (4)$$

3.1.2 Control Group

In *control group* techniques, submissions of the main group of workers are controlled by a separate group (Figure 4). The simplest forms of controlling are *voting* and *rating*. Voting is the act of indicating a choice among a set of similar options. In crowdsourcing voting refers to a separate task that is carried out by a different group of people than the ones performing the main task. Generally voting is done at a binary nominal scale (Yes/No, Pass/Fail, Like/NA, Selected/Unselected). Rating is defined as *classification or ranking based on a comparative assessment*.

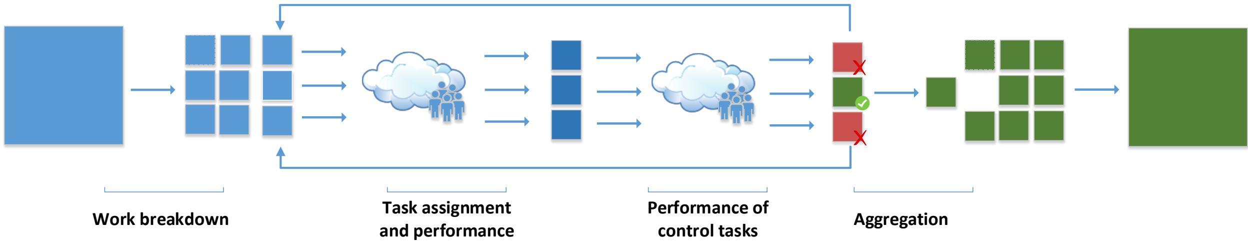


Figure 4. Control group quality assurance process

Direct cost of any task is assumed to be C_0 and the cost of controlling the outputs of one task is C_1.

Conformance costs in *control group* (CoC_{CG}) techniques are incurred by control tasks (5). Generally controlling outputs of a microtask is significantly less complex and thus costs less.

$$CoC_{CG} = N \cdot C_1 \qquad (5)$$

When the controlling workers decide that the submission does not comply with quality criteria, the task output is rejected and rework and retest are needed to replace that product. *Control group* either identifies poor quality work correctly or incorrectly giving the probability of a work output to be rejected as $P_{FN} + P_{TN}$. The cost of rework and retest is $C_0 + C_1$:

$$C_{IF} = N \cdot (P_{FN} + P_{TN}) \cdot (C_0 + C_1) \qquad (6)$$

An erroneous work output can be placed among the end product only if the *control group* incorrectly accepts it. The cost, C_{err} is incurred when an EF occurs in the end product. Whenever the *control group* fails to detect a poor quality submission (P_{FP}) or identifies a good quality output of a microtask as invalid (P_{FN}), damages occur to the trust mechanisms and worker community (C_{dmg}) :

$$C_{EF} = N \cdot ((P_{FP} + P_{FN}) \cdot C_{dmg} + P_{FP} \cdot C_{err}) \qquad (7)$$

3.1.3 Gold Standard

Also referred to as ground truth seeding (A. Quinn & Bederson, 2011) *gold standard* is basically a set of trusted inputs (labels, annotations, etc.) inserted among the data, which constitute expected results for certain tasks. If contributions of a worker deviate significantly from the trusted, -gold standard- result, measures are taken to improve quality (Huang, Zhang, & Parkes, n.d.; McCann, Shen, & Doan, 2008; Sorokin & Forsyth, 2008). The worker can be provided with immediate feedback including the gold standard response to ensure that expectations are understood clearly (Ipeirotis et al., 2010). This has an improving effect on submission quality, whether the gold standard comparison is made for training the user before moving on to the real tasks (Le & Edmonds, 2010), or randomly carried out within the task performing process (Figure 5). Incompatible submissions of workers are tracked to reveal a potential pattern in order to identify cheaters. Submission patterns of workers are used to define individual reputation which can be used to establish a trust evaluation infrastructure for the crowdsourcing system or platform (Voyer et al., 2010).

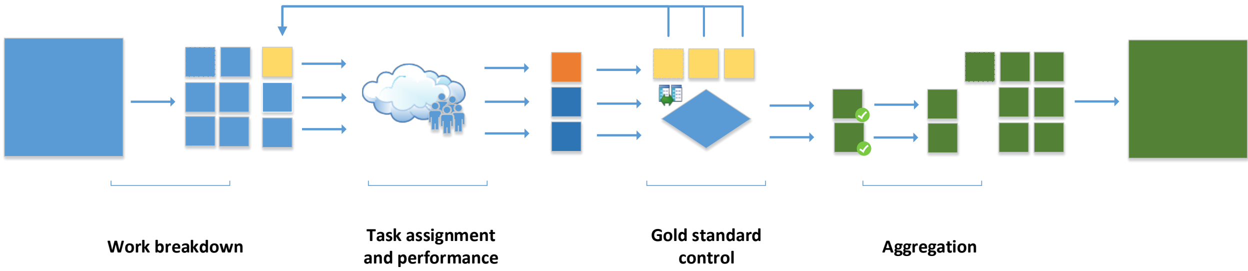


Figure 5. Gold standard quality assurance process

The sample size of gold standard tasks must be large enough, so that probability of the same worker being assigned with the same gold standard tasks within the process is quite low. However, establishing a large gold standard data set can result in a significant cost increase. In some cases the gold standard task pool can be enriched by dynamically altering the pool content (Oleson et al., 2011; von Ahn & Dabbish, 2008).

Gold standard technique can be used asynchronously or synchronously. In asynchronous usage gold standard tasks are assigned to the workers separately from regular tasks, usually in the form of qualification or training. In synchronous usage, gold standard tasks are assigned together with a number of regular tasks.

Direct cost of any task is assumed to be C_0 and cost of introducing one gold standard task into the system is C_{exp}.

Expression (8) represents the conformance costs (CoC_{GS}) for synchronous *gold standard* usage where $(k / t - k)$ is the ratio of the number of gold standard tasks to the number of regular tasks which are assigned together and X is the total number of tasks in the gold standard pool. k is the number of gold standard tasks assigned to a batch of t tasks.

$$CoC_{GS} = X \cdot C_{\exp} + N \cdot (\frac{k}{t-k}) \cdot C_0 \qquad (8)$$

Internal failure occurs with probability $(1 - (P_P)^k)$ when a worker submits an incorrect result for at least one of the gold standard tasks in a batch. The impact of this is the cost of rework and retest of $(t - k)$ regular tasks and k gold standard tasks:

$$C_{IF} = N \cdot (\frac{k}{t-k}) \cdot (1 - (P_P)^k) \cdot t \cdot C_0 \qquad (9)$$

EF occurs when the worker submits a valid result for gold standard tasks while providing poor quality contributions for regular tasks. The probability of a worker making an invalid submission for a regular task is P_W. Similar to the other quality assurance techniques EF costs also include the damage inflicted to the worker community when contributors' submissions are falsely evaluated. The cost of EF for *gold standard* techniques is given in (10). P_W is not represented in terms of P_{FP} and P_{TN} because the expression covers various situations in which the number of gold standard tasks and the number of regular tasks differ. For instance, using 1 gold standard task with 1 regular task results in producing 4 outcomes (TP, FP, TN, FN) and P_W can be expressed as the sum of P_{FP} and P_{TN}. In other cases using P_{FP} and P_{TN} to express P_W increases the complexity of the model.

$$C_{EF} = N \cdot (\frac{k}{t-k}) \cdot (P_P)^k \cdot P_W \cdot (t-k) \cdot (C_{err} + C_{dmg}) \qquad (10)$$

3.2 Cost Estimation Process

In this section we define a process which describes the utilization of cost models in practice (Figure 6). The first step in cost estimation is to identify suitable quality assurance techniques. This decision should be based upon good practices or practitioner's experience. For instance, if the main task is significantly more complex than the control task, utilizing a *control group* technique is reported to be more cost effective than *redundancy* (Hirth et al., 2013).

The second step is to obtain the values (P_P, P_{IC}, P_W, P_{FP}, P_{FN}, P_{TN}, P_{TP}). In order to measure probability values (P values), a pilot study can be conducted in an environment which closely represents real life. Another option is to use P values reported in the results of other projects. Before constructing the cost model, C_{err} and C_{dmg} values need to be determined based on specific characteristics of the work and the practitioner's management style.

The last step consists of constructing the model based on P, C_{err} and C_{dmg} and estimating the resulting costs. This can be done by simply calculating the cost model formula with obtained parameters, or by applying techniques such as Program Evaluation and Review Technique (PERT) (Malcolm, Roseboom, Clark, & Fazar, 1959), or Monte Carlo (Fishman, 1996) to obtain optimistic, pessimistic and most-likely estimates.

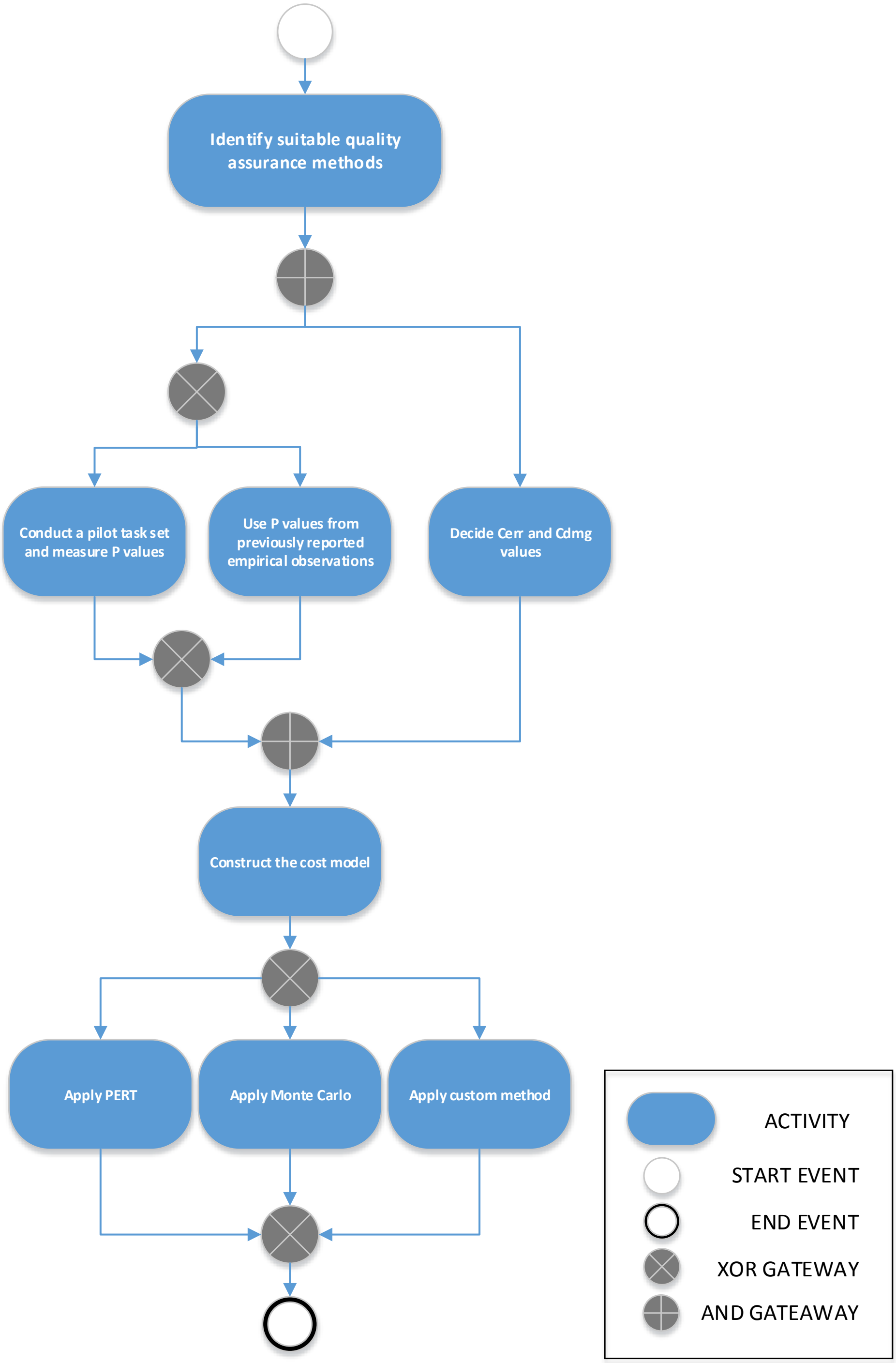

Identify suitable quality assurance methods
Conduct a pilot task set and measure P values
Use P values from previously reported empirical observations
Decide Cerr and Cdmg values
Construct the cost model
Apply PERT
Apply Monte Carlo
Apply custom method
ACTIVITY
START EVENT
END EVENT
XOR GATEWAY
AND GATEAWAY

Figure 6. Sample cost model utilization process

4. COQ EVALUATION

In this study we used action research method which is a special type of case research in which the researchers not only observe but also participate in solution process (Wieringa & Morali, 2012). In order to assess the validity and applicability of the proposed CoQ techniques, we conducted multiple action research, based on three different real-life crowdsourcing cases. Each case covers different design-time characteristics of *nature of task* and *crowd type* as shown in Table 9. The same run-time quality assurance techniques were utilized in all three cases.

The primary research goal at this stage was to determine the probabilities of quality assurance outcomes. These probability values were used to construct the cost models. Cross validation techniques were used to determine if the probability values come from the same distribution, thus being predictable and not random.

The first case involves image illustration and evaluation tasks performed by an external crowd on Amazon Mechanical Turk (AMT) ("Amazon Mechanical Turk," n.d.). These tasks can be classified as subjective. The second case covers a big data cleaning solution which consists of objective tasks carried out by an external crowd on AMT. The third case comprises a phonebook registry update problem which includes subjective tasks performed by an internal crowd.

Measurements and calculations made in action research cases use common parameters. Definitions of common parameters are given in Section 1 and below in this section.

Taking the definition of cheating as *the act of a contributor to make poor quality submissions whether because of malevolent intentions or simply an attempt of maximizing personal gain*, cheat probabilities are measured simply by comparing individual submissions against the expert evaluation. Cheat probability is the sum of P_{FP} and P_{FN} and is denoted by $P_{W.}$

Redundancy quality assurance process reaches an IC outcome only if the number of elements in the result set is not less than the number of redundant submissions or when the number of redundant submissions is even. In all three cases the number of redundant tasks is odd (*m*=3), thus reaching an IC is not possible. Nevertheless this parameter is preserved for completeness.

Observed P values of all outcomes are reported in respective tables presented at the end of the section describing the study on each case.

Cost models of *gold standard* quality assurance techniques include the parameter of cost of an expert introducing 1 gold standard task into the system (C_{exp}), which is common to all action research cases. This parameter is assumed to be $10 \, . \, C_0$.

Finally, C_{prod} represents the total cost of product excluding all quality related costs. C_{prod} is used to adjust C_{err} values and to normalize CoQ for comparison. The effectiveness of the quality assurance techniques were assessed according to the *Decision Fitness* (DF) measure (11). CoQ values and DF are used together to compare cost effectiveness of quality assurance mechanisms.

$$DF = P_{TN} + P_{TP} \qquad (11)$$

The validity of the observations was checked via the V-fold cross validation technique (Arlot & Celisse, 2010).

Cross validation is a simple and universal method used to estimate risk of an estimator and model selection. The basic idea behind v-fold cross validation is to split the data into v subsamples. Each subsample successively acts as the validation portion whereas the others are used for training. This process is repeated until all subsamples are used once as the validation portion. We applied cross validation on the probability observations of quality assurance technique outcomes. In each repetition Magnitude of Relative Error (MRE) was calculated. Mean MRE (MMRE) values were used to evaluate the validity of the observations.

4.1 Case 1: Illustration and Evaluation of Simple Images: CoQ of Subjective Microtasks on AMT

4.1.1 *Description*

This action research addressed the production process of a large number of hand-drawn simple images to be used in design of brand merchandise with the concept of *lizards*. The business goal of this action research was to produce at least 200 illustrations which unmistakably resemble lizards. Rather than using artists to draw the illustrations, the job was assigned to the crowd in order to reflect the perception of a wide variety of people and produce a diverse set of images. The research goal was to observe the process outcomes of common crowdsourcing quality assurance techniques when applied on subjective tasks. In the first phase workers were asked to draw an illustration of a lizard. At the end of this phase, the image set produced by the crowd was expected to contain many good and poor quality illustrations. Therefore, in the second phase separate groups of workers were asked to evaluate the images in terms of resemblance to a lizard. Three different crowdsourcing designs were used employing various common crowdsourcing quality assurance techniques. All user actions were logged for analyzing the costs and the quality. The quality of both primary (lizard drawing) and secondary (image evaluation) tasks were decided by comparing the submissions against the expert judgment. The details of this action research can be found in (Author, 2014a).

4.1.2 *Method*

Both primary and secondary tasks were published on AMT. Workers performing the primary task were provided with an online, open-source canvas editing utility (“Literally Canvas,” n.d.) and were asked to draw an illustration of a lizard. Upon successful completion of each task workers were paid $0.15. Task success was determined based on expert evaluation. The entire image set, consisting of 504 images, was evaluated by the researcher. Three separate groups of workers performing secondary tasks were provided with links to three different external web applications according to the group they belonged. Each worker was restricted to submitting one judgment only. The instructions specified that correct judgments were to be paid $0.01 and others were to be rejected. The correctness of the control tasks was decided based on comparison against the expert evaluation.

In *Control Group Voting* (CG voting), workers were shown a random image from the lizard image data set and asked if the image resembles a lizard or not. The evaluations were made in binary scale; *yes* or *no*.

Control Group Rating (CG rating) is almost the same as CG voting, with the only difference being that the evaluations were performed on a 5-level Likert scale rather than binary. In the analysis, 4 and 5 were considered as positive and 1, 2 and 3 as non-positive ratings.

In *Gold Standard Rating* (GS rating) workers were shown two different images at the same time. One of the images came from the lizard image set while the other was from the gold standard image set. The gold standard image set consisted of 40 images; half of them were good examples of lizard illustrations and the other half were clearly not lizard images. Evaluations were made on a 5 point Likert scale, for both images separately. If the worker failed to provide a valid rating for a gold standard image then the system rejected the submission and displayed a warning to the worker.

Workers continued performing the secondary tasks until all images in the lizard image set were evaluated three times, applying the *redundancy* technique. These redundant evaluations were used to derive a majority decision. Therefore, all three designs were evaluated both with and without *redundancy*.

4.1.3 *Measurements*

In total, 504 images were submitted by the workers. 27 obvious cheat attempts were detected by expert review in primary tasks. A total of 5,183 control tasks were performed which consisted of 504 expert evaluations, 1,512 CG voting, 1,512 CG rating and 1,655 GS rating submissions. 143 invalid gold standard submissions were received.

Cheat probabilities for each design were measured by comparing individual submissions against the expert evaluation. Denoted by P_W, cheat probability for the primary task was reported to be 0.34. P_W values for secondary tasks are shown in Table 4.

Probability outcome values of quality assurance processes are shown in Table 4. Table 4 omits the P_{FP} value for single GS rating design, because rather than this probability value, P_P and P_N values are used in cost models for GS rating design. Representing the probability of a worker to submit a negative result to a gold standard task, P_N is observed to be 0.09 for GS rating, and P_P is 0.91 as expected.

Table 4.** **Probability values of quality assurance process outcomes

		P_W	P_{IC}	P_{FP}	P_{FN}	P_{TN}	P_{TP}
CG Voting	Single	**0.25**	N/A	**0.16**	**0.09**	**0.19**	0.55
	With Redundancy	-	0.00	**0.15**	0.05	0.20	0.58
CG Rating	Single	**0.34**	N/A	**0.06**	**0.28**	**0.29**	0.37
	With Redundancy	-	0.00	**0.03**	0.26	0.32	0.39
GS Rating	Single	**0.31**	N/A	-	-	-	-
	With Redundancy	-	0.00	**0.08**	-	-	-
Expert Evaluation		**0.00**	**N/A**	**0.00**	**0.00**	**0.35**	**0.65**

4.1.4 Validation

V-fold cross-validation (Arlot & Celisse, 2010) of the observed probability outcomes reported in Table 4, yields the following average MMRE values, where V is 15 and group size is 100:

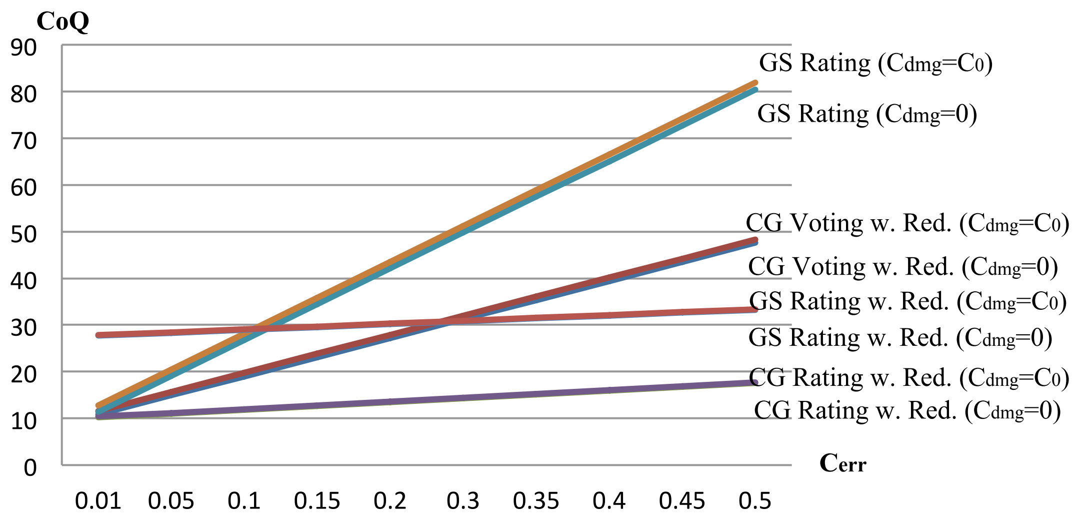
- $MMRE_{CG\ voting} = 0.12$
- $MMRE_{CG\ rating} = 0.15$
- $MMRE_{GS\ rating} = 0.14$

MMRE values smaller than 0.2 are considered acceptable for prediction models (Conte, Dunsmore, & Shen, 1985).

4.1.5 CoQ Calculations

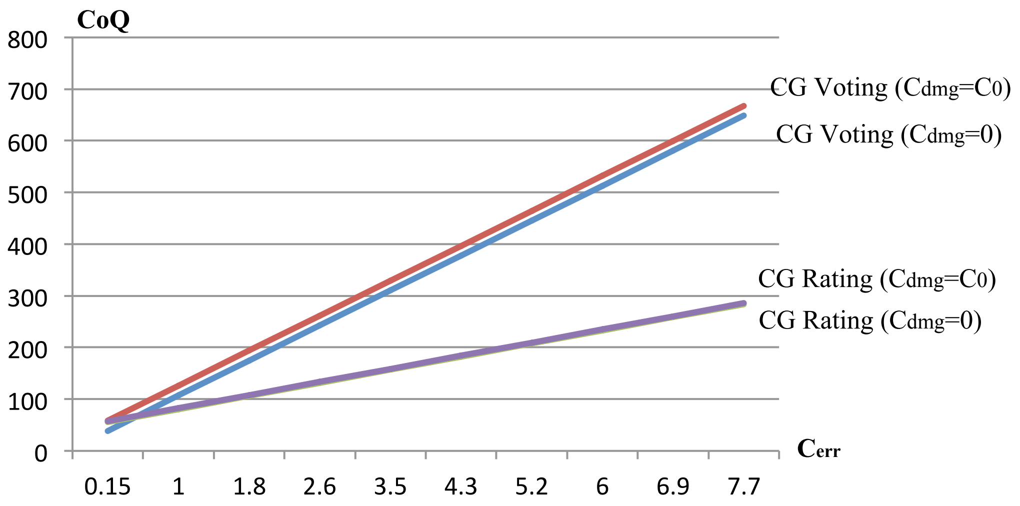
Cost formulas presented in Section 3.1 are used to calculate CoQ for three designs: CG voting, CG rating and GS rating, both with and without redundancy. Probability values provided in Table 4 were used as parameters in CoQ formulas.

In this particular case two separate values were used for C_{dmg} ($C_{dmg} = 0$, $C_{dmg} = C_0$) and a value interval was provided for C_{err} with the lower and upper limits of ($C_{err} = C_0$, $C_{err} = 0.1 \ . \ C_{prod}$) while C_{prod} is the total direct cost of producing the complete product, which is calculated as $C_{prod} = 504 \ . \ 0.15{=}75.16$). C_{prod} varies for primary and secondary (control) tasks. C_{err} and C_{dmg} values are used to observe the effect of changing impact on total CoQ and the results are displayed in Figure 7.



GS Rating	$= 40 . C_{exp} + 5.95 + 142,18 . C_{err} + 142,18 . C_{dmg}$
CG Voting w. Red.	$= 10,08 + 75,6\ C_{dmg} + 75,6\ C_{err}$
GS Rating w. Red.	$= 40 . C_{exp} + 27,62 + 11,37 . C_{err} + 11,37 . C_{dmg}$
CG Voting w. Red	$= 10,08 + 15,12\ C_{dmg} + 15,12\ C_{err}$

(a)



CG Voting	$= 27.62 + 126 . C_{dmg} + 80.64 . C_{err}$
CG Rating	$= 51.01 + 171.36 . C_{dmg} + 30.24\ C_{err}$

(b)

Figure 7. ***The effect of changing Cerr and Cdmg on CoQ of various crowdsourcing designs***

4.1.6 Findings

Figure 7a shows CoQ of GS rating, GS rating with redundancy, CG voting with redundancy and CG rating with redundancy designs. 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 CoQ, due to high initial quality costs of GS rating with redundancy design. Using *redundancy* in GS rating leads to a higher CoQ 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 fewer errors to remain undetected.

Figure 7b shows the CoQ of CG voting and CG rating designs for varying C_{err} values. We observe that CG rating design is more likely to detect a submission as invalid, compared to CG voting ($P_{(TN+FN)\ CG\ rating} = 0.57$ and $P_{(TN+FN)CG\ voting} = 0.28$). This makes rating a more strict method of controlling than voting which may lead to less undetected errors. According to these findings it is concluded that a rating scheme is preferable to voting when EF tolerance is low but IF is more acceptable.

4.2 Case 2: Big Data Analysis: CoQ of Objective Microtasks on AMT

4.2.1 Description

This action research addresses a data cleaning and migration project recently undertaken in the Middle East Technical University (METU). Recently a project was initiated to integrate key components of the university's IT structure as automated business processes. This major overhaul caused some of the legacy data to be migrated to newly developed systems. METU employs over 2,500 academic personnel who are actively engaged in research and produce a large amount of publications. The records of academic accomplishments of METU personnel are kept in a legacy application. This application was designed to allow users to enter their publication records in free text format. Thus, the data contained many duplicates and typographical errors. Initially there were 53,822 records in the legacy database. The business goal of this action research was to normalize the data, to clean the duplicates, to fix typographical errors and to migrate the data to the newly developed system. The details of this action research can be found in (Author, 2014b).

4.2.2 Method

In order to solve this data cleaning and migration problem, a multistage, hybrid solution approach was taken. First, CrossRef ("CrossRef," n.d.) external Digital Object Identifier (DOI) web service was used to tag the publications with matching DOIs. As a result of the DOI resolution process 5,681 (10.56% of entire record set) records were matched with a DOI.

The second stage consisted of executing custom developed string similarity algorithms to detect the records that are either identical or clearly distinct. Primarily, DOI tags were used in comparison. If the record did not have a DOI, the title, authors, publisher and publication date fields were used. Upon completion of this stage, 4,558 records were identified as the same while 38,830 records were clearly distinct. These records were removed from the data set.

The remaining 10,434 records could not be classified either by querying the external web services or string similarity algorithms, still leaving too many records to be processed manually.

The crowdsourcing stage aimed at leveraging the strengths of human cognition in order to identify the duplicates and errors within the residual record set.

First, all similar records were gathered in pairs. This increased the number of tasks to be crowdsourced due to recurring records in multiple pairs. This arrangement enabled the researchers to ask the question in a way which limits the workers with binary answers: "Is the following record pair the same or different?" The total number of tasks was 9,308.

These tasks were posted on AMT as Human Intelligence Tasks (HIT). In each HIT, workers were asked to evaluate 4 record pairs. Upon successful completion they were paid 0.02$.

In the crowdsourcing stage multiple quality assurance techniques were utilized. These techniques included *redundancy*, *control group* and *gold standard*.

In order to apply the *gold standard* technique, a set of 100 gold standard pairs were constructed. 50 of these pairs consisted of identical pairs whereas the remaining 50 were unmistakably different. Each HIT contained 1 gold standard pair and 3 regular pairs, appearing in random order each time a HIT is displayed. Each microtask was assigned to 3 different workers for quality assurance purposes. Additionally, the majority decision was controlled by a separate group of workers.

4.2.3 Measurements

Worker activities were logged. 9,308 pairs were evaluated by the workers, judging the pair equality. Each pair was evaluated by 3 distinct workers. In total 29,844 tasks were performed including 1,920 gold standard failures. The results of these tasks were controlled by a different set of workers. 9,938 control tasks were performed including 630 gold standard failures.
As the results of majority decision, 6,225 pairs were decided as equal and 3,083 pairs were decided as different.

Finally, 6,102 pairs were evaluated manually by experts for validation purposes. The outcomes of quality assurance techniques were examined by comparing the decisions against expert judgments. The occurrence counts of observed quality assurance process outcomes are shown in Table 5.

***Table 5.* The occurrence counts of quality assurance process outcomes**

	FP	*FN*	*TN*	*TP*
Control Group	**131**	**1189**	**261**	**4521**
Redundancy	**392**	N/A	N/A	**5710**
Gold Standard	-	-	-	-

The probability values of quality assurance process outcomes are derived by calculating the percentage of particular occurrence of an outcome within all possible outcomes and shown in Table 6. P_W of *gold standard* is not the ratio of workers failing the gold standard question, but is the ratio of passing the gold standard and failing to provide a good quality submission. In this case P_N value for gold standard tasks was observed as 0.06 while P_P was observed as 0.94.

***Table 6.* Probability values of quality assurance process outcomes**

	P_W	P_{IC}	P_{FP}	P_{FN}	P_{TN}	P_{TP}
Control Group	**0.22**	N/A	**0.02**	**0.20**	**0.04**	**0.74**
Redundancy	**-**	N/A	**0.06**	N/A	N/A	0.94
Gold Standard	**0.17**	N/A	-	-	-	-
Expert Evaluation	**0.00**	**N/A**	**0.00**	**0.00**	**0.00**	**0.00**

4.2.4 Validation

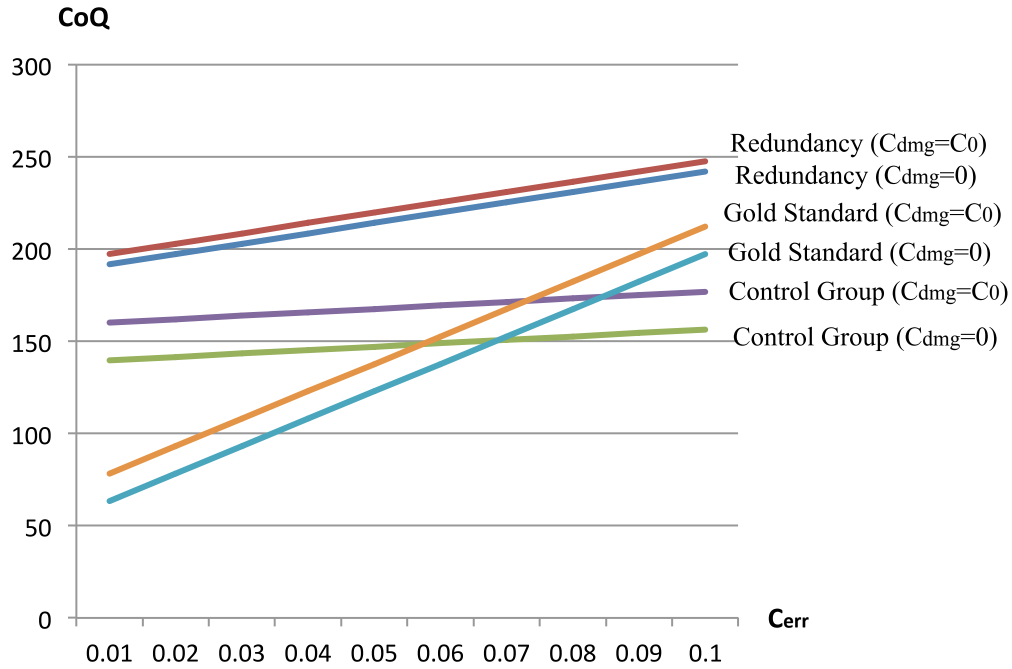
The observations were validated by using V-fold cross validation technique which is explained in the Section 4. V-fold cross validation yields the following MMRE results, where V is 15 and group size is 400:

- $MMRE_{CG} = 0.10$
- $MMRE_{Red} = 0.07$
- $MMRE_{GS} = 0.09$

MMRE values smaller than 0.2 are considered acceptable for prediction models (Conte et al., 1985).

4.2.5 CoQ Calculations

Probability values in Table 6 are used for CoQ calculations. In order to observe the change of CoQ of quality assurance techniques, two separate C_{err} and C_{dmg} values are used (Figure 8). In this particular case C_{dmg} is assumed to be equal to 0 or C_0. Considering the simplicity of the task and low level of criticality, lower bound of C_{err} is assumed to be C_0 and the upper bound is equal to $10 . C_0$.



Redundancy $= 186.16 + 558.48 . C_{dmg} + 558.48 . C_{err}$

Gold Standard $= 100 . C_{exp} + 38.47 + 1487.42 . C_{err} + 1487.42 . C_{dmg}$

Control Group $= 137.76 + 2047.76 . C_{dmg} + 186.16 . C_{err}$

***Figure 8.** The effect of changing Cerr and Cdmg on CoQ of various crowdsourcing designs*

4.2.6 Findings

In this setting microtasks were objective. Control tasks and primary tasks had similar complexity, thus the costs of primary and secondary tasks were equal. In such a setting, with given parameters, *control group* technique was observed to be the most robust technique against increasing values of C_{err}. However, when C_{err} is smaller than 0.6 *gold standard* produces lower CoQ results. On the other hand CoQ of *redundancy* is the highest and increases significantly at a higher rate than other quality assurance techniques, when C_{err} increases.

4.3 Case 3: Campus Phonebook Registry Update: CoQ of Objectıve Wisdom of Crowds Type Tasks

4.3.1 Description

This action research was also conducted in METU. In late 2011 a project was initiated to establish the corporate identity of the university. The project mainly consisted of developing social media identities and transferring websites to a corporate content management system. The project also included a work package for updating the phonebook registry. METU has two separate phonebook applications owned by different administrative units. Both applications contain outdated information and no automated mechanism exists to keep the phonebook registry up to date. Currently METU employs over 2,500 academic and 3,100 administrative personnel. There are more than 5,500 phone numbers assigned to the personnel. The business goal in this case was to update the corporate phonebook with accurate assignments.

4.3.2 Method

To solve the phonebook registry update problem an application with social features was developed, deployed on the university intranet and made available to all university personnel through the university portal application. With an email sent to the organization-wide mailing list, all personnel were asked to update their phone numbers. By using this application users were able to update their own phone number entry or submit phone numbers of their colleagues. The software kept detailed logs of user actions for data analysis.

In this enterprise crowdsourcing setting the crowd consisted of 5,500 university personnel. The microtasks were objective. Rather than the cognitive capacity of workers, this type of crowdsourcing aims at utilizing the collective knowledge residing within the crowd. Therefore it can be classified as *wisdom of crowds* (Surowiecki, 2005) type crowdsourcing.

Redundancy, control group and *gold standard* quality assurance techniques were used and the outcomes of quality assurance processes were observed from the user action logs.

4.3.3 Measurements

Data collection phase lasted two weeks and then terminated. 743 unique personnel were tagged with at least 1 phone number by the crowd workers whereas 328 of them were tagged 3 times. Upon agreement of multiple workers, these tags were finalized. After completion, all 328 records were controlled by the crowd workers through the same user interface.

In this case an asynchronous *gold standard* technique was also used. Workers were asked the phone numbers of well-known and frequently used phone numbers such as their department secretaries, deans' offices or campus entrance gates. The system was designed to display 1 gold standard task for 2 regular tasks. If the workers provided incorrect answers for the gold standard question their previous two answers were discarded. Only 4 instances of gold standard task failure were observed out of 164.

A subset of the results which consisted of 328 records was manually checked by experts. Correctness of user answers was decided based on expert evaluation. Observed quality assurance process outcomes are presented in Table 7.

***Table 7.** The occurrence counts of quality assurance process outcomes*

	FP	*FN*	*TN*	*TP*
Control Group	**26**	**5**	**21**	**276**
Redundancy	**18**	N/A	N/A	**310**
Gold Standard	-	-	-	-

The probability values of quality assurance process outcomes are presented in Table 8. In this case P_N for *gold standard* process outcome was observed as 0.02 while P_P was observed as 0.98.

***Table 8.** Probability values of quality assurance process outcomes*

	P_W	P_{IC}	P_{FP}	P_{FN}	P_{TN}	P_{TP}
Control Group	**0.10**	N/A	**0.08**	**0.02**	**0.06**	**0.84**
Redundancy	**-**	N/A	**0.06**	N/A	N/A	**0.94**
Gold Standard	**0.10**	N/A	-	-	-	-
Expert Evaluation	**0.00**	**N/A**	**0.00**	**0.00**	**0.00**	**0.00**

4.3.4 Validation

The observations were validated by using V-fold cross validation technique which is explained in the Section 4. V-fold cross validation yields the following MMRE results, where V is 10 and group size is 32:

- $MMRE_{CG} = 0.38^{*}$
- $MMRE_{Red} = 0.15$
- $MMRE_{GS} = 0.31^{*}$

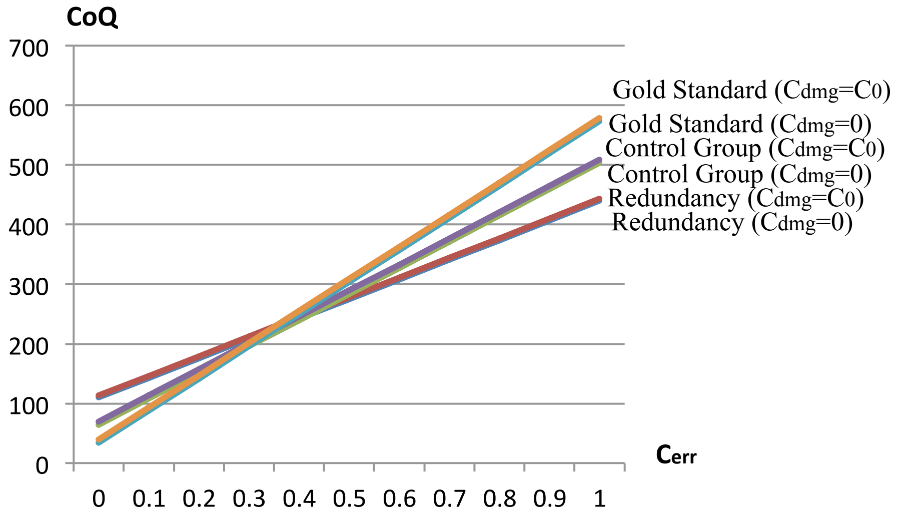
MMRE values smaller than 0.2 are considered acceptable for prediction models (Conte et al., 1985).

(*) Due to small sample size a large variance in cross validation error occurs, which may lead to statistically unreliable results (Rao, Fung, & Rosales, 2008). Therefore validation results for this case are not considered statistically reliable. Cross validation needs to be repeated when more data become available.

4.3.5 CoQ Calculations

Probability values in Table 8 are used for CoQ calculations. In this case an internal crowd was used thus many parameters differ. Even though crowd workers were not paid upon task completion, C_0 and C_1 were assumed to be \$0.01. The entire job consisted of 5,500 tasks but the work was terminated before completion. However, in order to calculate the CoQ for the whole job, total number of tasks was assumed to be 5,500. The number of gold standard phone numbers introduced to the system was 50. Cost of introducing 1 gold standard task into the system was assumed to be equal to 10 times of C_0.

Two different C_{err} and C_{dmg} values were used for observing the impact of change on total CoQ (Figure 9). In this particular case C_{dmg} is assumed to be equal to 0 or C_0. Lower bound of C_{err} is assumed to be C_0 and the upper bound is assumed to be 10 times C_0.



Gold Standard $= 50 . C_{exp} + 29.15 + 539 . C_{err} + 539 . C_{dmg}$
Control Group $= 63.8 + 550 . C_{dmg} + 440 . C_{err}$
Redundancy $= 110 + 330 . C_{dmg} + 330 C_{err}$

<u>Figure 9.</u> The effect of changing Cerr and Cdmg on CoQ of various crowdsourcing designs

4.3.6 Findings

This case is different from the first two because the analysis was conducted before all the tasks were completed. Thus, a smaller amount of data could be collected. However this situation perfectly reflects the real life scenario in which the whole cost of the job needs to be estimated depending on a limited number of initial measurements.

Another difference of this case is that work was done by an internal crowd rather than an anonymous external crowd. Therefore, monetary payments were not made upon task completion. Motivation to participate was different. The observed number of poor quality submissions was significantly lower compared to the other cases. This can be explained by the fact that the identities of the workers were known and workers completed the tasks with a higher sense of accountability compared to anonymous workers.

According to Figure 9, *redundancy* displays a slightly more robust profile against the changing values of C_{err}. However when C_{err} is lower than 0,4 both *control group* and *gold standard* techniques produce lower CoQ results than *redundancy*.

4.4 Discussion

Present research covers various crowdsourcing scenarios which differ in task type and crowd type. Various quality assurance mechanisms were used in each design. Costs of these mechanisms were estimated with an estimation process using probabilistic cost models. This estimation process, which is based on CoQ analysis, can be used for decision making while selecting quality assurance methods for crowdsourcing.

In order to compare the cost of quality assurance techniques, the calculations were normalized to reflect the ratio of CoQ to the total cost of work, excluding the cost of all quality related activities. For this analysis C_{err} is assumed to be equal to C_0. Both normalized CoQ and DF calculations are presented in Table 9.

Table 9. Normalized CoQ calculations and DF values

Case	**Crowd Type**	**Task Type**	**C_{prod}**	**CoQ**		**CoQ/C_{prod}**	**DF**
1	AMT Workers	Subjective	75.6	CG Voting	39.72	0.53	0.74
				CG Rating	55.55	0.73	0.66
			5.04	CG Voting w. Red	10.84	2.15	0.78
				CG Rating w. Red	10.23	2.03	0.71
				GS Rating	11.37	2.26	0.63
2	AMT Workers	Objective	93.08	Control Group	139.62	1.50	0.78
				Redundancy	191.75	2.06	0.94
				Gold Standard	63.34	0.68	0.78

3	Internal Crowd	Objective	55	Control Group	68.2	1.24	0.90
				Redundancy	113.3	2.06	0.94
				Gold Standard	39.54	0.72	0.88

The DF values of Case 3 are significantly higher than Case 1 and Case 2. This can be explained by the fact that Case 3 utilizes an internal crowd with a better sense of accountability compared to the AMT workers. Even though Case 1 and Case 2 uses AMT workers, DF values of Case 2 are higher than Case 1 due to the difference in task types. Obviously, quality assurance techniques applied on objective tasks lead to more effective results. Furthermore, using an internal crowd increases the effectiveness of quality assurance techniques. Thus, the practitioners can invest less on quality assurance when they use an internal crowd.

In Case 2 and Case 3, *redundancy* is observed to be the most expensive technique, while *gold standard* is the least expensive in terms of CoQ/C_{prod}. Using *control group* in these cases lead to lower CoQ compared to *redundancy*, but at the expense of effectiveness.

The CoQ changes according to C_{err} value decided by the practitioners. Some quality assurance techniques provide better CoQ when C_{err} is high. The robustness of a technique against increasing C_{err} values can easily be understood by looking at the slope of the CoQ / C_{err} graph or the coefficient of C_{err} in the cost model formulas. The lower the coefficient, the more robust is the technique.

In the first action research case the *redundancy* and *gold standard* techniques were applied on the secondary task, while *control group* was applied on the primary task. In this sense, the instrumentation slightly differs from the other cases. In order to preserve internal validity, quality assurance techniques of the first case were only compared to the ones which were applied on the same type of task.

None of the participants were individually selected by the researcher. Workers simply answered an open call for participating. Without doubt one of the most important parameters which influences quality assurance process outcomes is crowd characteristics. When a crowd with different characteristics does the work, different results can be expected. In the first two cases, AMT workers were used. The cross validation produced MMRE values which are smaller than 0.2. This indicates that outcomes with similar error rate can be expected if the study is to be repeated, supporting the generalizability claim.

5. **CONCLUSION**

Introducing and maintaining quality assurance techniques inevitably increase project costs. However the crowdsourcing literature lacks defined procedures to estimate the cost of quality assurance. Such procedures may benefit crowdsourcing practitioners as guidelines for selecting and using techniques which provide higher cost effectiveness. Furthermore, massive inefficiencies in resource utilization at a global scale can be avoided through widespread usage of these cost models.

Hirth et al. compare cost effectiveness of two different quality assurance techniques based on simulation data under different cost assumptions. Their approach consists of executing simulations with various cost and cheat probability parameters. As a result, they reported that the control group technique is more cost effective compared to the majority decision when the tasks are complex and high priced. The present study similarly uses outcome probabilities in cost models but differs significantly as it utilizes empirically observed probability outcomes in cost models and these cost models include cost items for all possible quality assurance process outcomes satisfying the needs of a CoQ approach. Furthermore, the present study introduces a generic cost estimation process which can be applied in any crowdsourcing scenario which utilizes run-time quality assurance techniques, including some cases of citizen science.

The main contributions of this research are the cost models of common quality assurance techniques and the CoQ estimation process. The applicability of this estimation process and cost models for different crowdsourcing scenarios was evaluated with multiple action research. CoQ was determined by using the cost models and various quality assurance techniques were compared based on CoQ. The secondary contribution is the observations of probabilistic outcomes of quality assurance processes for different types of work and crowd. These values can be used by other practitioners and researchers as guidelines.

The cost models proposed in this study empower crowdsourcing practitioners with a defined cost estimation procedure which they may use instead of unstructured methods and expert judgment. Enabling formal planning by basing decisions on procedural calculations is especially valuable in enterprise projects which have a low tolerance for uncertainty. As exemplified in the multiple action research cases, the cost estimation process is applicable to various crowdsourcing scenarios. Crowdsourcing practitioners can use the proposed estimation process to build cost models satisfying the specific needs of their projects, or they may use the cost models as they are introduced in the present study especially for projects with characteristics similar to the ones covered in the multiple action research cases.

The impact of this study can be better grasped when the current status of crowdsourcing and citizen science is considered. The crowdsourcing market is still growing. Even though practitioners use crowdsourcing to access inexpensive and scalable workforces, inevitably, the market will eventually saturate. Therefore it is imperative to develop ways to achieve efficiency. As an example, when compared to software engineering, CoQ of crowdsourcing is significantly high. For instance, it has been reported that the Motorola Global Software Group managed to decrease an initial 35% CoQ to 25% through software process improvement (Laporte, Berrhouma, Doucet, & Palza-Vargas, 2012). In this study we report CoQ ratings in a range of 68% to 226%. These tremendous ratings can also be decreased by developing ways to optimize quality costs. This study paves the way for future research aiming at quality and cost optimization for crowdsourcing and eventually some citizen science scenarios.

6. REFERENCES

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality Control in Crowdsourcing Systems: Issues and Directions. *Internet Computing, IEEE*, *17*(2), 76–81. doi:10.1109/MIC.2013.20

Amazon Mechanical Turk. (n.d.). Retrieved from http://www.mturk.com/

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Conte, S. D., Dunsmore, H. E., & Shen, V. Y. (1985). Software effort estimation and productivity. *Advances in Computers*, *24*, 1–60.

Crosby, P. B. (1979). *Quality is free: The art of making quality certain* (Vol. 94). McGraw-Hill New York.

CrossRef. (n.d.). Retrieved from www.crossref.org

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are Your Participants Gaming the System ? Screening Mechanical Turk Workers, 0–3.

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, *38*(2), 189–200.

Feigenbaum, A. V. (1956). Total Quality-Control. *Harvard Business Review*, *34*(6), 93–101.

Fishman, G. S. (1996). *Monte Carlo*. Springer.

Geiger, D., & Seedorf, S. (2011). Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. In … (pp. 1–11).

Grier, D. A. (2011). Foundational Issues in Human Computing and Crowdsourcing. In *Position Paper for the CHI 2011 Workshop on Crowdsourcing and Human Computation. CHI.*

Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, *57*(11-12), 2918–2932. doi:10.1016/j.mcm.2012.01.006

Ho, C.-J., & Vaughan, J. W. (2012). Online Task Assignment in Crowdsourcing Markets. In *AAAI*.

Hossfeld, T., Hirth, M., & Tran-Gia, P. (2011). Modeling of Crowdsourcing Platforms and Granularity of Work Organization in Future Internet. In *Proceedings of the 23rd International Teletraffic Congress* (pp. 142–149). International Teletraffic Congress. Retrieved from http://dl.acm.org/citation.cfm?id=2043468.2043491

Hsueh, M.-C., Tsai, T. K., & Iyer, R. K. (1997). Fault injection techniques and tools. *Computer*, *30*(4), 75–82.

Huang, E., Zhang, H., & Parkes, D. C. (n.d.). Toward Automatic Task Design : A Progress Report Categories and Subject Descriptors, 77–85.

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on Amazon Mechanical Turk. *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, 64. doi:10.1145/1837885.1837906

Karger, D. R., Oh, S., & Shah, D. (2011). Budget-optimal crowdsourcing using low-rank matrix approximations. *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 284–291. doi:10.1109/Allerton.2011.6120180

Kazai, G., Kamps, J., & Milic-Frayling, N. (2011). Worker types and personality traits in crowdsourcing relevance labels. *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, 1941. doi:10.1145/2063576.2063860

Kern, R., Thies, H., Bauer, C., & Satzger, G. (2010). Quality assurance for human-based electronic services: A decision matrix for choosing the right approach. In *Current Trends in Web Engineering* (pp. 421–424). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-16985-4_39

Kern, R., Zirpins, C., & Agarwal, S. (2009). Managing quality of human-based eservices. In *Service-Oriented Computing--ICSOC 2008 Workshops* (pp. 304–309).

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., … Horton, J. (2013). The future of crowd work. *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, 1301. doi:10.1145/2441776.2441923

Laporte, C. Y., Berrhouma, N., Doucet, M., & Palza-Vargas, E. (2012). Measuring the Cost of Software Quality of a Large Software Project at Bombardier Transportation.

Law, E., & Ahn, L. von. (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *5*(3), 1–121.

Le, J., & Edmonds, A. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *… crowdsourcing for search …* (pp. 17–20). Retrieved from http://ir.ischool.utexas.edu/cse2010/materials/leetal.pdf

Literally Canvas. (n.d.). Retrieved from http://literallycanvas.com/

Malcolm, D. G., Roseboom, J. H., Clark, C. E., & Fazar, W. (1959). Application of a technique for research and development program evaluation. *Operations research*, *7*(5), 646–669.

McCann, R., Shen, W., & Doan, A. (2008). Matching Schemas in Online Communities: A Web 2.0 Approach. *2008 IEEE 24th International Conference on Data Engineering*, 110–119. doi:10.1109/ICDE.2008.4497419

Okubo, Y., Kitasuka, T., & Aritsugi, M. (2013). A Preliminary Study of the Number of Votes under Majority Rule in Crowdsourcing. *Procedia Computer Science*, *22*, 537–543. doi:10.1016/j.procs.2013.09.133

Oleson, D., Sorokin, A., Laughlin, G., & Hester, V. (2011). Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human computation*, 43–48. Retrieved from http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewFile/3995/4267

Quinn, A., & Bederson, B. (2011). Human computation: a survey and taxonomy of a growing field. In *… Conference on Human Factors in Computing …*. Retrieved from http://dl.acm.org/citation.cfm?id=1979148

Quinn, A. J., & Bederson, B. B. (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1403–1412). New York, NY, USA: ACM. doi:10.1145/1978942.1979148

Rao, R. B., Fung, G., & Rosales, R. (2008). On the Dangers of Cross-Validation. An Experimental Evaluation. In *SDM* (pp. 588–596).

Ross, J., Irani, L., & Silberman, M. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended …* (pp. 2863–2872). Retrieved from http://dl.acm.org/citation.cfm?id=1753873

Rouse, A. (2010). A preliminary taxonomy of crowdsourcing. Retrieved from http://aisel.aisnet.org/acis2010/76/

Schenk, E., & Guittard, C. (2011). Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics & Management*, (1), 93–107.

Schiffauerova, A., & Thomson, V. (2006). A review of research on cost of quality models and best practices. *International Journal of Quality & Reliability Management*, *23*(6), 647–669. doi:10.1108/02656710610672470

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 614–622). New York, NY, USA: ACM. doi:10.1145/1401890.1401965

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on* (pp. 1–8). doi:10.1109/CVPRW.2008.4562953

Surowiecki, J. (2005). *The wisdom of crowds*. Random House LLC.

Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 57. doi:10.1145/1378704.1378719

Voyer, R., Nygaard, V., Fitzgerald, W., Copperman, H., Suite, B. S., & Francisco, S. (2010). A Hybrid Model for Annotating Named Entity Training Corpora, (July), 243–246.

Vukovic, M. (2009). Crowdsourcing for Enterprises. In *Congress on Services - I* (pp. 686–692). doi:10.1109/SERVICES-I.2009.56

Vukovic, M., & Bartolini, C. (2010). Towards a research agenda for enterprise crowdsourcing. In *Leveraging applications of formal methods, verification, and validation* (pp. 425–434). Springer.

Welinder, P., & Perona, P. (2010). Online crowdsourcing: Rating annotators and obtaining cost-effective labels. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 25–32. doi:10.1109/CVPRW.2010.5543189

Wieringa, R., & Morali, A. (2012). Technical action research as a validation method in information systems design science. In *Design Science Research in Information Systems. Advances in Theory and Practice* (pp. 220–238). Springer.

Wiggins, A., & Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on* (pp. 1–10).

Wikipedia. (n.d.). Retrieved from www.wikipedia.org