COST OF QUALITY FOR CROWDSOURCING MANAGEMENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

DENİZ İREN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

JUNE 2014

COST OF QUALITY FOR CROWDSOURCING MANAGEMENT

Submitted by **DENİZ İREN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Information Systems, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Informatics Institute**

Prof. Dr. Yasemin Yardımcı Çetin _____
Head of Department, **Information Systems**

Prof. Dr. Semih Bilgen _____
Supervisor, **Information Systems, METU**

Examining Committee Members:

Assoc. Prof. Dr. Altan Koçyiğit _____
Information Systems, METU

Prof. Dr. Semih Bilgen _____
Electrical and Electronics Engineering Dept., METU

Assist. Prof. Dr. Erhan Eren _____
Information Systems, METU

Assoc. Prof. Dr. Banu Günel _____
Information Systems, METU

Assist. Prof. Dr. Özgür Tanrıöver _____
Computer Engineering, Ankara University

Date: 18.06.2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Deniz İren

Signature : _____

ABSTRACT

COST OF QUALITY FOR CROWDSOURCING MANAGEMENT

İren, Deniz

Ph. D., Department of Information Systems

Supervisor: Prof. Dr. Semih Bilgen

June 2014, 104 pages

Abstract. Crowdsourcing is a business model which allows practitioners to access a rather cheap 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 way of doing work. Thus crowdsourcing practitioners use certain techniques to make sure the end product complies with the quality requirements. Each quality assurance technique used in crowdsourcing impacts the project cost and schedule. A well-defined method is needed to estimate these impacts in order to manage the crowdsourcing effectively and efficiently. This study proposes a cost of quality approach for analyzing quality related costs in crowdsourcing and introduces the cost models of common quality assurance techniques.

Keywords: Crowdsourcing; project management; cost of quality; cost models.

ÖZ

KİTLE KAYNAKLI ÇALIŞMA YÖNETİMİ İÇİN KALİTE MALİYETİ

İren, Deniz

Doktora, Bilişim Sistemleri

Tez Yöneticisi: Prof. Dr. Semih Bilgen

Haziran 2014, 104 sayfa

Öz. Kitle kaynaklı çalışma uygulayıcılar için görece ucuz ve ölçeklenebilir iş gücüne erişim sağlayan bir iş modelidir. Ancak gevşek işçi-işveren ilişkisi, kitlenin sahip olduğu yetenek çeşitliliği ve katılımcıların belirsizliği dolayısıyla geleneksel iş yapış biçimlerine göre daha düşük kalite ile sonuçlanma eğilimindedir. Bu nedenle kitle kaynaklı çalışma modelini kullananlar, erişmek istedikleri iş ürününün kalite gereksinimlerini karşıladığını garanti altına almak için belli kalite güvence yöntemleri kullanmaktadırlar. Kullanılan her bir kalite güvence yöntemi projenin maliyet ve takvimini etkilemektedir. Bu etkileri kestirerek kitle kaynaklı çalışma işini etkili ve verimli bir şekilde sürdürmek için iyi tanımlı yöntemler gereklidir. Bu çalışma, kitle kaynaklı çalışmalarda kalite ile ilişkili maliyetlerin analizi için kalite maliyeti yaklaşımı önermekte ve yaygın olarak kullanılan kalite güvence yöntemleri için maliyet modelleri sunmaktadır.

Anahtar Kelimeler: Kitle kaynaklı çalışma; proje yönetimi; kalite maliyeti; maliyet modelleri.

ACKNOWLEDGEMENTS

I dedicate this thesis to my family and friends. The time I spent on this dissertation was the time I did not spend with them. I am thankful for their understanding and patience. I hope they accept this apology and take solace in the fact that the time I stole from us was not spent in vain.

I would like to express my deepest gratitude to my supervisor Dr. Semih Bilgen. Without the wisdom he offered and the supervision he provided, I truly would be lost. He provided invaluable guidance, not only for this thesis but for my life. I will always strive to be worthy of his trust and appreciation.

I am grateful for all the supports of thesis committee members, Dr. Altan Koçyiğit and Dr. Banu Günel. Their insights significantly determined the direction of this study.

I also thank Dr. Erhan Eren and Dr. Özgür Tanrıöver for honoring me by accepting to be jury members of this thesis.

I am blessed by having Matthias Hirth as a colleague and friend as not only he inspired me in many aspects of my academic career but also supported me in various stages of my research. I also thank Dr. Tobias Hossfeld and Dr. Tran Gia for the opportunities they presented and guiding advices they offered. I sincerely wish that I may find the opportunity to work with them in the future.

I cannot ever be grateful enough for the support of the hero I always looked up to, the only constant in my life, the origin of my passion: my mother. When I reached to a point that I could not motivate myself anymore she was there to strengthen me and support me.

I thank my aunt, Dr. Zerrin İren, for being my lifelong source of inspiration. I think she has more influence on me than I can imagine for becoming the person I became.

I will surely fail to express the gratitude I feel for knowing Dr. Banu Aysolmaz, my mentor, my guide, my teammate, my best friend and my counterpart, as I know no words to represent my true emotions. She did not just witness all my troubled moments but shared them as well. I am truly blessed by her existence in my life for countless reasons. As she always challenged and inspired me in all aspects of my life, I know that I will be a better person because of her and for her.

I am also thankful for contributions of Gökhan Kul, Engin Bozkurt, Eylem Elif Maviş and my colleagues at the Computer Center for their support.

Throughout this process I always felt the support of my family, Gökçe Cüceoğlu, Doğa Cüceoğlu, Gülen Çalış, Özer Şendir, Set İren, Gaia İren, Odin İren and my friends, Denizhan Divanlıoğlu, Gerçek Dorman, Emrah Arslan, Ozan Çivit, Yener Balcı, Onur Süer, Altuğ Uçar, Volkan Rüzgar, Ulaş Kürüm, Doruk Balkan, Yusuf İncekara, Barış Salttürk, Murat Tüfekçioğlu, Dr. Oumout Chouseinoglou, Dinçer Özoran, Emre Sezgin, Tayfun Asker and Suna Yılmaz.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMT : Amazon Mechanical Turk

CG rating : Control Group Rating

CG voting : Control Group Voting

CoC : Cost of Conformance

CoQ : Cost of Quality

DF : Decision Fitness

DOI : Digital Object Identifier

EF : External Failure

FLIRT : Focus, Language, Incentive, Rules, Tools

FN : False Negative

FP : False Positive

GS rating : Gold Standard Rating

HIT : Human Intelligence Task

IC : Inconclusive

IF : Internal Failure

METU : Middle East Technical University

MMRE : Mean Magnitude of Relative Error

MRE : Magnitude of Relative Error

NA : Not Applicable

OCR : Optical Character Recognizer

PERT : Program Evaluation and Review Technique

TN : True Negative

TP : True Positive

CHAPTER 1

INTRODUCTION

Crowdsourcing has become a valid means of producing value in large- as well as small-scale projects. However due to certain characteristics of crowds such as loose employer – worker relationship, crowdsourcing continues to pose unique challenges for practitioners.

An overview of crowdsourcing and related concepts along with the context of this research are presented in this chapter. A brief definition of the problem addressed in this dissertation and an overview of the basic idea underlying the proposed solution are given together with the scope and research methods applied.

1.1. The Context

Crowdsourcing is an umbrella term for various value creation approaches with the shared characteristic of using a large group of people as resource (Howe, 2008). Crowdsourcing is first defined as "*the act of outsourcing a job which is traditionally done by designated agents, to a large group of people, in the form of an open call*" (Howe, 2006).

Despite the fact that the coining of the term "*crowdsourcing*" dates back to 2006 (Howe, 2006), examples of crowdsourcing have begun emerging almost at the same time with the founding of the Internet. However this new massive collaboration phenomenon has started to be broadly utilized as a business enabler following the rise of the Web 2.0. By introducing interactive features to the intertwined world of Internet, Web 2.0 empowered users with the ability to participate in content creation. Thus, this vast number of connected individuals, first became a global market, then transformed into a huge resource of human labor. Naming this phenomenon has triggered large numbers of research initiatives leading to development of novel, innovative and effective business models offering significant benefits.

Outsourcing is used to achieve higher quality at lower costs (Rouse, 2010). As a special form of outsourcing, crowdsourcing also offers low costs. In addition crowdsourcing

enables practitioners to access a scalable workforce consisting of individuals with a diverse skill set. Furthermore, costs are mostly associated with products, lifting the burden of organization activities, salaries, insurance and other overheads from the shoulders of practitioners. These characteristics make crowdsourcing a desirable business approach for risk prone, entrepreneurs.

Crowdsourcing is used for solving many problems which differ in size, type and importance. Large and complex tasks can be broken down into smaller, more manageable tasks and assigned to a crowd of workers as microtasks (Kittur, Smus, & Kraut, 2011). Contents can be organized to foster open innovation, collaborative problem solving and collective creativity even for solving world's most demanding scientific problems ("Innocentive," n.d.). These include content generation ("Youtube," n.d.), product development ("Threadless Inc.," n.d.), funding ("Sell-a-Band," n.d.), software alpha-beta tests ("Mob4hire," n.d.), marketing ("Amazon Inc.," n.d.), innovation ("My Starbucks Idea," n.d.), complex expert problems ("Innocentive," n.d.), image tagging (von Ahn & Dabbish, 2004), music tagging (von Ahn & Dabbish, 2008), establishing stock photography repositories ("iStockPhoto," n.d.) and massive data analysis ("Help-find-Jim," n.d.; A. J. Quinn & Bederson, 2011).

1.2. Problem Definition

In contrast to traditional business models, crowdsourcing lacks a clearly defined pact or a binding service level agreement between the workers and the employer. The crowd which participates in crowdsourcing usually consists of individuals with a certain degree of anonymity and who voluntarily choose the tasks to perform. These characteristics of crowdsourcing make the control of crowd-based production process uniquely challenging (Kittur et al., 2013). This lack of control raises concerns about the quality of the end product (Kern, Zirpins, & Agarwal, 2009). Thus, crowdsourcing researchers and practitioners have developed techniques to ensure that end products satisfy quality requirements. While some of these techniques are similar to traditional quality assurance techniques, others are completely different and cannot be practically applied in traditional production processes. For instance, having separate workers control the quality of the products created by a crowd is similar to having a dedicated assessment team performing quality checks in traditional production settings. On the other hand, a frequently used crowdsourcing quality assurance technique; *redundancy*, is not applicable to traditional production as it is impractical to manufacture, say, multiple cars, select the best and discard the rest. This inefficiency indicates the need to improve quality assurance processes in crowdsourcing. Such process improvement is possible either by developing better quality assurance techniques or by supporting decision making regarding which quality assurance technique to apply under certain circumstances.

Introducing and maintaining quality assurance techniques inevitably increase project costs. However the crowdsourcing literature lacks defined procedures for estimating

quality assurance costs. Such procedures may benefit crowdsourcing practitioners as guidelines for selecting and using quality assurance 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.

Hence the problem to be addressed in this dissertation is the formulation of a well-defined method to assess the level and cost of quality achievable in crowdsourcing projects. The method to be proposed will be validated in experimental as well as real-life settings.

1.3. Purpose of the Study

The purpose of this study is to contribute to the solution of the inefficiency of quality assurance techniques by introducing a method for estimating costs of common quality assurance techniques, which can be used as a guideline by crowdsourcing practitioners.

As a total quality management approach, Cost of Quality (CoQ) has been used in various domains frequently and successfully since 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 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 needs to be managed in order to optimize quality costs.

Crowdsourcing can still be considered as an emerging business enabler. Although crowdsourcing makes accessing a scalable workforce very easy, a crowd is a scarce resource and it is reasonable to expect shortcomings in the near future with increasing demands on globally interconnected crowds. In order to minimize inefficiencies regarding quality assurance, selection of quality assurance techniques should be based on defined practices and empirical results rather than solely on instincts.

In this study we introduce cost models for common quality assurance techniques used in crowdsourcing. These models are developed through CoQ approach, which emphasize the distinction between conformance and non-conformance costs. We also examine the effectiveness of quality assurance techniques to enable practitioners to conduct cost - effectiveness analysis, when used in combination with cost models.

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.

1.4. Research Strategy

In this study initially we performed a thorough review of the crowdsourcing literature. As we identified the gap in the literature regarding crowdsourcing management, we focused our research efforts on this area.

First we examined crowdsourcing taxonomies. Based on categorizations presented in various studies we clarified the definitions of crowdsourcing and related concepts. Then we constructed a representational anatomy of crowdsourcing which includes cost centers of generic crowdsourcing models. We identified cost of quality assurance, which displays significant differences in crowdsourcing settings, as an essential aspect of crowdsourcing, requiring specific approaches to be managed. Therefore we drilled through the literature and real life examples to sort out common quality assurance techniques used in crowdsourcing. The literature search revealed that the number of studies conducted specifically on crowdsourcing quality costs was not very high. Therefore we developed cost models and tailored these models to represent specific characteristics of common crowdsourcing quality assurance techniques.

To construct the cost models, we applied CoQ analysis and utilized observed process outcomes. We applied these models in a multiple action research which covered an experiment and various real-life crowdsourcing scenarios with different characteristics. In all action research cases we employed common quality assurance techniques. We logged crowd worker activities and compared the products against expert judgment. This comparison revealed the outcomes of individual quality assurance processes. By using the observed probability values of quality assurance outcomes we improved the cost models to represent quality costs accurately. All measurements were validated via v-fold cross validation techniques.

Finally we conducted semi-structured interviews with stakeholders of the action research projects. We evaluated and discussed the qualitative findings of those interviews along with the quantitative results of multiple action research.

1.5. Organization of the Thesis

This thesis is organized as follows. This chapter provides an introduction to the concept of crowdsourcing, describes the research problem and introduces the basic idea underlying the proposed solution. Chapter 2 sets the background and reviews the existing literature on crowdsourcing and quality assurance. Chapter 3 describes the research methodology used in this study. In Chapter 4 cost models associated with common crowdsourcing quality assurance techniques are introduced. Chapter 5 covers the multiple action research and presents the observations and findings. Finally in Chapter 6, the results are discussed and concluding thoughts are shared.

CHAPTER 2

LITERATURE REVIEW

The term crowdsourcing has been coined relatively recently, but the underlying concept of massive collaborative connected work dates back to the founding of the Internet. This chapter presents a review of the crowdsourcing literature. In Section 2.1 we present different, and at times conflicting definitions of crowdsourcing proposed by researchers. In Section 2.2 we review crowdsourcing taxonomies in the literature and propose a simplified taxonomy to support the present study. In Section 2.3 we introduce a component model which refers to the cost items associated with crowdsourcing process. In Section 2.4 we review the literature on management of crowdsourcing. In Section 2.5 we present the studies in the literature regarding crowdsourcing quality assurance and in Section 2.6 we present common crowdsourcing quality assurance techniques. In Section 2.7 we introduce the CoQ concept and in Section 2.8 we review existing cost models that fit the CoQ approach.

2.1. Crowdsourcing and Related Concepts

Many different definitions have been proposed for crowdsourcing and related concepts. Definitions suggested by researchers often conflict with one another. However, it is imperative to have a universal definition of crowdsourcing and clear distinction between related concepts in order to develop a robust and comprehensive management methodology.

In this research we sorted out various well-known crowdsourcing initiatives and studied a large number of research articles to elicit various definitions of crowdsourcing and related concepts. Since we identified conflicting definitions, we focused our effort on finding indicators of key characteristics of crowdsourcing which distinguish it from similar phenomena.

The first definition of crowdsourcing was proposed by Howe: "*Crowdsourcing is the act of taking a job traditionally performed by a designated agent and outsourcing it to an undefined, generally large group of people in the form of an open call*" (Howe, 2006).

This broad definition matches many Internet-based businesses and organizations such as Wikipedia or open source software development.

Brabham leaves out open source software development, and states that Wikipedia and open source software development are not crowdsourcing due to the inexistence of benefitting organizations in these examples. His definition of crowdsourcing is: "*an online, distributed problem-solving and production model already in use by for profit organizations such as Threadless, iStockphoto*". Brabham also emphasizes that Internet and Web 2.0 are the essential media for crowdsourcing. Brabham's definition suggests that crowdsourcing is always directed by an organization and only the sponsor organization benefits from the work performed (Brabham, 2008).

Rouse's preliminary taxonomy of crowdsourcing (Rouse, 2010) includes a dimension for categorization of crowdsourcing according to the *distribution of benefits*, clearly conflicting with Brabham's definition which states that only a sponsoring organization may benefit from the work performed.

We consider that the essence of crowdsourcing does not lie on the distinction of the benefitting parties. Rather, the value creation characteristic should be focused on. Doan, Ramakrishnan and Halevy's initial definition supports our viewpoint, while conflicting with Brabham's: "*crowdsourcing initiatives are the systems that enlist a crowd of users to explicitly (and implicitly) collaborate to build a long-lasting artifact that is beneficial to the whole community".* They consider this definition to be too strict, excluding many examples of crowdsourcing. Therefore, they state that "*a system is a crowdsourcing system if it enlists a crowd of humans to solve a problem defined by the system owners*" (Doan, Ramakrishnan, & Halevy, 2011).

The concept of "*creation of value by consumers*" is also emphasized in the concept of co-creation (Zwass, 2010). Brabham argues that contributors are mostly amateurs in crowdsourcing initiatives (Brabham, 2008) and Ipeirotis' findings support this statement as a large percentage of Amazon Mechanical Turk (AMT) ("Amazon Mechanical Turk," n.d.) participants allocate their free time as amateur hobbyists aiming to earn extra income (P. Ipeirotis, n.d.). This, however, excludes crowdsourcing examples such as AMT and Microworkers ("Microworkers," n.d.).

Crowdsourcing has also been used synonymously with the terms *human computation*, *collective intelligence* and *social computing/systems*. However nuances exist among these concepts:

According to Quinn and Bederson, the term human computation has been in use since 1838, in the field of philosophy and psychology literature (A. J. Quinn & Bederson, 2011). Contemporary definitions of human computation were provided by Law and von Ahn; human computation is "*simply computation that is carried out by humans*" or "*intelligent systems that organize humans to carry out process of computation*" (Law &

Ahn, 2011). Gentry et al. define distributed human computation to refer to the way of solving problems which are difficult for computers but easy for humans to solve (Gentry, Ramzan, & Stubblebine, 2005).

According to Quinn and Bederson, human computation must tackle the problems which fit the general computation paradigm, and thus can be solved by computers in the future. Additionally human computation requires human participation to be directed by computers (A. J. Quinn & Bederson, 2011).

Social computing / social systems are defined as "*systems which facilitate collective action and social interaction online with rich exchange of multimedia information and evolution of aggregate knowledge*" (Parameswaran & Whinston, 2007). Social computing has a broad coverage including almost any system which combines computing and social behavior of people.

Collective intelligence refers to a broad spectrum of phenomena of intelligent behavior among groups of individuals. Individuals do not even need to be human or living things.

According to Law and von Ahn, human computation must employ explicit control of computers on the process. This means that research focus in human computation is on algorithms instead of human behavior (Law & Ahn, 2011).

It is clear that crowdsourcing has a different meaning than these related concepts. In crowdsourcing a large number of individuals must be used as a resource. On the other hand, human computation does not necessarily require a crowd. Individuals participating in crowdsourcing may be isolated from each other so that social behavior of individuals has limited effect on the phenomenon. Collective intelligence is a broader term which entirely covers crowdsourcing.

In our efforts to reach a common definition of crowdsourcing we identified key characteristics of crowdsourcing and arrived at a new definition. We propose the following crowdsourcing definition which we base our research upon:

Crowdsourcing is a process of value creation by a generally anonymous mass consisting of voluntary non-professionals, as a result of an outsourcing initiative, in which the interactive features of the Internet are utilized.

2.2. Taxonomy of Crowdsourcing

In order to resolve conceptual conflicts, to clarify the definition of crowdsourcing and to draw a borderline which separates the related terms, researchers have developed taxonomies (Geiger & Seedorf, 2011; A. J. Quinn & Bederson, 2011; Rouse, 2010; Schenk & Guittard, 2011). These taxonomies provide a typology of existing crowdsourcing initiatives and their characteristics which are essential for researchers to build their studies upon. The present research also requires certain distinctions to be made. As an

example, quality assurance techniques vary in accuracy and effectiveness for different types of crowdsourcing tasks. Therefore identifying the type of task being used in a crowdsourcing setting is important to make a decision about which quality assurance technique suits best. Taxonomies provide this information to practitioners and researchers.

According to LaVechia, crowdsourcing initiatives are classified into three groups: *contest*, *marketplace* and *bid* (La Vecchia & Cisternino, 2010). Contest type of crowdsourcing is the type in which the work is announced to the crowd in the form of an open call. Submissions are evaluated either by experts or democratic voting of participants. For example, Threadless ("Threadless Inc.," n.d.) organizes a t-shirt design contest each week. At the end of the week designs which obtain the majority of votes are sent to production and designers are awarded with a royalty fee from all sold t-shirts. Similarly Innocentive ("Innocentive," n.d.) announces highly complex problems and offers significant amounts of prize money to anybody providing an acceptable solution. *Marketplace* crowdsourcing type operates by breaking down the work into microtasks and assigning them to crowd members. One by one microtasks are completed by participants and aggregated by the system into an end product. Analyzing satellite photos and transcribing voice recordings are examples of marketplace type crowdsourcing. *Bid* type crowdsourcing refers to the case when the work is announced in the form of an open call and participants submit their price and capability proposals.

Another classification derived from LaVechia's approach, similarly categorizes crowdsourcing initiatives as *marketplace* and *contest* (Vukovic, 2009). This categorization differs in defining *crowdsourcing mode* as an additional dimension. Vukovic's categorization also emphasizes the function within an enterprise where crowdsourcing initiative is undertaken. Major functions are listed as *innovation*, *design*, *development*, *test*, *sales*, *marketing* and *support*.

In order to support academic research, a more detailed taxonomy was proposed by Rouse (Rouse, 2010). Rouse's taxonomy examines crowdsourcing initiatives from three dimensions: *the nature of the work done*, *distribution of benefits* and *motivational tools* used.

Quinn and Bederson's taxonomy (A. J. Quinn & Bederson, 2011) referred earlier examines crowdsourcing initiatives in six dimensions: *motivation*, *quality control*, *aggregation*, *human skill need*, *procedural order*, *work demand cardinality*.

Geiger and Seedorf propose a taxonomy which differs from the prior taxonomy frameworks in that it focuses only on the organizational perspective (Geiger & Seedorf, 2011). The taxonomy they developed consists of four dimensions: *pre-selection of contributors*, *accessibility of peer contributions*, *aggregation of contributions* and *remuneration* for contributions.

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. Being objective means that the same result is produced each time the task is completed complying to its definition (Kern, Thies, Bauer, & Satzger, 2010a). For example, counting the number of road junctions on a satellite image of a town is an objective task. Each worker assigned with the same instance of this task reaches the exact same number, if s/he does the job successfully and in good faith. Results of objective tasks can be checked automatically. On the other hand 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. Even if the workers performing the same instance of this task are looking at the same image, they may reach different conclusions. The potential outcome of this task is binary, either positive or negative, thus, the frequency of the votes casted 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 few sentences is another example of subjective task, yet with an infinite set of potential results. In this case the results can only be aggregated manually. Table 1 provides further examples of task categorization which is essential to the present research.

Table 1 - Task categories and examples

Objective tasks	**Subjective tasks with finite potential result set**	**Subjective tasks with infinite (or very large) potential result set**
Transcribing an image of a distorted text	Judge an image's relevance to a text and map	Annotating a data object
Grouping similar items in a set of items	Answering a demographics survey	Tagging an image
Extract purchased items from a shopping receipt	Rating the traffic jam on a video stream of a road	Drawing an illustration of a cat
Finding duplicate items in a list	Choosing the best picture among a few pictures	Recommending a book or a movie related with given tags
Audio transcription of a news clip	Rating a product	Providing textual review about a product

Crowd type emphasizes the difference between an *internal* and an *external crowd*. Internal crowd consists of individuals who belong to the same organization such as a company or an association. Such 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, external crowd refers to

online individuals with a certain degree of anonymity. Thus, it is expected that both crowd types have different effects on many aspects of a crowdsourcing project such as quality, costs and motivational tools utilized in the project. Both crowd types have different uses. For instance, utilizing an internal crowd is more effective for a "wisdom of crowds" (Surowiecki, 2005) type crowdsourcing scenario because of the shared characteristics and the availability of common knowledge among the individuals of the internal crowd.

Wisdom of crowds can be used only when the required knowledge resides within the crowd. Levy explains the essence of wisdom of crowd solutions as: *"since 'no one knows everything, everyone knows something, [and] all knowledge resides in humanity', digitization and communication technologies must become central in this coordination of far-flung genius"* (Lévy & Bonomo, 1999) .

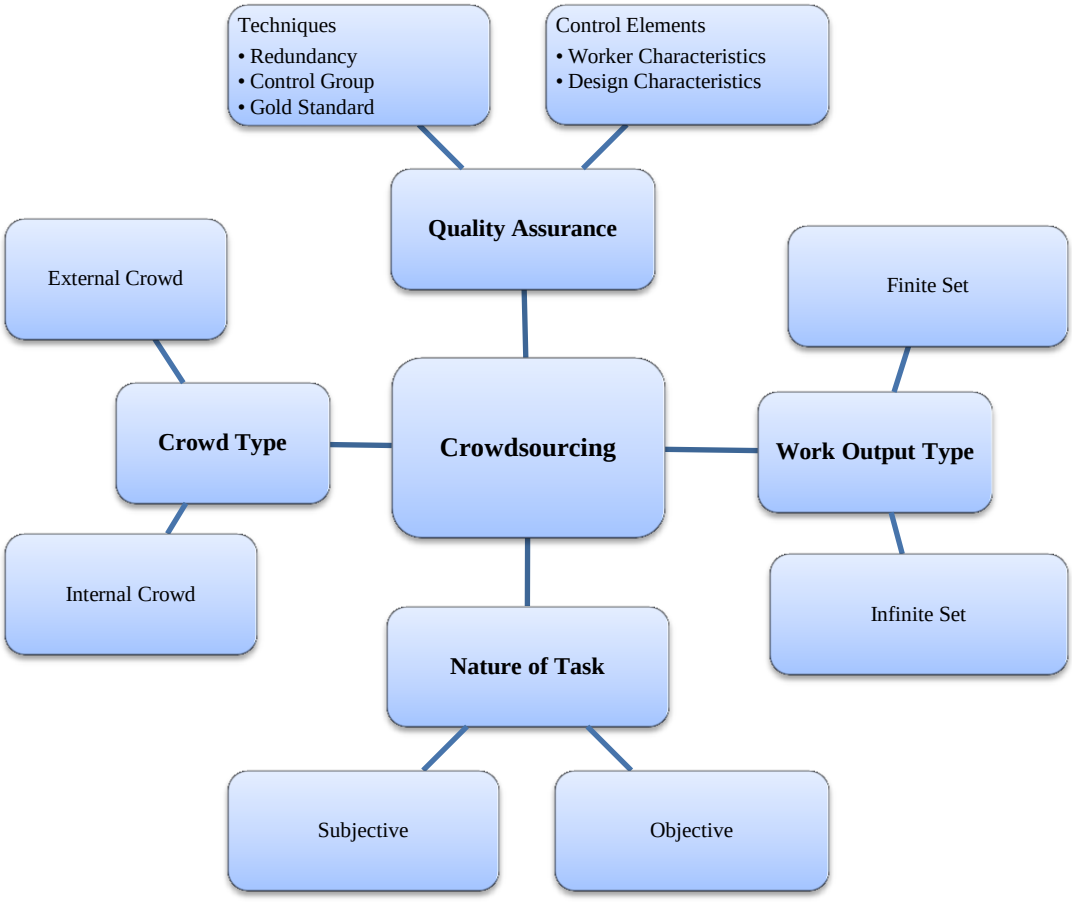


Figure 1 - Basic crowdsourcing taxonomy dimensions

Diversity and anonymity of a crowd have both advantages and disadvantages. Using crowdsourcing is effective especially in scientific or market research in which a random pool of participants is needed, as a crowd already satisfies this requirement (Kittur, Chi,

& Suh, 2008a). On the other hand, these crowd characteristics make the work performed rather difficult to govern.

2.3. Component Model of Crowdsourcing

Crowdsourcing process consists of distinct phases. Phases consist of activities and each activity involves certain costs. In order to comprehensively analyze cost items, a component model can be utilized. Component model is the abstract representation of activity components of a particular system. Component model for a generic crowdsourcing business model is used to gain a comprehensive understanding of the crowdsourcing concept and guiding activities and good practices in designing and operating a crowdsourcing initiative.

La Vecchia and Cisternino's formula representing crowdsourcing task decomposition and crowd resource allocation is considered an early example of crowdsourcing component models (La Vecchia & Cisternino, 2010):

T represents the task which can be split into multiple sub-tasks. t_j denotes sub-tasks that can be assigned to multiple workers. W_i represents an individual worker who is selected in the whole set of workers *W.*

A special purpose crowdsourcing system can be defined as $C_T(W)$ which assigns tasks to individual workers for a specific task T. On the other hand, a general purpose crowdsourcing system is defined as C(T,W) which has the ability to orchestrate the whole crowdsourcing process. There may be a set of constraints associated with the system, denoted by K, for example a microtask requiring special knowledge to be completed. Constraints are used in restrictions and permissions when associating tasks to workers.

Even though La Vecchia and Cisternino's formula is not comprehensive, it led the way for researchers who developed more advanced component models. Kittur, Smus and Kraut propose an approach to divide a complex work into more manageable and verifiable microtasks and assign the tasks to a crowd (Kittur, Smus, & Kraut, 2011). This approach also utilizes a similar component model consisting of the following components:

- Splitting project into microtasks,
- Assigning microtasks to workers,
- Training the workers,
- Timing the microtasks,
- Coordination of the microtasks,
- Inputting outputs of some microtasks to others,
- Quality assurance,
- Integrating microtask results.

2.4. Crowdsourcing Management

Studies aiming to enhance manageability of crowdsourcing projects exist in the literature. Researchers defending that open source software development is a form of crowdsourcing focus on application of management techniques used for open source projects on crowdsourcing initiatives. Jain examined governance techniques for major, successful open source software development projects and gathered her findings as an analysis framework for crowdsourcing initiatives (Jain, 2010). Viitamaki's FLIRT (Focus, Language, Incentives, Rules, Tools) model is another guideline which can be used to ensure manageability of crowdsourcing (Viitamäki, 2008). According to this model, a successful crowdsourcing initiative must have a clear focus, a common language used by the crowd, proper incentive mechanisms, rules defining the terms of participation and tools to make participation easier. Hirth proposes two distinct cost models for two broad categories of cheat detection mechanisms (Hirth, Hoßfeld, & Tran-Gia, 2011). Majority decision type cheat detection mechanisms use statistical consistency for quality assurance, while control group type cheat detection mechanisms rely on comparing contributions to trusted inputs. Both the approach and the results of Hirth's studies have provided significant contributions to the measurement and management of crowdsourcing.

2.5. Crowdsourcing Quality Assurance

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 of crowdsourcing (Grier, 2011), minimizing costs (Vukovic & Bartolini, 2010) while improving quality to a level of perfection (Kittur et al., 2013) still need to be satisfactorily addressed.

Due to anonymity and limited accountability of workers and lack of control over crowds, quality assurance is an essential part of crowdsourcing. All crowdsourcing initiatives involve ways to detect or prevent poor quality contributions and most publications about crowdsourcing consider the application of some quality assurance technique.

Since tasks are performed by a crowd, the quality of results is directly influenced by crowd characteristics. In order to set realistic quality goals it is imperative to know these 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 (J. Downs & Holbrook, 2010), and other worker characteristics (Kazai, Kamps, & Milic-Frayling, 2011). When crowd characteristics are known, practitioners can foresee the rate of poor quality task results provided by the crowd, and thus, they can decide on the extent of investment required on quality assurance.

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 problem causes can be handled with different approaches. 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 (J. S. Downs, Holbrook, Sheng, & Cranor, 2010). On the other hand, identifying cheaters and removing them from the crowd requires tighter quality assurance techniques.

It is also important to know what motivates the crowd to participate and to complete the tasks in good faith. Crowdsourcing platforms such as Microworkers or AMT are businesses based on participants' completion of microtasks in exchange of a small payment. Research suggests that raising the payment increases the quantity of the work (Horton, Chilton, Paul, & Way, n.d.; Sorokin & Forsyth, 2008) and increases completion speed (Mason & Watts, 2010). However, it has frequently been observed that payment does not improve the quality of work performed by a crowd (Mason & Watts, 2010; Rogstadius, Kostakos, Kittur, & Smus, 2011). Thus, increasing payments to workers should not be expected to lead to high quality nor quality assurance cost savings.

Shaw et al. studied the effects of both financial and social incentive schemes on quality (Shaw, Hall, Horton, & Chen, 2011). Their findings suggest that while some financial incentives have positive effect on quality, none of the social incentives lead to increased quality. It has also been reported that better quality is achieved when intrinsic motivators are used instead of extrinsic motivators (Rogstadius et al., 2011). Thus, quality expectations are related with the motivators used. Practitioners have to consider which type of motivator to use when designing crowdsourcing tasks.

In summary cost factors of quality consist of crowds' characteristics, task design, motivators utilized and the rate of existence of the cheaters in the crowd.

2.6. Common Crowdsourcing 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 and selective worker assignment. Run-time quality assurance covers various techniques which can be applied while the task is being performed, often requiring additional actions from workers or practitioners. 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 2 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 design-time characteristics as independent variables of research, and we propose CoQ models for common run-time quality assurance techniques.

Table 2 - Common quality assurance techniques used in crowdsourcing

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

2.6.1. Redundancy

Quality assurance techniques which involve assigning multiple instances of the same task to contributors in order to produce interchangeable results are classified as *redundancy*.

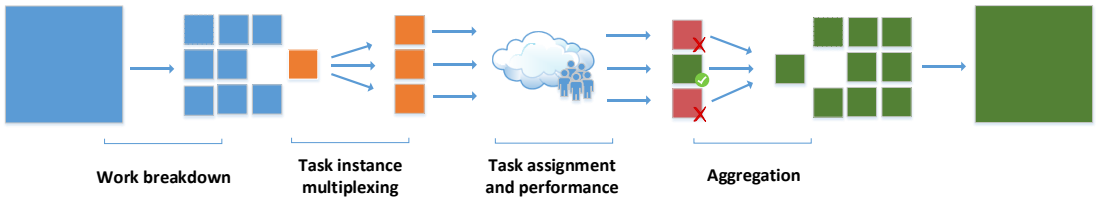


Figure 2: Redundancy quality assurance process.

To achieve quality assurance via *redundancy* (Figure 2), 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.

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 tasks are *objective* or *subjective with a finite potential result set*. This way results can be compared and the frequency of each submission can be determined automatically, and the most frequent submission can be accepted as the best result. Manual aggregation can be performed by a different set of workers or domain experts. This is basically utilization of *control group.*

The names given to *redundancy* techniques of quality assurance vary according to the business domain of practitioners, the aggregation mechanism utilized in the technique or various operational nuances.

For instance, Sorokin et al. use the term *multiple annotations* (Sorokin & Forsyth, 2008) whereas Sheng et al. use *repeated labeling* (Sheng et al., 2008) to express a scenario of collecting multiple labels/annotations from a crowd in the knowledge discovery domain.

Emphasizing the aggregation approach, the terms *majority voting* (Eagle, 2009) and *majority decision* (Hirth et al., 2013) are the most common ones which refer to *redundancy*.

Some *redundancy* techniques seek agreement of multiple contributors synchronously. For instance, *output agreement* which is used in ESP Game (von Ahn & Dabbish, 2004) requires two players to submit identical labels for the same image synchronously. On the other hand, a similar technique, *input agreement* used in Tag-a-Tune (von Ahn & Dabbish, 2008), uses an asynchronous scheme, in which submissions are evaluated after task completion.

All *redundancy* techniques have the common characteristic that a number of instances of the same task are assigned to multiple contributors. The number of redundant submissions varies due to many factors such as quality requirements, cost considerations, crowd characteristics or domain constraints. Sheng et al. show that increasing redundancy level is only beneficial if the probability of correctness of individual submissions (p) is greater than 0.5. The level of benefit of adding more contributors changes according to this value (Sheng et al., 2008).

Zhai et al. use an iterative approach to assign weights to user votes when deciding on crowd consensus. Certain workers have more influence on the consensus based on their former accuracy (Zhai, Hachen, Kijewski-Correa, Shen, & Madey, 2012). Using a weighted voting scheme may result in decreasing the votes needed therefore decreasing the costs.

Redundancy, by design, can lead to decreased resource efficiency to a great extent. Thus, using cost models when designing crowdsourcing tasks is vital for optimizing resource utilization.

2.6.2. Control Group

In *control group* techniques, submissions of the main group of workers are controlled by a separate group (Figure 3). 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 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) accepting or rejecting work items (tasks, products, etc.). Rating is defined as *classification or ranking something based on a comparative assessment* ("Oxford Dictionary: definition of rating," n.d.). Rating can be done at an ordinal scale where the notion of ordering is meaningful.

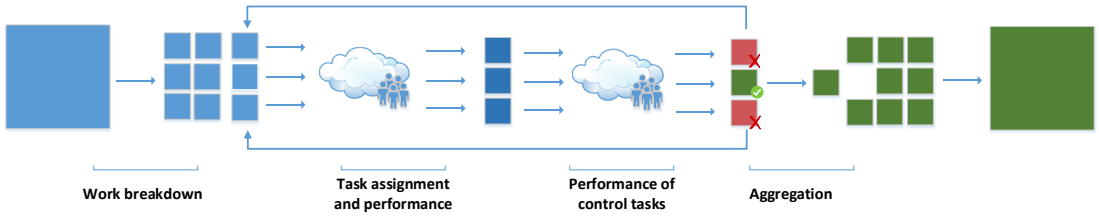


Figure 3 - Control group quality assurance process

When the controlling party consists of more than one individual, the controlling group needs to reach a consensus. By design, these cases pose redundancy, and the same mechanisms of aggregation apply in control tasks.

Generally controlling the outputs of a task is far less complex than performing that task. In those cases, control tasks may cost significantly less than the main task (Kern et al., 2010a). However when the primary task is extremely simple and small, time and cost spent on verifying the task outputs become comparable with the resources used for the primary task (P. G. Ipeirotis et al., 2010). Hirth et al. show that using *Control group* techniques is more cost effective when the primary task is significantly more complex than the control task (Hirth et al., 2013).

A *control group* may not only be responsible for accepting or rejecting submissions but also providing feedback, rationale for the decision made or improving submissions (Kern et al., 2010a). Obviously, these additional efforts result in increased costs.

Voting and rating can be applied in reviewing outputs of both simple and complex tasks. Kittur et al. exemplify the usage of voting mechanism as a way of evaluating the quality of multiple Wikipedia articles which are similar in content (Kittur, Smus, Khamkar, & Kraut, 2011). Most of the contest-type crowdsourcing initiatives use voting

and rating to select the best submissions. For instance, Threadless, a popular crowdsourcing initiative which focuses on t-shirt design, uses rating mechanism for selecting designs to be produced ("Threadless," n.d.). Rating is used in almost all online marketplaces which utilize a recommender system.

2.6.3. 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 (P. G. Ipeirotis et al., 2010). This has an improving effect on submission quality, whether the gold standard comparison is made for training users before moving on to real tasks (Le & Edmonds, 2010), or randomly carried out within the task performing process. 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).

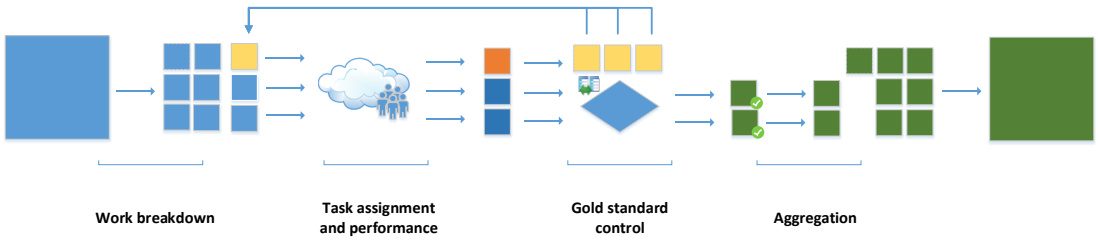


Figure 4 - Gold standard quality assurance process

Gold standard verification can be applied at different stages in a crowdsourcing process. The most frequent usage is asynchronous, in which gold standard tasks are assigned to workers randomly in the task sequence. McCann et al. define a mechanism for identifying trusted and untrusted workers via gold-standard questions (McCann et al., 2008). In synchronous usage, the main task and the gold standard task are assigned at the same time (Figure 4). As an example, Re-Captcha presents images of two words together to a user. One of the images displays a control word which is known in advance. If that word is submitted correctly by the user, only then is the submission for the unknown word considered valid. The second word is the one which is expected to be digitized (von Ahn & Dabbish, 2008). Gold standard tasks can be assigned to workers before the main tasks as a method for training them or evaluating their competency.

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

2.6.4. Worker Characteristics

Since most of the poor quality work comes from a small percentage of workers (Kittur, Chi, & Suh, 2008b; A. Quinn & Bederson, 2011), by identifying and removing this small portion from the system, overall quality can be increased.

Quality assurance techniques based on worker characteristics can only be used in cases which workers do not have total anonymity. Researchers have focused their attention on a large spectrum of areas to develop ways to improve crowdsourcing quality by studying workers. These areas include but are not limited to crowd demographics (P. Ipeirotis, n.d.; Ross et al., 2010), participation inequality(Stewart, Lubensky, & Huerta, 2010), contributor biases (Antin, 2012), worker character stereotypes (Kazai et al., 2011) and motivation (Rogstadius et al., 2011; Shaw et al., 2011).

Reputation which is a measure of worker trustworthiness is calculated according to former submissions by an individual. Reputation can be used as a criterion for selecting crowd members or identifying and banning cheaters. Establishing reputation tracking infrastructure requires workers to be identified by the system. Crowdsourcing platforms such as Microworkers and AMT keep worker accounts. Wikipedia uses a reputation system to choose reputable workers as reviewers and editors (Stvilia & Twidale, 2008).

Ad hoc, temporary reputation systems may also be developed. Callison-Burch used a few initial -gold standard- questions to judge if workers are trustworthy or not. Workers were assigned trust scores according to the extent to which their answers matched expert answers (Callison-Burch, 2009).

Recently an increasing number of researchers have started working on quality assurance techniques based on worker characteristics. These studies aim at managing worker skills, biases and trustworthiness to select the most appropriate workers for specific tasks (Ho & Vaughan, 2012). Furthermore, workers' social media profiles and networks are used for worker recommendation (Difallah, Demartini, & Cudré-Mauroux, 2013).

2.6.5. Design Characteristics

Quality assurance can be achieved through designing user friendly and robust tasks. Defensive task design in crowdsourcing suggests designing tasks so that cheating is not easier than completing the task in good faith (Kittur et al., 2008b; A. Quinn & Bederson,

2011). It is also recommended to include verifiable parts in tasks (Kittur et al., 2008b), so that statistical quality control becomes possible.

Through a statistical approach, Ipeirotis emphasizes the distinction between a predictable error (bias) and unrecoverable error (spam submission). Based on an algorithm using a confusion matrix and soft labeling technique they are able to calculate the error rate and expected cost of a contribution of a particular worker. Identifying bias patterns make recovery possible, thus decreases costs of making non-true contributions (P. G. Ipeirotis et al., 2010).

In certain situations where a time consuming task such as reading a long text or hand-drawing an item, monitoring the time-to-complete the task may be a valid way to detect cheaters. In an experiment involving reading and grading Wikipedia articles, Kittur et al. used time-to-complete measurements to differentiate participants who are cheating, during post analysis of the submissions (Kittur et al., 2008b). Tasks can be designed to last no less than a certain amount of time, and submissions made faster can either be flagged for further quality control or denied automatically (Xia, Zhang, Xie, & Li, 2012).

Task size can make a difference in the quality of worker contributions (Hossfeld et al., 2011). Thus, optimal granularity level should be achieved by dividing complex tasks into smaller, simpler, shorter microtasks.

Aside from these, crowdsourcing practitioners have developed good practices and guidelines for effective task design. Tasks must be clearly described. User interfaces must be simple and user friendly. In paid microtask crowdsourcing projects, payments must be fair. Workers tend to choose to work on tasks which they are able to perform multiple times, to maximize their gains. Enabling workers to complete tasks over and over again can result in faster task completion but also can attract cheaters.

A submission which is aligned with the majority decision may not always be of high quality. Thus, denying payment for tasks which do not reflect majority decision may bias the crowd behavior. Being aware of such a payment scheme, participants may choose to make contributions which they think would align with other submissions, rather than what they think is correct.

2.6.6. Combining Multiple Quality Assurance Techniques

Using multiple quality assurance techniques is a common practice, especially when high quality is desired. However this may result in significant cost increases. Thus collective usage of quality assurance techniques should be optimized according to quality needs.

McCann et al. describe a series of quality assurance practices used together in an experiment. Acknowledging the fact that untrustworthy contributors exist in the crowd, first they try to select trusted users by asking them evaluation questions. If a

user provides a sufficient number of valid responses, they classify that user as *trusted*. They collect the answers submitted to other questions which are asked to multiple users. They select the answer which was submitted most frequently by contributors (McCann et al., 2008). In this example, using of the evaluation questions which have answers known in advance is basically a *gold standard* quality assurance technique. Establishing reputation scores for workers is *worker characteristics* type of quality assurance. Asking multiple instances of the same question to multiple people is *redundancy*.

2.7. 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. It is expressed as the sum of 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, workforce remains a scarce resource. Deciding to spend effort for quality assurance purposes rather than performing new tasks introduces an opportunity cost. Especially in enterprise crowdsourcing (Vukovic, 2009), significant hidden costs are incurred since crowds consist 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 percentage of the CoQ in an overall crowdsourcing job is generally higher compared to the traditional production process. Major CoQ categories and example crowdsourcing scenarios are listed in Table 3.

Table 3: 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

It is expected that various quality assurance techniques lead to different ratios of cost of conformance and cost of non-conformance. Since non-conformance may result in lost reputation and profit to an unknown extent, it is considered as the more risky portion, thus, practitioners initially often tend to minimize non-conformances. Utilization of additional quality assurance techniques cause the cost of non-conformance to decrease, while expectedly increasing the costs of conformance (Figure 5). Therefore, in order to optimize quality costs, analyzing conformance and non-conformance costs is imperative.

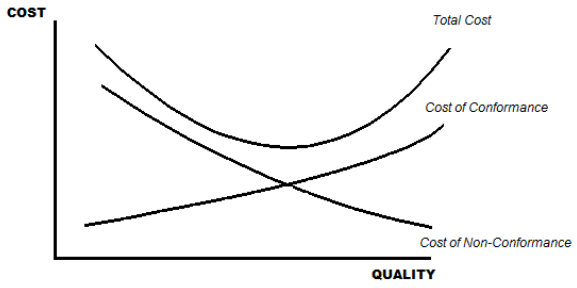


Figure 5: Relationship between cost of conformance and cost of non-conformance

2.8. Cost Modeling of Crowdsourcing

Cost modeling of crowdsourcing has rarely been studied in the literature. Hirth developed probabilistic cost models of control group and majority decision techniques and compared them, reporting that control group techniques yield better results when the primary task is more complex than the control task (Hirth et al., 2013). Kittur et al. analyzed crowdsourcing process aiming at identifying cost components. They mainly focus on coordination of crowdsourcing tasks at different levels (Kittur, Smus, & Kraut, 2011).

Hirth et al. examine *control group* and *majority decision* techniques in terms of cost effectiveness. By using probabilistic cost models, they show that both techniques offer the same cheat detection effectiveness but different costs and applicability based on simulation. While *control group* technique is more cost-effective for more complex and expensive tasks, *majority decision* is more cost-effective for simple and cheap tasks (Hirth et al., 2013).

The present study aims to extend and generalize Hirth et al.'s idea of using probabilistic cost models by introducing a CoQ approach. To develop cost models, we have used quality assurance process outcomes empirically observed in real life scenarios. The present study also differs from that by Hirth et al. in terms of the goal which can be stated as providing an adjustable, easy to use approach for crowdsourcing practitioners to be utilized in various crowdsourcing settings.

CHAPTER 3

METHODOLOGY

In this chapter we describe the scientific method which we applied in this study. Crowdsourcing, as a research area, is not currently well defined. Therefore we required basic definitions of the concepts and related taxonomy dimensions in order to base our research upon. First we conducted a comprehensive review of crowdsourcing literature. We examined taxonomy studies thoroughly and identified the factors potentially having the largest effect on quality and costs. Then we classified common crowdsourcing quality assurance techniques and derived cost models representing the characteristics of those techniques. Finally we applied these cost models in multiple action research cases. We primarily applied quantitative techniques, supported by qualitative methods throughout the cases.

The methodology used in this thesis research is presented in detail in this chapter. Section 3.1 states the research questions. Section 3.2 explains our epistemological stance. Our pluralist approach which combines qualitative and quantitative data collection and analysis is described in Section 3.3 and finally the details about the multiple action research method are provided in Section 3.4.

3.1. Research Questions

Application of quality assurance techniques impacts crowdsourcing costs. However, the crowdsourcing literature lacks a well-defined method to estimate the cost of quality. As we identified this problem, we directed our research efforts to analyze the problem, ask research questions and build propositions that may lead us to the solution of our research problem.

(Q1) How can the costs of quality assurance techniques be estimated?

As stated in Section 2.6, many quality assurance techniques exist which are commonly used in crowdsourcing. The proposed estimation method needs to be comprehensive, covering most quality assurance techniques and must be applicable to various crowdsourcing settings.

Therefore we propose that by defining cost models of common quality assurance techniques cost of quality assurance can be estimated accurately.

(Q2) Can cost of quality models be used to support decision making of practitioners regarding method selection and assist them to avoid inefficiencies?

Since the crowdsourcing literature lacks methods for estimating cost of quality, practitioners apply quality assurance techniques on crowdsourcing systems based on their instincts or experiences. This usually leads to inefficiencies such as overusing or misusing quality assurance techniques. These inefficiencies cause excessive increase in quality costs which leads to a massive loss of workforce at a global scale. We propose that cost models of quality assurance techniques can be used to accurately determine cost-effectiveness. Thus, by enabling practitioners to make decisions based on estimation techniques, cost models can be utilized for selecting efficient quality assurance techniques. Therefore widespread utilization of quality cost estimation methods may lead to significant cost savings for the entire crowdsourcing domain.

3.2. Epistemological Stance

In the course of this study we have held an interpretivist epistemological stance. The literature includes examples of case research conducted from both interpretivist and positivist viewpoints (Cavaye, 1996; Walsham, 1995). This study focuses on crowdsourcing which involves participation of a large number of workers. Our observations highly depend on the behavior of people. The proposed cost models employ parameters which reflect the submission quality profile of a large number of workers. We examine the social dynamics of participation and incentive mechanisms to support motivation of the crowd. Similarly, psychology and sociology study highly context-dependent, complex and hard to predict phenomena, generally from an interpretivist viewpoint. Therefore, a positivist stance which claims that scientific knowledge is unique, objective, universal and can only be gained by repeatable experiments and observations is not suitable for this research. This research differs from observations of natural phenomena because it is related with man-made, social systems and includes examinations of how this system is perceived by individuals and groups. For this reason, an interpretivist rather than positivist stance has been preferred one.

3.3. Pluralist Method (Quantitative and Qualitative)

By definition, crowdsourcing consist of many tasks performed by a large number of individuals via computational technologies. This empowers crowdsourcing researchers with the ability to run statistical techniques and derive generalizable conclusions. However crowdsourcing is also a social discipline since it is strongly related with social dynamics of groups of people working together and psychology of individual contributors. Thus, a comprehensive research carried out in crowdsourcing domain should apply both quantitative and qualitative methods.

Usually data collected in case studies and action research possess qualitative nature. Interviews, field observations and examination of written documents might enable access to richer and more relevant data. However, collecting and analyzing qualitative data is both difficult and time consuming. Moreover it is usually too difficult and inappropriate to make generalizations out of qualitative findings.

Supporting the usage of qualitative data with quantitative data is a widely recognized way of enhancing research validity and overcoming the problems caused by the shortcomings of both methods. In this research we adopt a pluralist approach which suggests the usage of both qualitative and quantitative methods together. Quantitative analysis was used when measuring work submissions of crowd workers and when seeking statistical consistency in model parameters and outcomes.

3.4. Multiple Action Research

In this study we use action research methodology (Figure 6).

A famous Chinese proverb says: "*tell me, I'll forget, show me, I'll remember, involve me, I'll understand*". Action research, which was originally promoted by Lewin (Lewin, 1946), is an approach which takes a pragmatic stance; suggesting that knowledge can be gained through experience. Therefore, this approach aligns well with our perspective in this study.

In case studies the researcher is not expected to be involved in solution development. The researcher's role in case studies is limited to being the external observer and analyst (Susman & Evered, 1978). However in this study, the researcher was not only an observer but also a worker collaborating with the personnel of the host organization in all three cases. As the action researcher is a participant in the work and the process of change is the main focus of research (Benbasat, Goldstein, & Mead, 1987) the main research method used in this study is action research rather than case study (Wieringa & Morali, 2012). Both the researcher and host organization agreed with the scope, roles, responsibilities and interests prior to the study. The action research lead to an immediate change in organization and the results of two action research cases (Section 5.2 and Section 5.3) were put to use.

Action research has often been criticized for not having a defined process (Avison, Lau, Myers, & Nielsen, 1999). As advocated by Susman and Evered, this study was carried out in a methodological fashion, with defined process steps (Susman & Evered, 1978). In order to address this issue we applied the canonical action research methodology (Davison, Martinsons, & Kock, 2004).

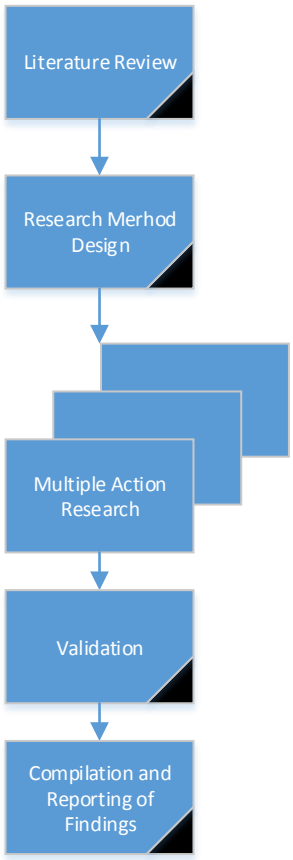


Figure 6 – Multiple action research methodology process

We finalized the multiple action research by conducting semi-structured interviews with project stakeholders who had the opportunity to observe the crowdsourcing process. We referred to the evaluations of the stakeholders for validating the research goal, the effectiveness of proposed solution and the results.

CHAPTER 4

COST OF QUALITY MODELS IN CROWDSOURCING

In this chapter we propose cost models for common crowdsourcing quality assurance techniques.

First in Section 4.1 we describe a generic quality assurance technique and potential outcomes of quality assurance processes. In Section 4.2 we formulate cost models of redundancy, control group and gold standard quality assurance techniques. Finally in Section 4.3 we propose a cost of quality estimation process which utilizes the proposed cost models. A more detailed description of common crowdsourcing quality assurance techniques and associated CoQ models can be found in (Iren & Bilgen, 2013a).

4.1. Anatomy and Outcomes of a Generic Quality Assurance Technique

The goal of any quality assurance technique is either to prevent or to detect low quality.

There is a finite set of potential outcomes of quality assurance process. Quality assurance techniques either correctly or incorrectly evaluate the quality of products, when in fact the product is either of good or poor quality. In certain special cases quality assurance techniques may also fail to reach a conclusion.

When the quality of the product is acceptable and the quality assurance technique correctly approves it, the outcome is True Positive (TP). When the quality assurance technique identifies poor quality correctly the outcome is True Negative (TN). However, when a good quality product is incorrectly detected as defected, the outcome is False Negative (FN). Likewise, when the quality assurance technique fails to detect a defect, and incorrectly approves a poor quality submission the outcome is False Positive (FP). In cases when the quality assurance technique fails to reach a decision about the submission, the outcome is Inconsistent (IC). The probabilities of these outcomes shall be denoted as P_{TP}, P_{TN}, P_{FN}, P_{FP} and P_{IC}, respectively

These outcomes are shown in Figure 7. Each outcome leads to different types of costs, which have a specific impact on the project. Thus this differentiation is essential for analyzing the CoQ. The outcomes differ for various quality assurance techniques.

Therefore, cost models of different quality assurance techniques also differ. 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 prevented or detected by the quality assurance technique, 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}.

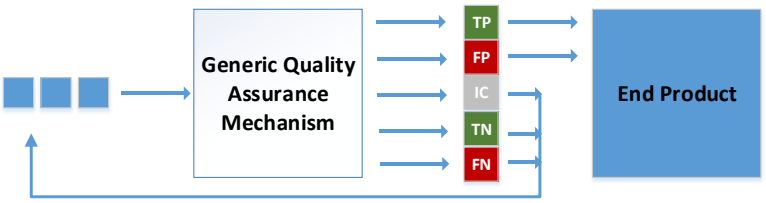


Figure 7 - Possible outcomes of a generic quality assurance process

4.2. Cost Models for Common Quality Assurance Techniques

Conformance costs (CoC) vary depending on quality assurance process design. These do not include direct costs. Direct cost is the cost of one task, without any quality assurance technique. Thus total direct cost is the total cost of the job only when assuming that all tasks are performed in perfect quality and no measures are needed for quality assurance. 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:

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

In order to achieve a complete end product, it is assumed that all outputs which fail to comply with 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 represent with monetary costs. In this study 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. For instance, a poor quality translation of a sentence may not cause a major trouble for a website operator, whereas a failure to flag a potential tumor on a computed tomography scan image may lead to a dreadful result.

Furthermore, when quality assurance techniques fail to distinguish between poor and good quality, long term problems may arise regarding trust mechanisms and crowd behavior. If workers' good quality submissions are being frequently denied by quality assurance techniques, workers 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} is a global variable and currently there is no way to estimate or control this type of damage and its long lasting, large spectrum 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 4 shows the outcomes of a generic quality assurance process and different categories of non-conformance raised by those outcomes.

Table 4 - 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 formulated 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 quality assurance techniques applicable in a given crowdsourcing scenario need to be identified.

Cost models are derived by multiplying the probability of an outcome with its estimated impact. Thus, certain probability values of outcomes need to be known in advance. These values can either be obtained from similar empirical experiments such as the ones covered in this study or a pilot project can be initiated and probability values can be observed. Then, these values have to be used as parameters of respective cost models.

4.2.1. Redundancy

The redundancy quality assurance process can produce three possible outcomes. Redundancy does not explicitly deny an output but rather assumes selecting the output with better perceived quality. Thus, the output is placed among the end product whether it fits the quality criteria or not. The only exception is the inconclusive outcome.

Figure 8 shows the potential outcomes of redundancy quality assurance mechanisms. When an output is selected among a few other outputs produced by different instances of the same microtask, it is assumed to be of high quality. The probability of redundancy quality assurance process selecting the output with truly high quality is P_{TP}. P_{FP} is the probability of the quality assurance mechanism to fail to filter out poor quality output and potentially erroneous output is placed among the end product. With

the probability of P_{IC} the redundancy quality assurance mechanism fails to achieve a conclusion about the quality of the submission. Inconclusive outcome can occur when none of the outputs of different instances of the same microtask can be selected. For example, if the number of redundant instances (m) is even, and the votes are in balance then a consensus cannot be reached.

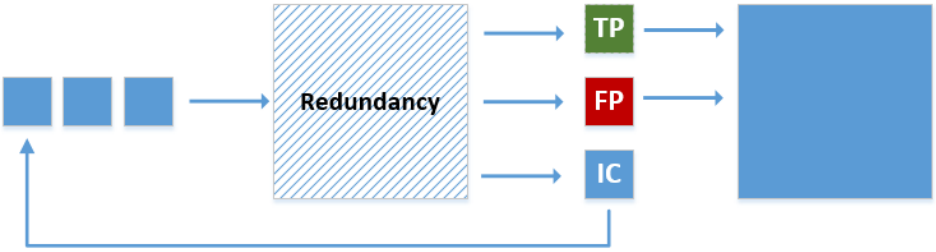


Figure 8 - Possible outcomes of redundancy quality assurance mechanisms

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}:

$$CoC_{Red} = N \cdot ((m-1) \cdot C_0 + C_{agg}) \qquad \textbf{(2)}$$

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

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

The probability of an EF is P_{FP}. EF leads to potential error in the end product (C_{err}) and damage done to the reputation and trust mechanisms and the worker community (C_{dmg}):

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

4.2.2. Control Group

Control group quality assurance process has four possible outcomes. Figure 9 shows these outcomes and the control group quality assurance process. P_{TP} is the probability of the worker submitting a high quality output and the control group correctly decides that it is valid. P_{TN} is the probability of the worker making a poor quality submission and the control group correctly decides that it is invalid. P_{FP} is the probability of control

group accepting a poor quality contribution and P_{FN} is the probability of control group to deny a good quality contribution by mistake.

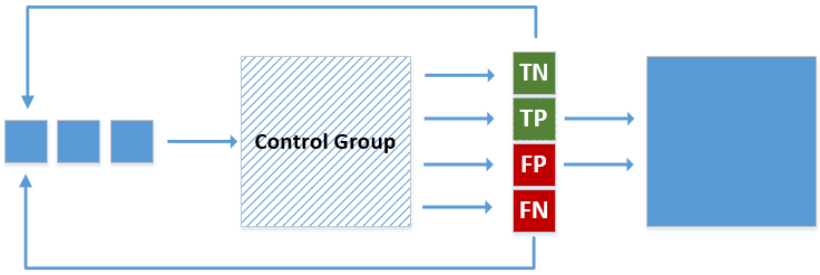


Figure 9 - Possible outcomes of control group quality assurance mechanisms

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

The conformance costs in *control group* (CoC_{CG}) techniques are caused by the control tasks as shown in (5). Generally controlling outputs of a microtask is significantly less complex and thus costs less.

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

When the controlling workers decide that the submission does not comply with quality criteria, the output of the task is denied 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 denied as $P_{FN} + P_{TN}$. The cost of rework and retest is $C_0 + C_1$:

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

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

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

4.2.3. Gold Standard

Gold standard techniques can be used asynchronously or synchronously. In asynchronous usage gold standard tasks are assigned to the workers separate from regular tasks, usually in form of qualification or training. In synchronous usage, gold standard tasks are assigned together with a number of regular tasks. In this study we examine the costs of synchronous usage of *gold standard* techniques.

In synchronous usage, gold standard tasks are provided to the user along with a number of regular tasks. In this case the decision to approve or deny the submissions is based on the comparison of the gold standard output and the predefined expected result. If the gold standard output is valid then the entire group of task outputs is accepted. Possible outcomes of a gold standard quality assurance mechanism are shown in Figure 10.

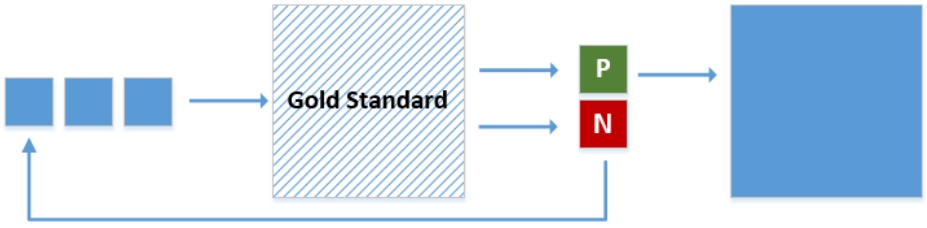


Figure 10 - Possible outcomes of gold standard quality assurance mechanisms

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}.

(8) shows 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 denotes the number of gold standard tasks assigned to a batch of t tasks.

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

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

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

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 shown in (10). P_W is not expressed in terms of P_{FP} and P_{TN} because the formula is generated to cover 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.

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

4.3. Cost Estimation Process

In this section we introduce a process through which cost models can be applied in practice (Figure 11). 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, based on reported findings of Hirth et al., if the main task is significantly more complex than the control task, utilizing a *control group* technique is more cost effective than *redundancy* (Hirth et al., 2013).

The second step is to obtain P values (P_P, P_{IC}, P_W, P_{FP}, P_{FN}, P_{TN}, P_{TP}). In order to measure P values, a pilot study can be conducted in an environment which represents the real life as good as possible. 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 needs to be decided based on the specific characteristics of the work and the practitioner's management style. Alternatively, these probabilities may be iteratively estimated using data collected throughout the course of a crowdsourcing project, thereby allowing a convergence process.

The next step consists of constructing the model by simply using the obtained P, C_{err} and C_{dmg} values. Finally, the cost models must be executed to calculate the final cost estimation. This can be done by simply calculating the cost model formula with obtained parameters, or by using supporting estimation techniques such as Program Evaluation and Review Technique (PERT) (Malcolm, Roseboom, Clark, & Fazar, 1959), or Monte Carlo (Fishman, 1996).

PERT is an estimation technique which is commonly used in project management. This technique requires three distinct values for each parameter. Optimistic, pessimistic and most-likely values for parameters are determined. These values can be obtained from the measurements made in the pilot project. The pessimistic value can be assigned with the lowest observed P value, while optimistic value represents the largest observed P value. Most likely is the mean of observed P values. Thus, the expected cost is calculated by using the PERT formula:

Expected Cost = (Optimistic + 4 Most Likely + Pessimistic) / 6 (11)

Monte Carlo is a statistical simulation practice used in various domains including the project management. Monte Carlo is a slightly more complex technique for practical usage. Monte Carlo operates by running the simulation (in this case the probabilistic cost models) many times and produces a probabilistic distribution of the outcomes. Tools exist which can assist the practitioners to run the given cost models in a Monte Carlo simulation and provides cost intervals along with the probability of these outcomes.

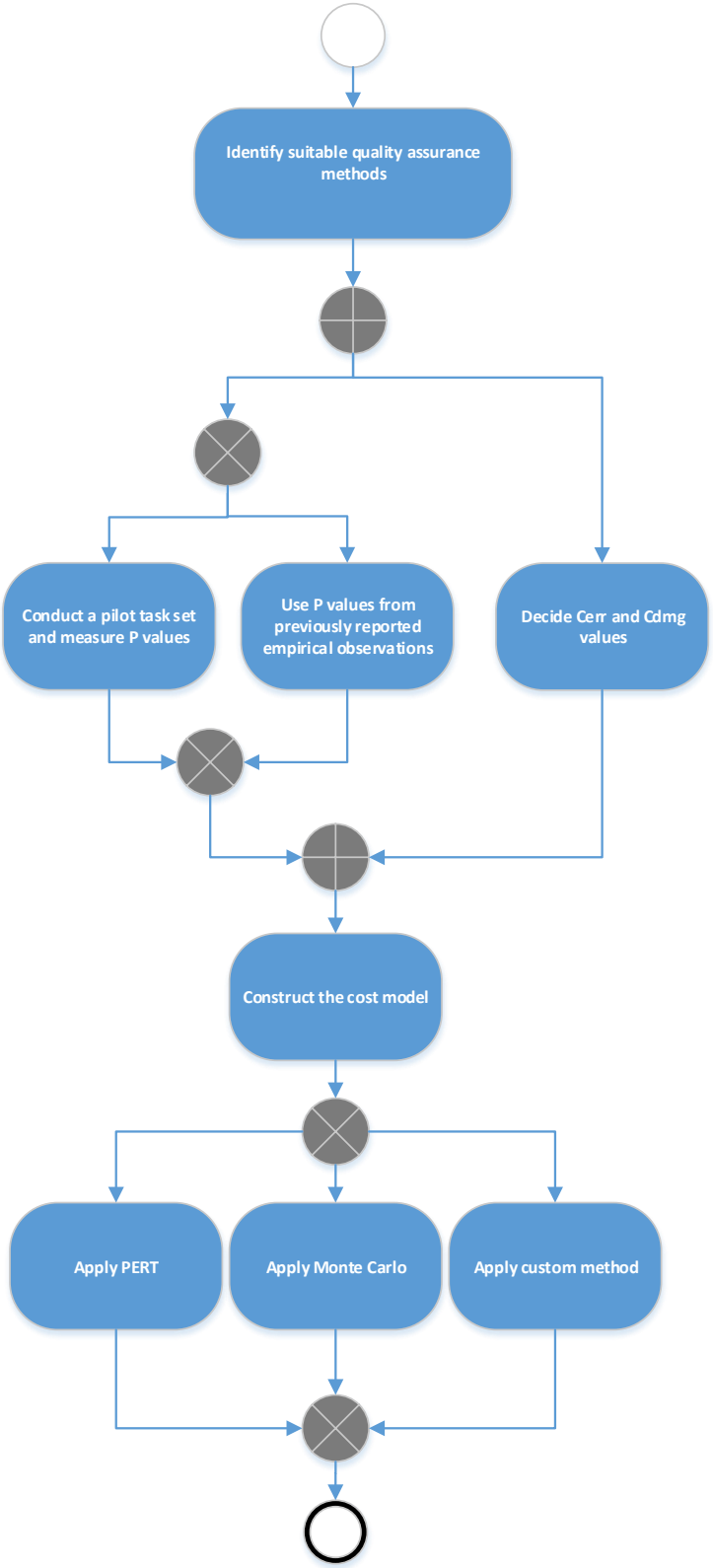


Figure 11 - Sample cost model utilization process

CHAPTER 5

CoQ EVALUATION: MULTIPLE ACTION RESEARCH

In this study, 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 covering different design-time characteristics of *nature of task* and *crowd type*. The same run-time quality assurance techniques were utilized in all three cases.

The primary research goals of these cases were to observe and measure the probabilities of quality assurance outcomes and to validate these probabilities using statistical cross validation techniques.

These probability values are used to construct the cost models. Cross validation techniques were applied to determine if the probability values come from the same distribution, thus ascertaining their predictability.

The first case constitutes an experiment involving image illustration and evaluation tasks performed by an external crowd on AMT. These tasks can be classified as subjective. The second case covers a real-life big data cleaning project which consists of objective tasks carried out by an external crowd on AMT. The third case focuses on 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 provided 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 to 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 of these cases the number of

redundant tasks is odd (m=3), thus reaching an IC is not possible. Nevertheless this parameter is kept for completeness.

Probability of reaching a positive outcome when the quality of the work is actually poor, is represented by P_{FP}.

P_{FN} represents the probability of the workers submitting a valid result but incorrectly decided as invalid by the *control group*. P_{TN} is the probability of the workers submitting an invalid result and correctly identified as invalid by the *control group*. Finally, P_{TP} is the probability of workers submitting a valid result and correctly identified as valid by the controllers.

Observed P values of all outcomes are reported in respective tables presented at the end of each individual action research section.

The cost models also include C_{err} and C_{dmg} parameters. These values should be determined according to the project environment, crowd characteristics and risk appetite of the practitioner. Therefore different values are used in each case and these values are reported in individual case sections.

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 validity of the observations was checked by using 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. As a result Mean MRE (MMRE) values were obtained and used to evaluate the validity of the observations.

5.1. Case 1: Illustration and Evaluation of Simple Images: CoQ of Subjective Microtasks

This action research experiment addressed the production of a large number of hand-drawn simple images to be used in the 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 of this action research was to observe the process outcomes of common crowdsourcing quality assurance techniques when applied on subjective tasks. The action research consisted of two phases. 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 which employ 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 determined by comparing the submissions against the expert judgment. The details of the task design, observed results and measurements for quality assurance technique effectiveness can be found in (Iren & Bilgen, 2013b) and (Iren & Bilgen, 2014).

5.1.1. 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 an expert evaluation. The entire image set 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) design, 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) design is almost the same as CG voting 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) design, workers were presented 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 secondary tasks until all images in the lizard image set were evaluated three times, making application of *redundancy* on secondary tasks possible. These redundant evaluations were used to derive a majority decision. Therefore, all three designs were analyzed both with and without *redundancy*.

5.1.2. 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 consists 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 is reported to be 0.34. P_W values for secondary tasks are shown in Table 5.

Probability outcome values of quality assurance processes are shown in Table 5. The meaning of each probability parameter was described in Chapter 4. Table 5 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 5 - 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	N/A	N/A	N/A	N/A
	With Redundancy	-	0.00	**0.08**	-	-	-
Expert Evaluation		**0.00**	**N/A**	**0.00**	**0.00**	**0.35**	**0.65**

Estimations are critical at the early phases of a project. Thus, not only accuracy of estimations but also their early availability is an important goal. We examined the change of our probability observation accuracy over time. Figure 16 to Figure 27 in Appendix A depict the variations of various probability outcomes throughout the course of the crowdsourcing project. We measured the probability of respective outcome for each group of consisting of 100 microtask instances.

Table 6 shows the summary of outcome frequency observations made in this case. The largest variation between the final mean of outcome frequency and timely observations was identified in CG rating P_{TN} with the variation of 0.10. The figures indicate that even early estimations performed with the proposed model do not deviate more than 0.10 and estimations converge to the final frequency mean in time.

Table 6 – Summary table of outcome frequency observations through time

		Observed Final Frequency	Maximum Variation
GS Rating	P_{TN}	0,27	0,03
	P_{TP}	0,42	0,05
	P_{FN}	0,23	0,03
	P_{FP}	0,08	0,01
CG Rating	P_{TN}	0,30	0,10
	P_{TP}	0,37	0,06
	P_{FN}	0,28	0,06
	P_{FP}	0,06	0,05
CG Voting	P_{TN}	0,19	0,08
	P_{TP}	0,56	0,03
	P_{FN}	0,09	0,02
	P_{FP}	0,16	0,09

5.1.3. Validation

We applied v-fold cross validation on the set of observed quality assurance outcomes as described in Section 5. The frequency of cross validation random partitioning outcomes and descriptive statistics are provided in Figure 28 to Figure 39 in Appendix A. Each figure displays the frequency of a quality assurance outcome per partition, the mean frequency and upper and lower limits with -/+ 2 standard deviations. Descriptive statistics indicate that each randomly generated partition has a similar outcome frequency distribution. The majority of outcome observations fall within defined upper and lower limits.

V-fold cross-validation (Arlot & Celisse, 2010) of the observed probability outcomes reported in Table 5, yields following MMRE values, where V is 15 and group size is 100:

- $MMRE_{CG\ Voting} = 0.12$
- $MMRE_{CG\ Rating} = 0.15$
- $MMRE_{GS\ Rating} = 0.14$

As widely used in software engineering (Foss, Stensrud, Kitchenham, & Myrtveit, 2003) MMRE values smaller than 0.2 are considered acceptable for prediction models (Conte, Dunsmore, & Shen, 1985). Since the MMRE calculations were below this threshold we confirm that the proposed cost model has significant predictive power.

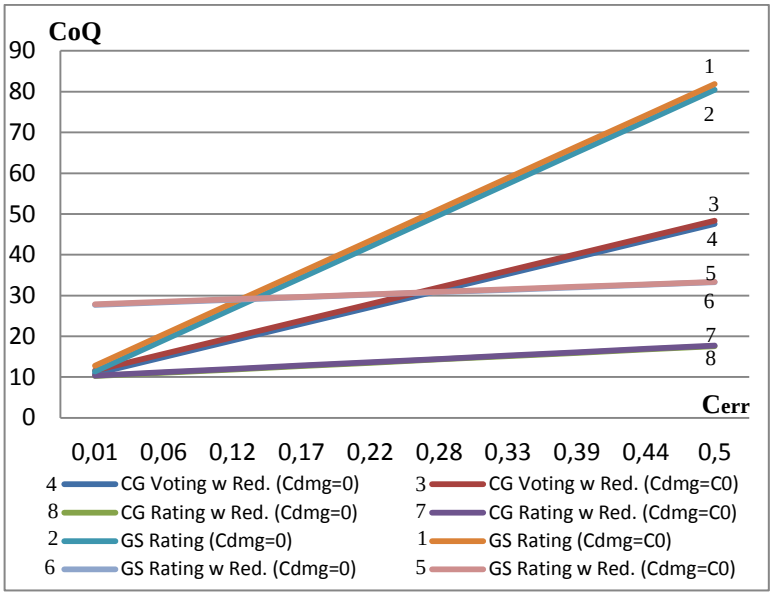
5.1.4. CoQ Calculations

Cost formulas presented in Section 4.2 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 5 were used as parameters in CoQ formulas. The results are summarized in Table 7.

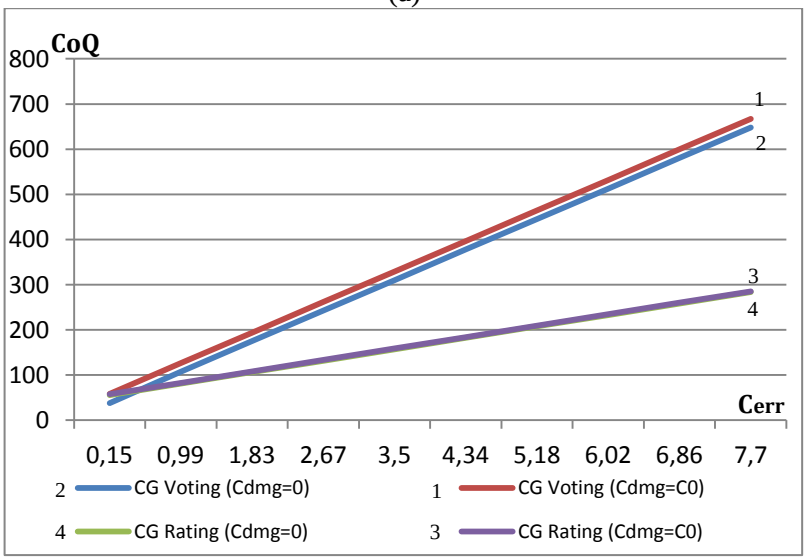
Table 7 - CoQ calculations

Design		CoQ
CG Voting	Single	$27.62 + 126 . C_{dmg} + 80.64 . C_{err}$
	w. Red.	$10.08 + 75.6\ C_{dmg} + 75.6\ C_{err}$
CG Rating	Single	$51.01 + 171.36 . C_{dmg} + 30.24\ C_{err}$
	w. Red.	$10.08 + 15.12\ C_{dmg} + 15.12\ C_{err}$
GS Rating	Single	$40 . C_{exp} + 5.95 + 142.18 . C_{err} + 142.18 . C_{dmg}$
	w. Red.	$40 . C_{exp} + 27.62 + 11.37 . C_{err} + 11.37 . C_{dmg}$

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 12.



(a)



(b)

Figure 12 - The effect of changing C_{err} and C_{dmg} on CoQ of various crowdsourcing designs

5.1.5. Findings

The most critical risk in the reported experimental action research setting was that the chosen tasks were subjective. However, subjectivity is a fact of crowdsourcing tasks and techniques exist which can be used to derive statistically significant results over subjective data. (Ribeiro, Florencio, & Nascimento, 2011).

Figure 12a shows CoQ of GS rating, GS rating with redundancy, CG voting with redundancy and CG rating with redundancy designs. According to the results 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 less error to remain undetected.

Figure 12b shows the CoQ of CG voting and CG rating designs for varying C_{err} values. According to the results, CG rating proves to be a more robust design against the impacts of undetected errors. Our observations indicate 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 technique of controlling than voting which may lead to less undetected errors. According to these findings it is concluded that a rating scheme is better than voting when EF tolerance is low but IF is more acceptable.

We observed that even the early estimations do not vary more than 0.10 compared to the final outcome frequency (Table 6). Therefore we consider early estimations usable for cost planning purposes.

5.2. Case 2: Big Data Analysis: CoQ of Objective Microtasks

This action research addresses a data cleaning and migration project recently undertaken in the Middle East Technical University (METU). The IT structure of METU combines many legacy applications and contains a large amount of data. Recently a project was initiated to integrate key components of this IT structure as automated business processes. This major overhaul caused some of the legacy data to be migrated to the newly developed systems. METU employs over 2,500 academic personnel who are actively engaged in research. As a result a large amount of publications are produced yearly. 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 is to normalize the data, to clean the duplicates, to fix typographical errors and to migrate the data to the newly developed system. The research goal is to apply common crowdsourcing quality assurance techniques in the solution and observe the probability of quality assurance process outcomes.

5.2.1. 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. Thus, a crowdsourcing solution was developed. Detailed information of the previous stages can be found in (Iren, Kul, & Bilgen, 2014).

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, the records were gathered in a combination of similar pairs, so that all similarity instances were represented in the pair set. Combining the records in pairs caused increase in the 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*". Thus, the total number of crowdsourcing 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 *gold standard* technique, a set of 100 gold standard pairs was 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.

5.2.2. 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 the expert judgments. The occurrence counts of observed quality assurance process outcomes are shown in Table 8. The meanings of the parameters are explained in Chapter 4.

Table 8- 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	N/A	N/A	N/A	N/A

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 9. 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. The probability outcomes are displayed in Figure 13.

Table 9 - Probability values of quality assurance process outcomes

	P_W	P_{IC}	P_{FP}	P_{FN}	P_{TN}	P_{TP}
Control Group	**-**	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	N/A	N/A	N/A	N/A
Expert Evaluation	**0.00**	**N/A**	**0.00**	**0.00**	**0.00**	**0.00**

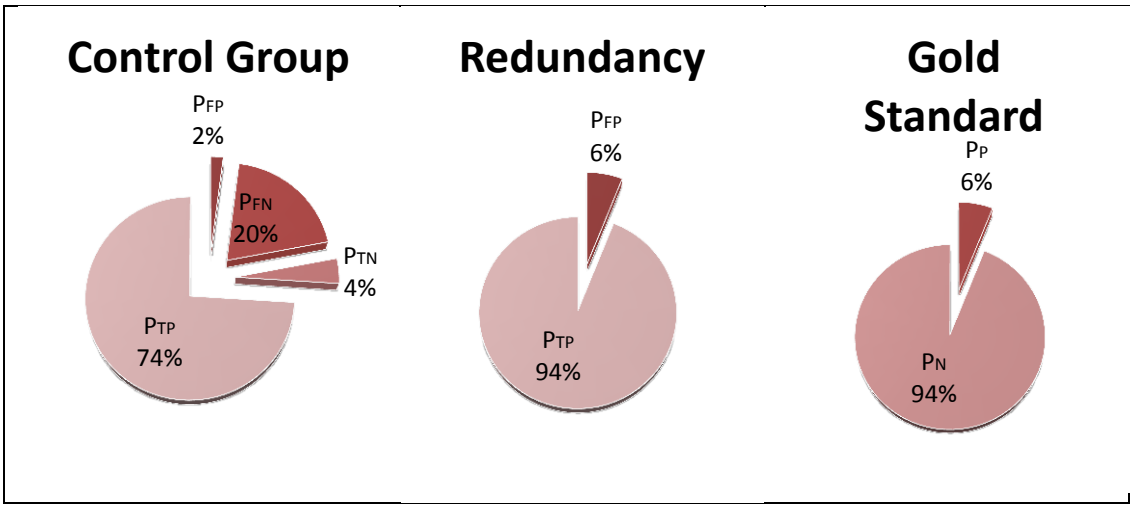


Figure 13 – Observed probabilities of quality assurance outcomes

In this action research case we examined the change of our probability observation accuracy over the time. Figure 40 to Figure 47 in Appendix B show our observations of various probability outcomes. We measured the probability of respective outcome for each group of consisting of 400 microtask instances.

Table 10 shows the summary of outcome frequency observations made in this case. The largest variation between the final mean of outcome frequency and timely observations was identified in gold standard P_P and gold standard P_N with the variation of 0.06. The figures indicate that even early estimations performed with the proposed model do not deviate more than 0.06 and estimations converge to the final frequency mean in time.

Table 10- Summary table of outcome frequency observations through time

		Observed Final Frequency	Maximum Variation
Control Group	P_{TN}	0,04	0,03
	P_{TP}	0,74	0,02
	P_{FN}	0,20	0,04
	P_{FP}	0,02	0,01
Redundancy	P_{TP}	0,94	0,02
	P_{FP}	0,06	0,01
Gold Standard	P_P	0,94	0,06
	P_N	0,06	0,06

5.2.3. Validation

The observations were validated by using V-fold cross validation technique which is explained in the Section 5. The frequency of outcomes of cross validation random partitioning and descriptive statistics are provided in Figure 48 to Figure 55 in Appendix B. Each figure displays the frequency of a quality assurance outcome per partition, the mean frequency and upper and lower limits with -/+ 2 standard deviations. Descriptive statistics indicate that each randomly generated partition has a similar outcome frequency distribution. Majority of outcome observations were within defined upper and lower limits.

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). The MMRE calculations were below this threshold, thus, we confirm the proposed model has significant predictive power.

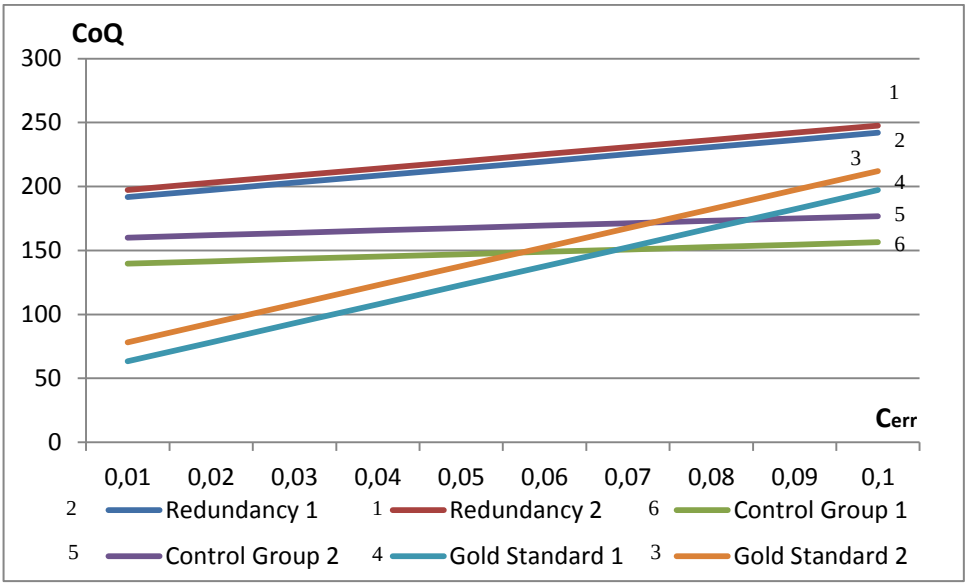
5.2.4. CoQ Calculations

Table 11 presents the CoQ calculations. In order to observe the change of CoQ of quality assurance techniques, two separate C_{err} and C_{dmg} values are used and thus Figure 14 is derived.

Table 11 - CoQ calculations

Design	CoQ
Control Group	$137.76 + 2047.76 \cdot C_{dmg} + 186.16 \cdot C_{err}$
Redundancy	$186.16 + 558.48 \cdot C_{dmg} + 558.48 \cdot C_{err}$
Gold Standard	$100 \cdot C_{exp} + 38.47 + 1487.42 \cdot C_{err} + 1487.42 \cdot C_{dmg}$

As described in Chapter 5, two different C_{err} and C_{dmg} values were used to observe the effect of changing impact on total CoQ and the results are displayed in Figure 14. 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 $10 \cdot C_0$.



Redundancy 1: (C_{dmg}=0), Redundancy 2: (C_{dmg}=C_0), Control Group 1: (C_{dmg}=0), Control Group 2: (C_{dmg}= C_0), Gold Standard 1: (C_{dmg}=0), Gold Standard 2: (C_{dmg}= C_0).

Figure 14 - The effect of changing C_{err} and C_{dmg} on CoQ of various crowdsourcing designs

5.2.5. Findings

In this action research 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.

5.3. Case 3: Campus Phonebook Registry Update: CoQ of Objective Wisdom of Crowds Type Crowdsourcing

This action research was also conducted in METU. In late 2011 a project was initiated to establish the corporate identity of METU. Project mainly consists of developing social media identities and transferring the websites to a corporate content management system. Project also includes 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 of this action research is to update the corporate phonebook with accurate assignments.

The research objective is to apply common crowdsourcing quality assurance techniques and to observe the quality assurance process outcomes.

5.3.1. Method

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

In this enterprise crowdsourcing setting the crowd consists of 5,500 university personnel. The microtasks are 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 are observed by examining the user action logs.

5.3.2. Measurements

Data collection phase lasted two weeks and then terminated. As a result 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 university entrance gates. System was designed to display 1 gold standard task for 2 regular tasks. If the workers provide wrong 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 consists of 328 records was controlled by experts manually. Correctness of user answers was decided based on expert evaluation. Observed quality assurance process outcomes are presented in Table 12.

Table 12 - 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	N/A	N/A	N/A	N/A

The probability values of quality assurance process outcomes are presented in Table 13. In this case P_N for *gold standard* process outcome was observed as 0.02 while P_P was observed as 0.98.

Table 13 - Probability values of quality assurance process outcomes

	P_W	P_{IC}	P_{FP}	P_{FN}	P_{TN}	P_{TP}
Control Group	-	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	N/A	N/A	N/A	N/A
Expert Evaluation	0.00	N/A	0.00	0.00	0.00	0.00

In this action research case we examined the change of our probability observation accuracy over the time. Figure 56 to Figure 63 in Appendix C show our observations of various probability outcomes. We measured the probability of respective outcome for each group of consisting of 32 microtask instances.

Table 14 shows the summary of outcome frequency observations made in this case. The largest variation between the final mean of outcome frequency and timely observations were observed in control group P_{TP} with the variation of 0.07. The figures indicate that even early estimations performed with the proposed model do not deviate more than 0.07 and estimations converge to the final frequency mean in time.

Table 14- Summary table of outcome frequency observations through time

		Observed Final Frequency	Maximum Variation
Control Group	P_{TN}	0,07	0,04
	P_{TP}	0,84	0,07
	P_{FN}	0,02	0,02
	P_{FP}	0,08	0,02
Redundancy	P_{TP}	0,94	0,04
	P_{FP}	0,06	0,03
Gold Standard	P_P	0,99	0,02
	P_N	0,01	0,02

5.3.3. Validation

We applied v-fold cross validation technique on the set of observed quality assurance outcomes as explained in the Chapter 5. The frequency of outcomes of cross validation random partitioning and descriptive statistics are provided in Figure 64 to Figure 67 in Appendix C. Each figure displays the frequency of a quality assurance outcome per partition, the mean frequency and upper and lower limits with -/+ 2 standard deviations. Descriptive statistics indicate large variances and figures display outliers in the observations. This can be explained by the small size of the data sets.

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 data sets 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.

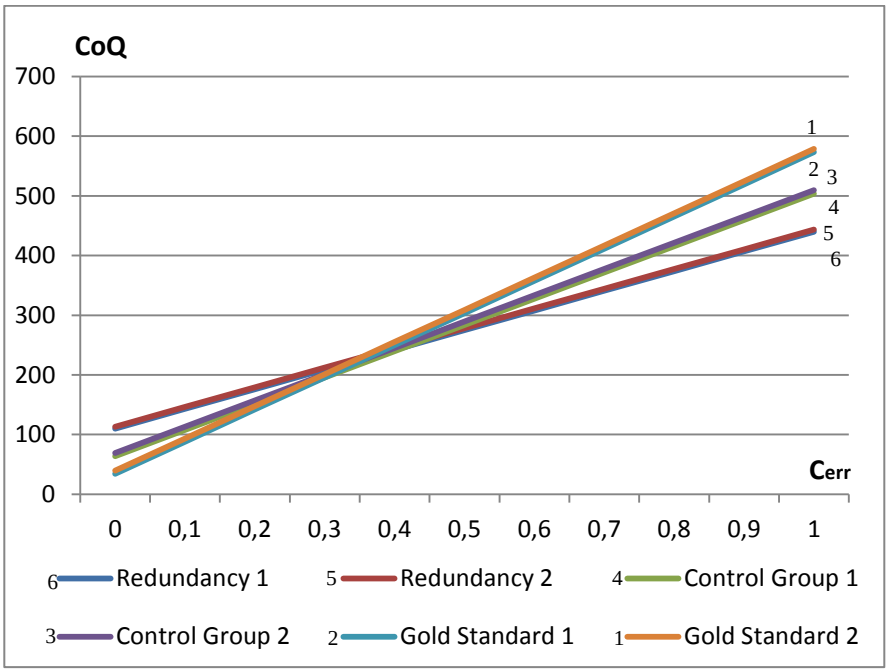
5.3.4. CoQ Calculations

CoQ calculations are presented in Table 15. 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. 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, which is equal to $0.1.

Table 15 - CoQ calculations

Design	CoQ
Control Group	$63.8 + 550 \cdot C_{dmg} + 440 \cdot C_{err}$
Redundancy	$110 + 330 \cdot C_{dmg} + 330\ C_{err}$
Gold Standard	$50 \cdot C_{exp} + 29.15 + 539 \cdot C_{err} + 539 \cdot C_{dmg}$

Two different C_{err} and C_{dmg} values were used for observing the impact of change on total CoQ and the results are displayed in Figure 15. 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.



Control Group 1: (C_{dmg}=0), Control Group 2: (C_{dmg}= C_0), Redundancy 1: (C_{dmg}=0), Redundancy 2: (C_{dmg}= C_0), Gold Standard 1: (C_{dmg}=0), Gold Standard 2:(C_{dmg}= C_0).

Figure 15 - The effect of changing C_{err} and C_{dmg} on CoQ of various crowdsourcing designs

5.3.5. Findings

This case is different than 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 in this case an internal crowd was used in place of 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 other cases. This can be explained with the fact that the identities of the workers are known and workers complete the tasks with a higher sense of accountability compared to anonymous workers.

According to the Figure 15, *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*.

5.4. Interviews

We conducted semi-structured interviews with stakeholders who were directly involved with the action research projects. We interviewed the vice president of METU, director of computer center, project manager, the project sponsor and two software developers. Project sponsor is the vice president of METU. During the interview sessions we recorded the conversations. Questions and summaries of answers which were gathered from the raw voice recordings are provided in Table 20.

Both the Project Manager and the Software Engineer 1 reported that they observed a higher motivation to participate in crowdsourcing compared to the traditional way of performing the task. The Director of Computer Center, Project Manager and Software Engineer 2 emphasized that crowdsourcing enabled them to access a scalable workforce which is otherwise inaccessible. Project Manager and Software Developer 2 mentioned that persuading the upper management about the effectiveness of crowdsourcing was a challenge for them. Vice President of METU stated that the execution of crowdsourcing sufficiently fitted the plans indicating the estimations were accurate. All participants reported that faster completion times and lower costs were achieved via crowdsourcing. Director of Computer Center stated that a major challenge of crowdsourcing was unreliable crowd workers which leads to application of excessive quality control techniques which in turn, increases the quality costs significantly. Project Manager stated that she was able to estimate and plan crowdsourcing by using the proposed estimation methods. All participants stated that they consider crowdsourcing as a valid way of problem solving and that they would use crowdsourcing in the future. All participants approved that they would use the proposed cost estimation methods in the future.

5.5. Results

In order to compare the cost of quality assurance techniques, the calculations were normalized to reflect the ratio of CoQ to total cost of the product, excluding the cost of all quality related activities. The effectiveness of the quality assurance techniques were calculated by using *Decision Fitness* (DF) measure (12). For this analysis C_{err} is assumed to be equal to C_0.

$$DF = P_{TN} + P_{TP} \qquad \textbf{(12)}$$

Both normalized CoQ and DF calculations are presented in Table 16.

Table 16 - 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. In summary, 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 technique 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.

We performed sensitivity analysis on the proposed CoQ models to determine the quality assurance outcomes with greater impact on CoQ. Since all CoQ models we propose are linear, the sensitivity of the model against the change in individual quality assurance outcomes as parameters can be determined directly by the coefficients of particular model as displayed in Table 17.

Table 17 – Model parameter coefficients that reflect model sensitivity

	Redundancy	Control Group	Gold Standard
m	$N . C_0 + N. P_{IC}. C_0$	-	-
P_{IC}	-		-
P_{TN}	-	$N . (C_0 + C_1)$	-
P_{TP}	-	-	-
P_{FN}	-	$N . (C_0 + C_1 + C_{dmg})$	-
P_{FP}	$N . (C_{err} + C_{dmg})$	$N . (C_{err}+ C_{dmg})$	-
P_P'	-	-	$N . (k/(t-k)) . (t-k) . (C_{err} + C_{dmg})$ *
P_N'	-	-	$N . (k/(t-k)) . t . C_0$ *
P_W	-	-	$N . (k/(t-k)) . (t-k) . (C_{err} + C_{dmg})$

* Gold standard cost models include the expression $(P_P)^k$. For ease of calculation, we denote this expression as P_P'.

We assume that C_{dmg} is either equal to 0 or C_0 as we assumed when calculating CoQ in action research cases throughout this study. We also examined the cases in which C_{err} is significantly greater than C_0 and vice versa. For each case model sensitivity was determined by comparing the coefficients with given cost parameters and quality assurance outcomes with greater impact on CoQ were displayed in Table 18.

Table 18 – Outcome parameters with greater impact on CoQ in different cases

$C_{dmg} = 0$		$C_{dmg} = C_0$	
$C_{err} >> C_0$	$C_0 >> C_{err}$	$C_{err} >> C_0$	$C_0 >> C_{err}$
P_{FP}, P_P', P_W	P_{FN}, P_{TN}, P_N', m	P_{FP}, P_P', P_W	P_{FN}

Hence, in cases when the impact of accepting a poor quality submission is high (C_{err}) decreasing FP outcomes becomes critical which may require utilizing stricter quality assurance techniques or introducing additional levels of quality assurance.

Finally, we calculated the specificity and sensitivity of the models by using following formulas:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

The results of the specificity and sensitivity calculation are shown in Table 19.

Table 19 – Results of specificity and sensitivity measurements

Case	Crowd Type	Task Type	Quality Assurance Technique	Specificity	Sensitivity
1	AMT Workers	Subjective	CG Voting	0.54	0.86
			CG Rating	0.83	0.57
			CG Voting w. Red	0.57	0.92
			CG Rating w. Red	0.91	0.60
			GS Rating	1.00	1.00
2	AMT Workers	Objective	Control Group	0.67	0.79
			Redundancy	0.00	1.00
			Gold Standard	1.00	1.00
3	Internal Crowd	Objective	Control Group	0.43	0.98
			Redundancy	0.00	1.00
			Gold Standard	1.00	1.00

5.6. Threats to Validity

We handled potential threats on construct validity as follows. All action research cases were planned and executed according to those plans. Outcome measurements and interview questions were reviewed and evaluated by peer researchers to ensure the appropriateness of these actions for answering the research questions and fulfilling the research goals.

In the first action research 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 action research cases. In order to preserve internal validity, the quality assurance techniques of the first action research were only compared to the ones which were applied on the same type of task. We acknowledge that there are additional factors affecting the CoQ besides the quality assurance outcomes. However our observations in multiple action research cases support that quality assurance process outcomes satisfactorily explain the effect on CoQ.

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

Worker activities were logged in database in all cases. All measurements and validity calculations were performed by software applications which were specifically

developed for this study. Therefore, if these measurements were to be conducted again by other research bodies, the same results would be reached, which supports the reliability claim.

CHAPTER 6

CONCLUSION

Due to unique characteristics of crowdsourcing, practitioners face certain problems regarding the quality and utilize various techniques for quality assurance. These techniques differ in terms of cost and effectiveness. This thesis introduces cost models of common quality assurance techniques, derived by using CoQ approach. The probability values which are used in the cost models were measured through observations on three different scenarios. These scenarios cover different task types (objective vs. subjective) which were performed by different types of crowds (internal vs. AMT workers).

6.1. Contributions

The main contributions of this research are the cost models of common quality assurance techniques and the CoQ estimation process. Applicability of this estimation process and cost models for different crowdsourcing scenarios were assessed within a multiple action research framework. The secondary contribution consists of the observations of probabilistic outcomes of quality assurance processes for different work and crowd types. These values can be used by other practitioners and researchers as a guideline.

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. By using the estimation process shown in Figure 11 and explained in Section 4.3 practitioners can calculate CoQ. Additionally, achievable quality levels of quality assurance techniques were provided in Table 16 represented with DF. Practitioners may derive cost effectiveness of quality assurance techniques by using CoQ and DF values together. Therefore they may use these values for decision making regarding quality assurance technique selection.

Enabling formal planning by basing decisions on procedural calculations is especially valuable in enterprise projects which have a low tolerance for uncertainty.

Quality assurance is a non-value added process. Especially in crowdsourcing, quality assurance techniques lead to massive amounts of wasted effort, significantly impacting project costs and durations.

The impact of this study can be better grasped when the current status of crowdsourcing is considered. The crowdsourcing market is still growing. Even though practitioners use crowdsourcing to access cheap and scalable workforces, inevitably, the market will eventually saturate. Therefore it is imperative to develop ways to achieve efficiency. 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. The cost models proposed in this study can be used to select quality assurance techniques which fit the job better or design efficient hybrid quality assurance techniques. We foresee that by enabling savings at microtask levels it is possible to make a significant improvement on crowdsourcing efficiency at a global scale. This study paves the way for future research aiming at quality and cost optimization.

6.2. Answers to Research Questions

(Q1) How can we estimate the costs of quality assurance techniques?

In this study we developed cost models of common quality assurance techniques used in crowdsourcing. We applied these models in multiple action research which covers different real-life crowdsourcing scenarios with various characteristics. We were able to estimate the impact of each quality assurance technique on CoQ.

(Q2) Can cost of quality models be used to support decision making of practitioners regarding technique selection and assist them to avoid inefficiencies?

The results of the multiple action research indicate that the cost models can be used to evaluate the cost effectiveness of quality assurance techniques. Therefore it is possible to use the models for selecting the more appropriate quality assurance technique in crowdsourcing. The interview results confirm the applicability and effectiveness of cost models in estimation and project planning.

6.3. Discussion

Even though most of the microtask crowdsourcing platforms operate with prepayment in which requesters pay money to obtain credits and later spend those credits to make payment to the workers, this does not necessarily mean that the cost of the job is previously known. Hidden opportunity costs exist. Especially in enterprise crowdsourcing, these hidden costs can become significant. As the crowd consists of an

organization's own personnel whose primary job is different from performing the crowdsourcing tasks, effort spent of crowdsourcing leads to lost revenue for the organization. Therefore modeling, estimating and measuring crowdsourcing CoQ is important.

The cost models introduced in this thesis can be used to estimate the costs that occur according to the quality assurance technique selection or design. The cost models include probabilistic parameters. These parameters depend on various characteristics such as the crowd, nature of work and incentive mechanisms. Crowdsourcing practitioners can use simulations to calculate cost estimations, which may guide them to make better quality assurance technique selections or designs. With more realistic probability values, the estimations will be more accurate. Thus crowdsourcing practitioners are advised to observe crowd behavior and the effects of design decisions on this behavior, and use the observed probabilities as parameters with the CoQ models.

When analyzing costs of potential outcomes of quality assurance techniques we considered costs of damages done to worker communities and trust mechanisms. We understand that these cost values may not be estimated accurately. However it is important for crowdsourcing practitioners to understand the long lasting side effects and indirect costs of quality assurance techniques they use, in order to enable crowdsourcing as a sustainable means of production.

We used linear models to represent the CoQ. Since we were able to validate these linear models with observations, we did not attempt to increase model complexity and try non-linear modeling.

Sensitivity and specificity of these findings have also been investigated, leading to the evaluations in Table 19. Sensitivity and specificity calculation is used in detecting type 1 and type 2 errors of a prediction method. Additionally this approach can be used to determine the predictive power of quality assurance techniques for detecting both defects and good quality inputs. Therefore, it enables the analysis and comparison of quality assurance techniques in terms of strictness. However, the outcomes of Redundancy and Gold Standard quality assurance techniques do not include all outcome set of TN, FN, FP and TP. Therefore sensitivity and specificity calculation yielded results which may not be effectively used in comparison, for this study. To determine the accuracy of a quality assurance technique we used DF which is a more suitable metric for our research design.

6.4. Limitations of the Study and Future Work

In this thesis we specifically focused on *run-time* common quality assurance techniques, developing cost models which represent these techniques. However, CoQ is also affected by *design-time* quality assurance approaches. In the future, techniques to estimate costs of design-time quality assurance approaches need to be developed to

achieve a more comprehensive control over crowdsourcing costs. We propose focusing on developing best practices and heuristics considering factors such as; task granularity, worker identification and monitoring and design of better user interfaces.

The cost of crowdsourcing depends on many parameters besides quality costs which are covered in this study. Unveiling the effects of these parameters on crowdsourcing costs is an important research goal to be pursued in the future. Furthermore, to make crowdsourcing more manageable, certain practices of project management domain can be exploited. A valid research agenda exists for crowdsourcing management, including measurement, estimation and optimization of cost, time and quality aspects of crowdsourcing.

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 Inc. (n.d.). Retrieved from www.amazon.com

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

Antin, J. (2012). Social Desirability Bias and Self-Reports of Motivation : A Study of Amazon Mechanical Turk in the US and India, 2925–2934.

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

Avison, D. E., Lau, F., Myers, M. D., & Nielsen, P. A. (1999). Action research. *Communications of the ACM*, *42*(1), 94–97.

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS quarterly*, *11*(3).

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: the international journal of research into new media technologies*, *14*(1), 75–90.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical ...* (pp. 286–295). Retrieved from http://dl.acm.org/citation.cfm?id=1699548

Cavaye, A. L. M. (1996). Case study research: a multi-faceted research approach for IS. *Information systems journal*, *6*(3), 227–242.

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

Davison, R., Martinsons, M. G., & Kock, N. (2004). Principles of canonical action research. *Information systems journal*, *14*(1), 65–86.

Difallah, D. E., Demartini, G., & Cudré-Mauroux, P. (2013). Pick-A-Crowd: Tell me what you like, and I'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 367–374).

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, *54*(4), 86–96.

Downs, J., & Holbrook, M. (2010). Are your participants gaming the system?: screening mechanical turk workers. In *… in Computing Systems* (pp. 0–3). Retrieved from http://dl.acm.org/citation.cfm?id=1753688

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

Eagle, N. (2009). txteagle: Mobile crowdsourcing. In *Internationalization, Design and Global Development*. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-02767-3_50

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

Foss, T., Stensrud, E., Kitchenham, B., & Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. *Software Engineering, IEEE Transactions on*, *29*(11), 985–995.

Geiger, D., & Seedorf, S. (2011). Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. In … (pp. 1–11). Retrieved from http://schader.bwl.uni-mannheim.de/fileadmin/files/schader/files/publikationen/Geiger_et_al._-_2011_-_Managing_the_Crowd_Towards_a_Taxonomy_of_Crowdsourcing_Processes.pdf

Gentry, C., Ramzan, Z., & Stubblebine, S. (2005). Secure distributed human computation. In *Proceedings of the 6th ACM conference on Electronic commerce* (pp. 155–164).

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*.

Help-find-Jim. (n.d.). Retrieved from http://www.helpfindjim.com/

Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2011). Cost-Optimal Validation Mechanisms and Cheat-Detection for Crowdsourcing Platforms. In *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing* (pp. 316–321). Ieee. doi:10.1109/IMIS.2011.91

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*.

Horton, J. J., Chilton, L. B., Paul, A. C., & Way, S. (n.d.). The Labor Economics of Paid Crowdsourcing, (1), 209–218.

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

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1–4.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

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.

Innocentive. (n.d.). Retrieved from www.innocentive.com

Ipeirotis, P. (n.d.). Demographics of Mechanical Turk.

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

Iren, D., & Bilgen, S. (2013a). *Cost models of crowdsourcing quality assurance mechanisms*.

Iren, D., & Bilgen, S. (2013b). *Validating cost of quality models in subjective non-deterministic microtask crowdsourcing*.

Iren, D., & Bilgen, S. (2014). Cost Models of Quality Assurance in Crowdsourcing. In *Proceedings of the 5th IEEE International Conference on Communications and Electronics (In publishing)*.

Iren, D., Kul, G., & Bilgen, S. (2014). Utilization of synergetic human-machine clouds: a big data cleaning case. In *Proceedings of the 1st International Workshop on CrowdSourcing in Software Engineering* (pp. 15–18).

iStockPhoto. (n.d.). Retrieved from www.istockphoto.com

Jain, R. (2010). Investigation of Governance Mechanisms for Crowdsourcing Initiatives. In *AMCIS*. Retrieved from http://www.virtual-communities.net/mediawiki/images/f/fd/Jain.pdf

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. (2010a). Quality Assurance for Human-Based Electronic Services : A Decision Matrix for Choosing the Right Approach, 421–424.

Kern, R., Thies, H., Bauer, C., & Satzger, G. (2010b). 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., Chi, E. H., & Suh, B. (2008a). Crowdsourcing user studies with Mechanical Turk. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, 453. doi:10.1145/1357054.1357127

Kittur, A., Chi, E. H., & Suh, B. (2008b). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (p. 453). New York, New York, USA: ACM Press. doi:10.1145/1357054.1357127

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

Kittur, A., Smus, B., Khamkar, S., & Kraut, R. (2011). Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ….* Retrieved from http://dl.acm.org/citation.cfm?id=2047202

Kittur, A., Smus, B., & Kraut, R. E. (2011). CrowdForge : Crowdsourcing Complex Work.

La Vecchia, G., & Cisternino, A. (2010). Collaborative workforce, business process crowdsourcing as an alternative of BPO. In *Current Trends in Web Engineering* (pp. 425–430). Springer.

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

Lévy, P., & Bonomo, R. (1999). *Collective intelligence: Mankind's emerging world in cyberspace*. Perseus Publishing.

Lewin, K. (1946). Action research and minority problems. *Journal of social issues*, *2*(4), 34–46.

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.

Mason, W., & Watts, D. (2010). Financial incentives and the performance of crowds. In *ACM SigKDD Explorations Newsletter*. Retrieved from http://dl.acm.org/citation.cfm?id=1809422

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

Microworkers. (n.d.). Retrieved from www.microworkers.com

Mob4hire. (n.d.). Retrieved from http://www.mob4hire.com/

My Starbucks Idea. (n.d.). Retrieved from www.mystarbucksidea.com

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

Oxford Dictionary: definition of rating. (n.d.). Retrieved from http://www.oxforddictionaries.com/definition/english/rating

Parameswaran, M., & Whinston, A. B. (2007). SOCIAL COMPUTING: AN OVERVIEW. *Communications of the Association for Information Systems*, *19*.

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).

Ribeiro, F., Florencio, D., & Nascimento, V. (2011). Crowdsourcing subjective image quality evaluation. In *Image Processing (ICIP), 2011 18th IEEE International Conference on* (pp. 3097–3100). doi:10.1109/ICIP.2011.6116320

Rogstadius, J., Kostakos, V., Kittur, A., & Smus, B. (2011). An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *ICWSM* (pp. 321–328). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2778/3295

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

Sell-a-Band. (n.d.). Retrieved from www.sellaband.com

Shaw, A. D., Hall, B., Horton, J. J., & Chen, D. L. (2011). Designing Incentives for Inexpert Human Raters Office 3020, 275–284.

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

Stewart, O., Lubensky, D., & Huerta, J. (2010). Crowdsourcing participation inequality: a SCOUT model for the enterprise domain. In *… of the ACM SIGKDD Workshop on …* (pp. 30–33). Retrieved from http://dl.acm.org/citation.cfm?id=1837895

Stvilia, B., & Twidale, M. (2008). Information quality work organization in Wikipedia. *... society for information ...*, *59*(6), 983–1001. doi:10.1002/asi

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

Susman, G. I., & Evered, R. D. (1978). An assessment of the scientific merits of action research. *Administrative science quarterly*, 582–603.

Threadless. (n.d.). Retrieved from https://www.threadless.com/

Threadless Inc. (n.d.). Retrieved from http://threadless.com/

Viitamäki, S. (2008). The FLIRT model of crowdsourcing.[viitattu 26.9. 2012]. *Saatavissa: http://www. scribd. com/fullscreen/20607704*.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, 319–326. doi:10.1145/985692.985733

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.

Walsham, G. (1995). Interpretive case studies in IS research: nature and method. *European Journal of information systems*, *4*(2), 74–81.

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.

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

Xia, T., Zhang, C., Xie, J., & Li, T. (2012). Real-time quality control for crowdsourcing relevance evaluation. *2012 3rd IEEE International Conference on Network Infrastructure and Digital Content*, 535–539. doi:10.1109/ICNIDC.2012.6418811

Youtube. (n.d.). Retrieved from www.youtube.com

Zhai, Z., Hachen, D., Kijewski-Correa, T., Shen, F., & Madey, G. (2012). Citizen Engineering: Methods for "Crowdsourcing" Highly Trustworthy Results. *2012 45th Hawaii International Conference on System Sciences*, 3406–3415. doi:10.1109/HICSS.2012.151

Zwass, V. (2010). Co-creation: Toward a taxonomy and an integrated research perspective. *International Journal of Electronic Commerce*, *15*(1), 11–48.

APPENDICES

APPENDIX A – SUPPLEMENTARY MATERIAL FOR ACTION RESEARCH 1

A1 - Observations of quality assurance outcomes through time

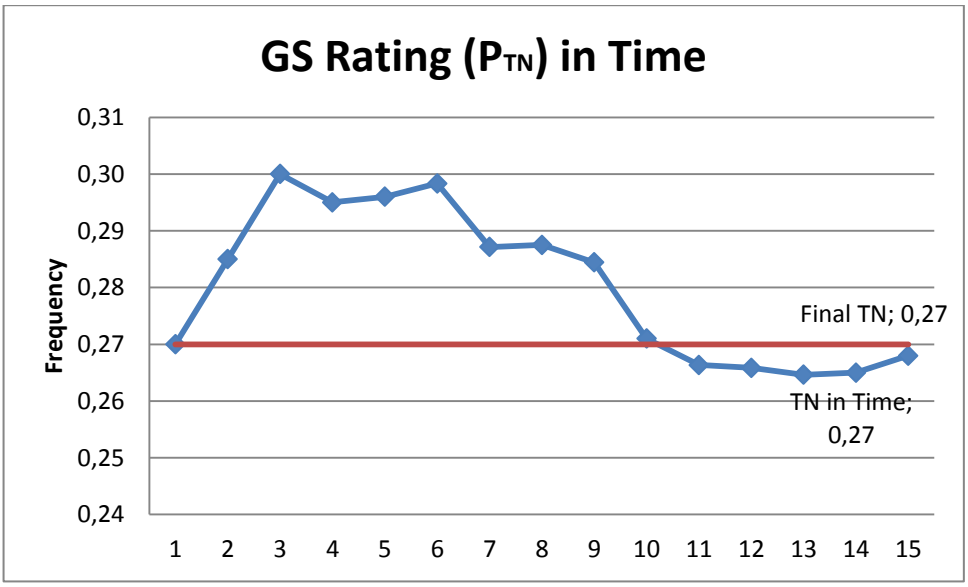


Figure 16 – Observations of P_{TN} for GS rating through time

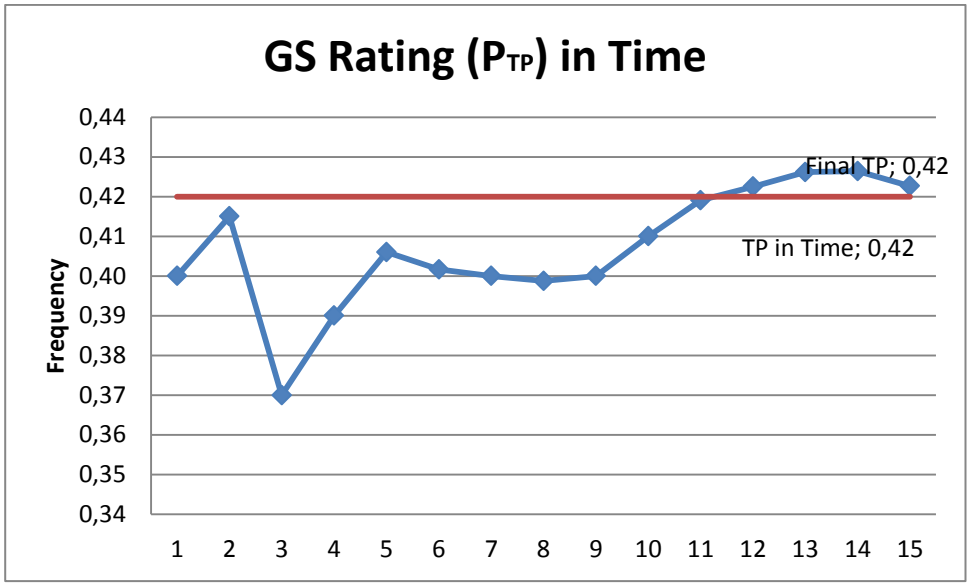


Figure 17– Observations of P_{TP} for GS rating through time

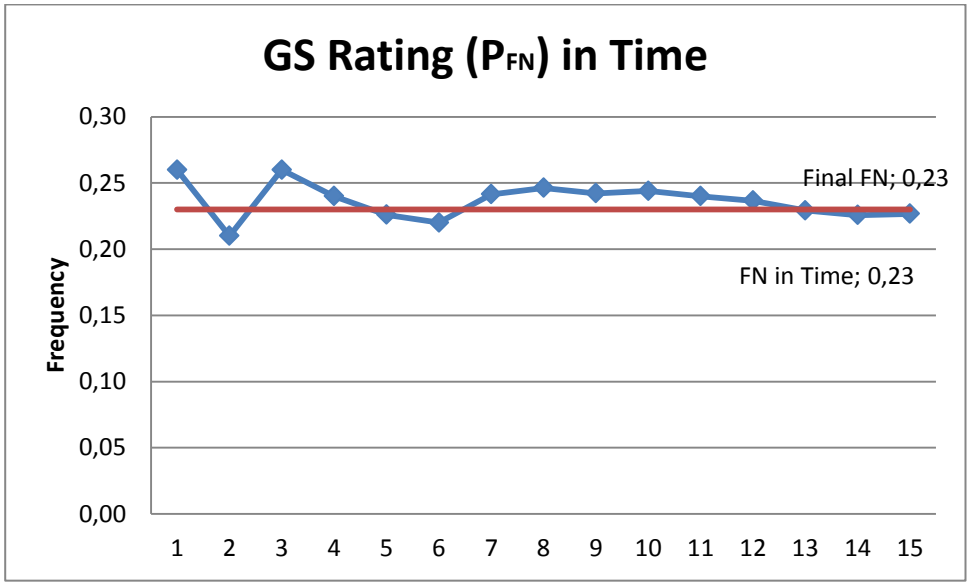


Figure 18– Observations of P_{FN} for GS rating through time

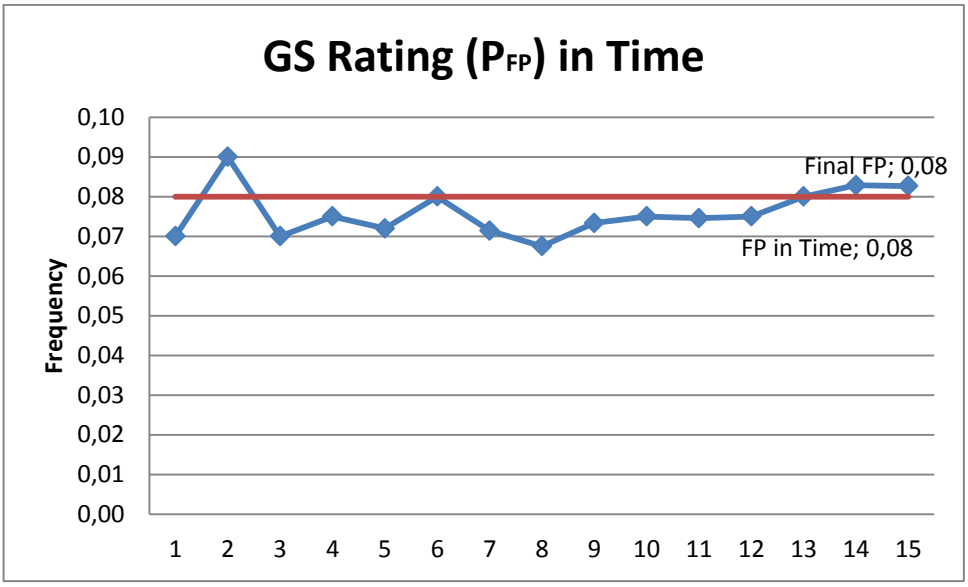


Figure 19– Observations of P_{FP} for GS rating through time

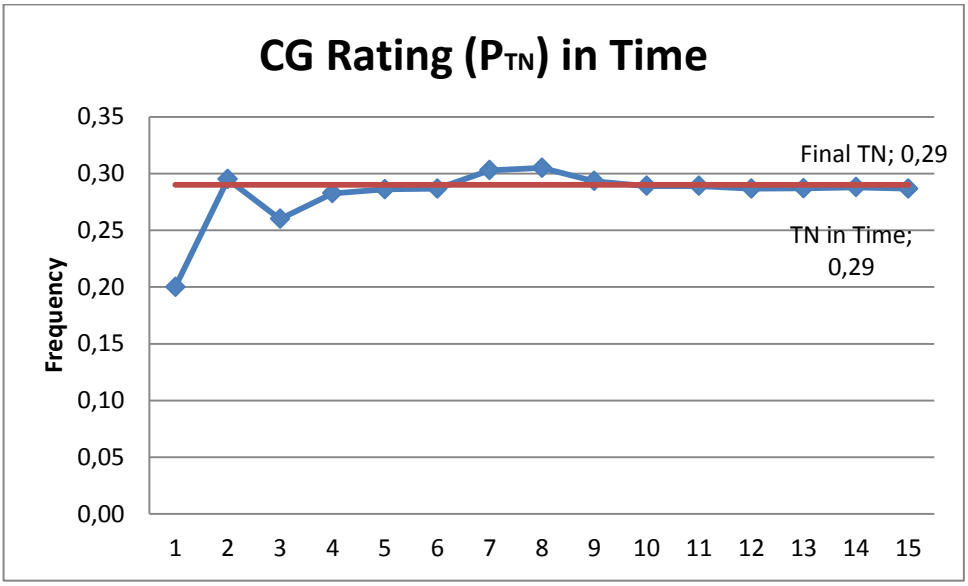


Figure 20– Observations of P_{TN} for CG rating through time

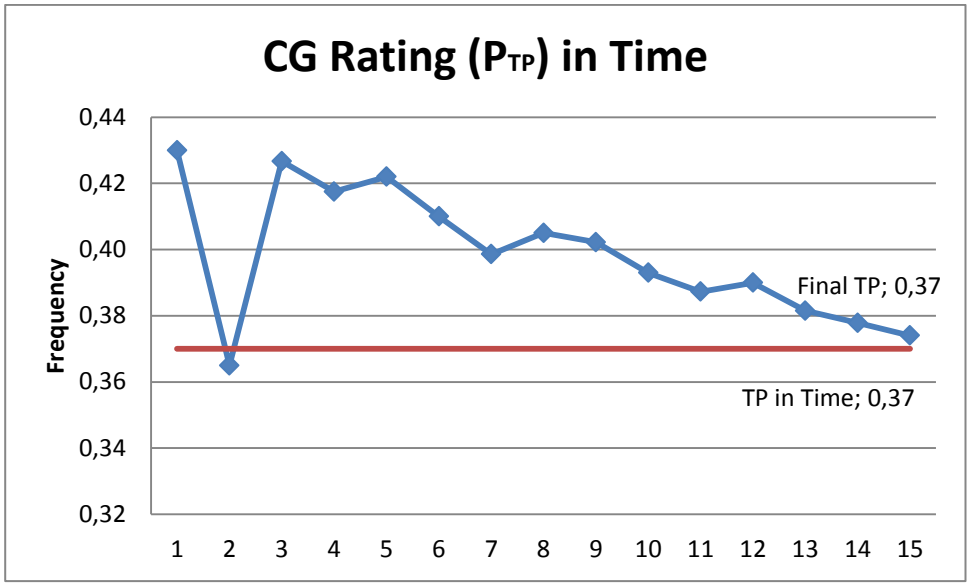


Figure 21– Observations of P_{TP} for CG rating through time

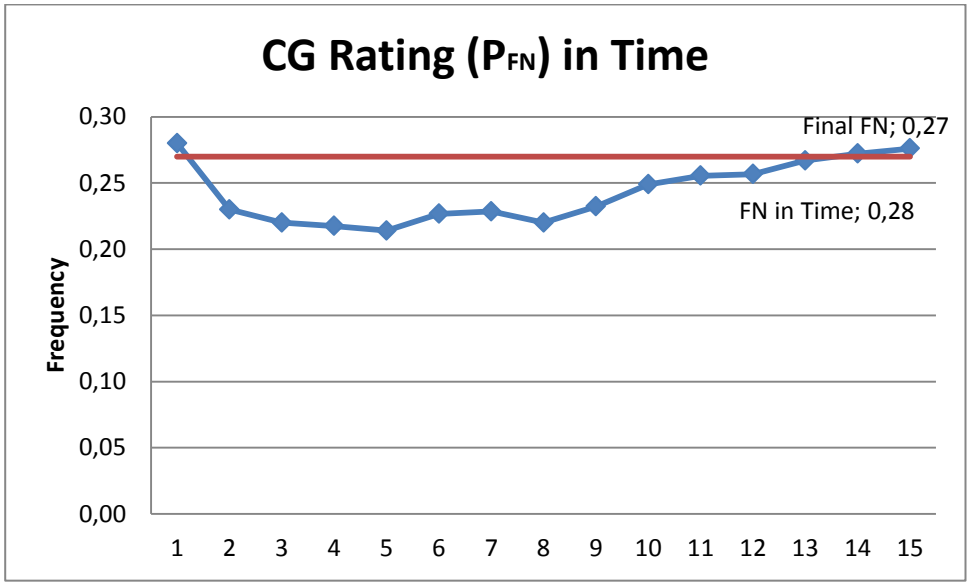


Figure 22– Observations of P_{FN} for CG rating through time

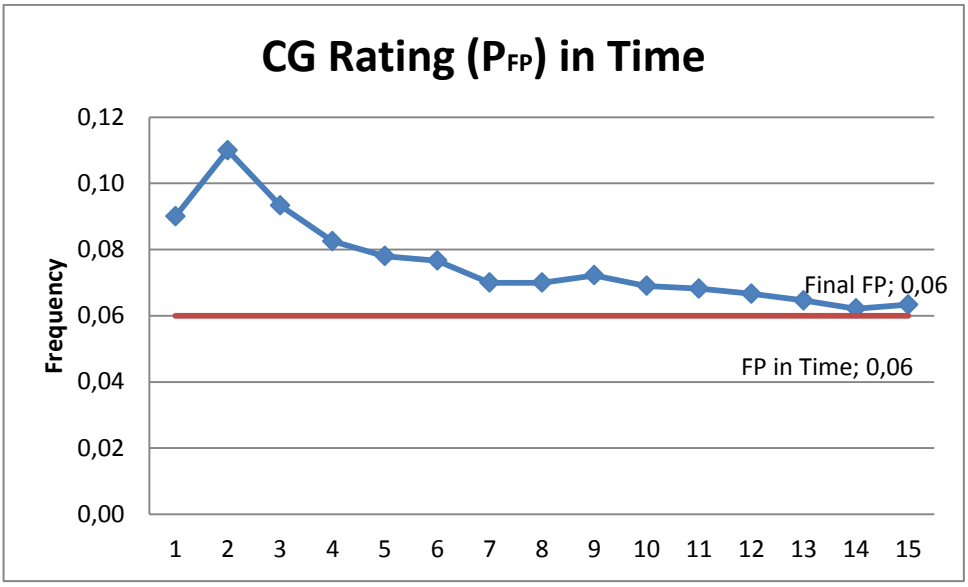


Figure 23– Observations of P_{FP} for CG rating through time

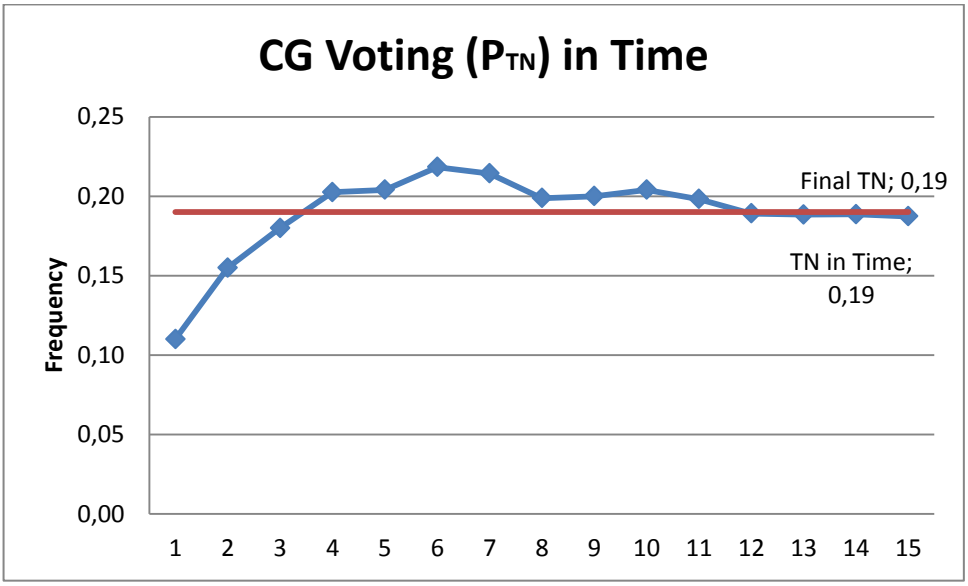


Figure 24– Observations of P_{TN} for CG voting through time

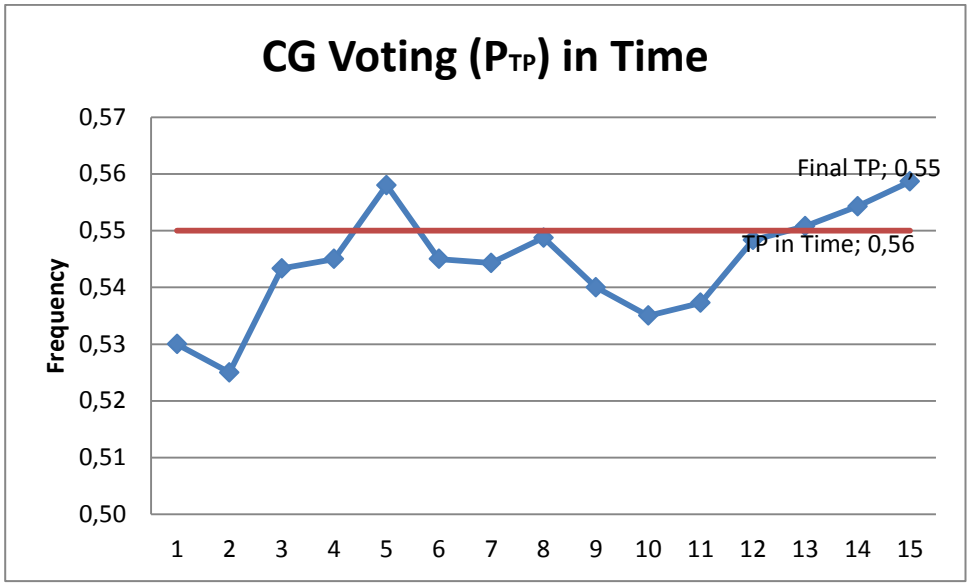


Figure 25– Observations of P_{TP} for CG voting through time

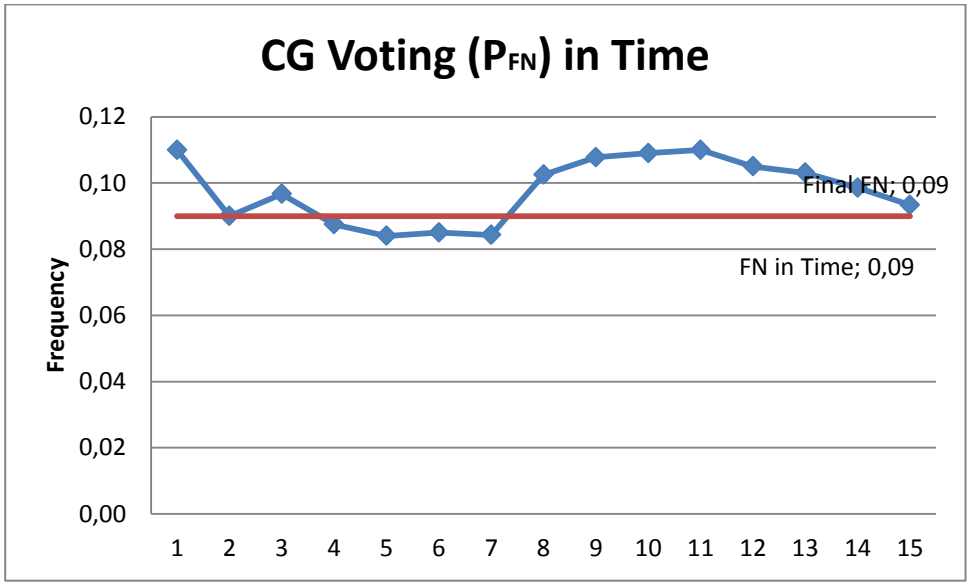


Figure 26– Observations of P_{FN} for CG voting through time

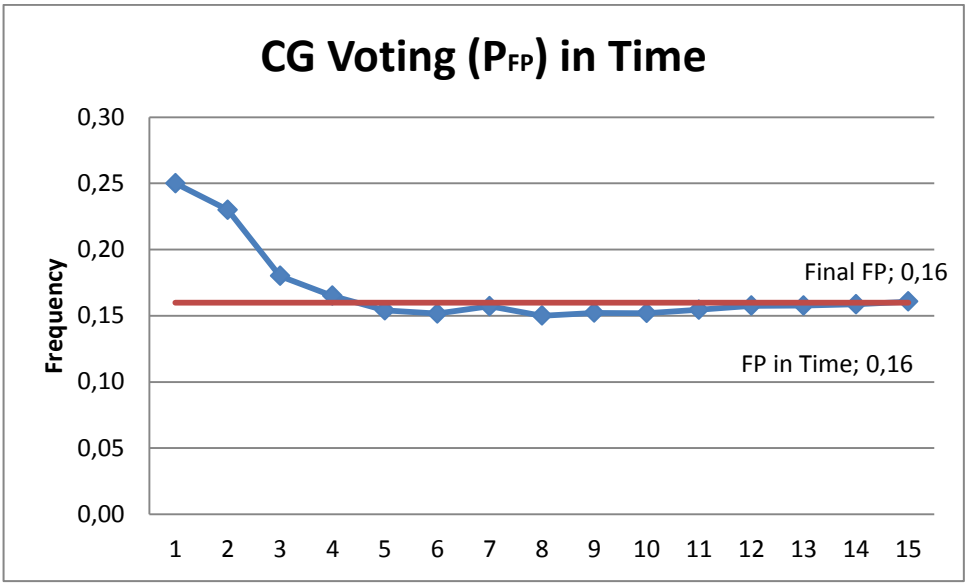


Figure 27– Observations of P_{FP} for CG voting through time

Observations of quality assurance outcome frequencies

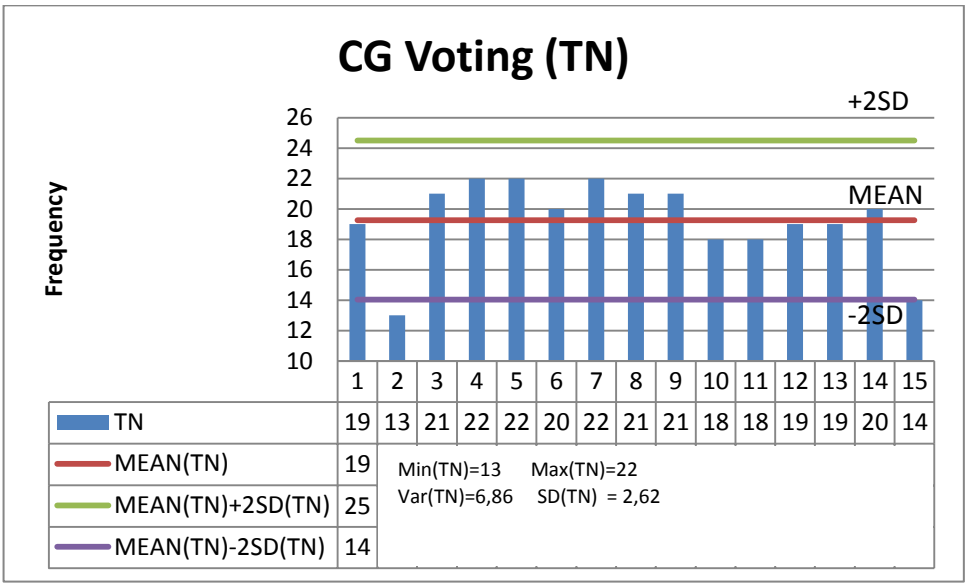


Figure 28 – Observed frequency of TN occurrences in CG voting

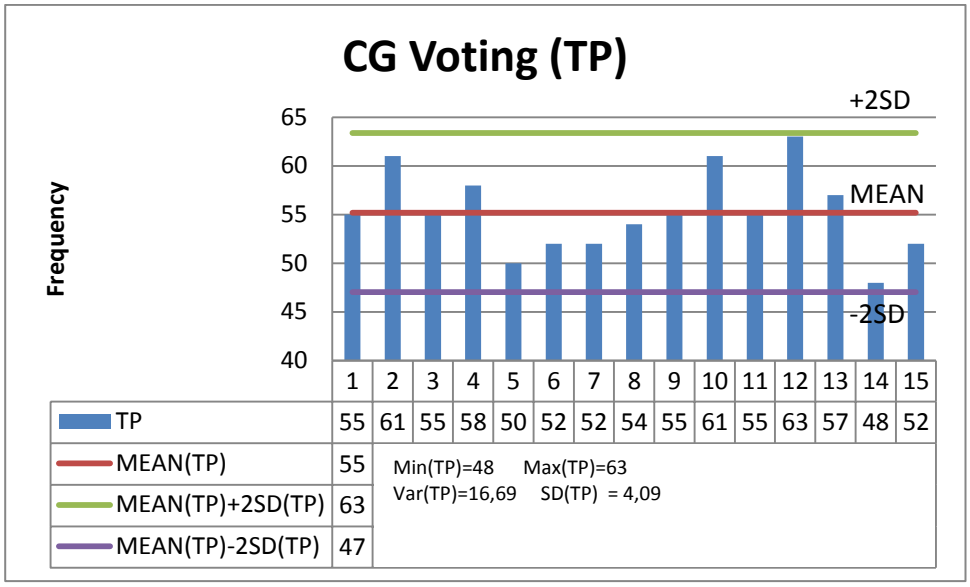


Figure 29 – Observed frequency of TP occurrences in CG voting

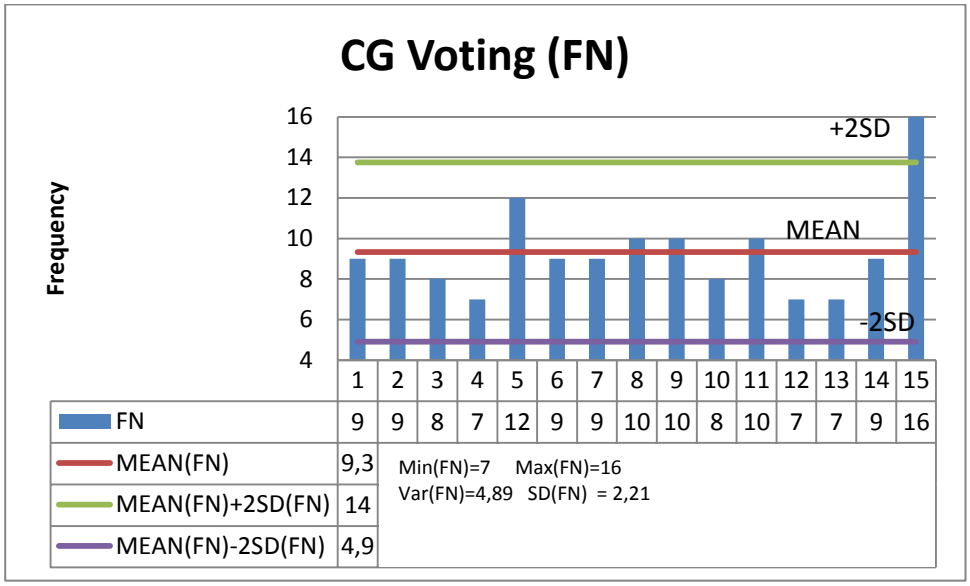


Figure 30– Observed frequency of FN occurrences in CG voting

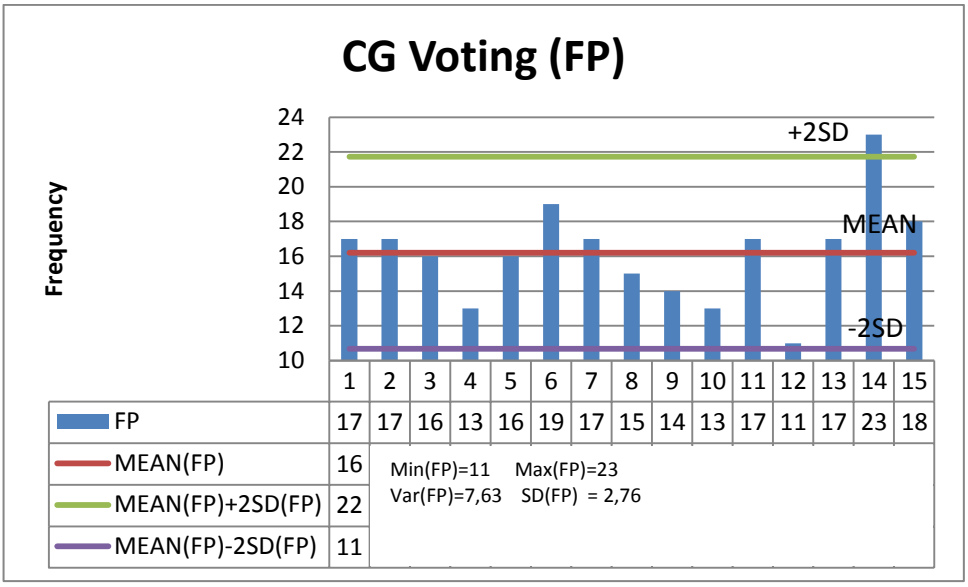


Figure 31– Observed frequency of FP occurrences in CG voting

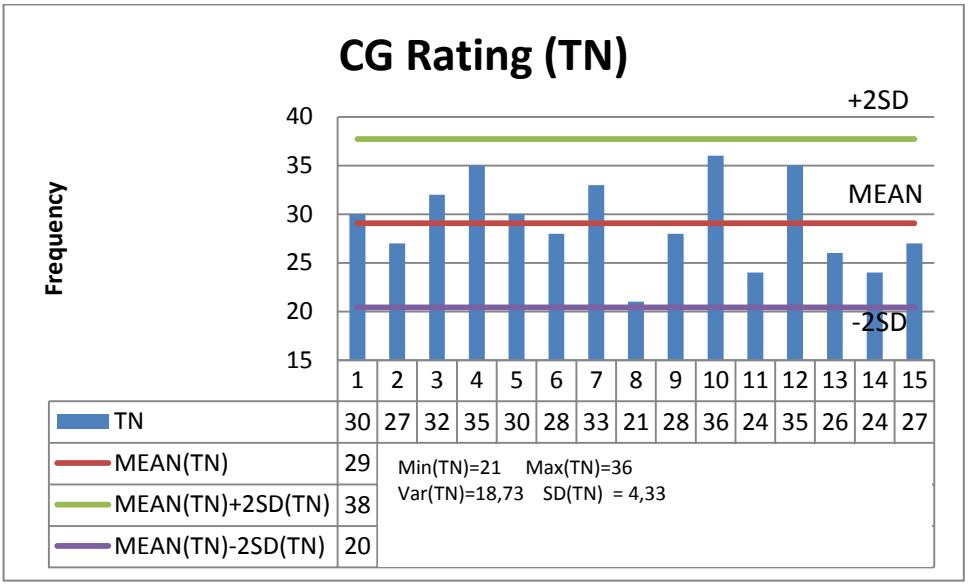


Figure 32– Observed frequency of TN occurrences in CG rating

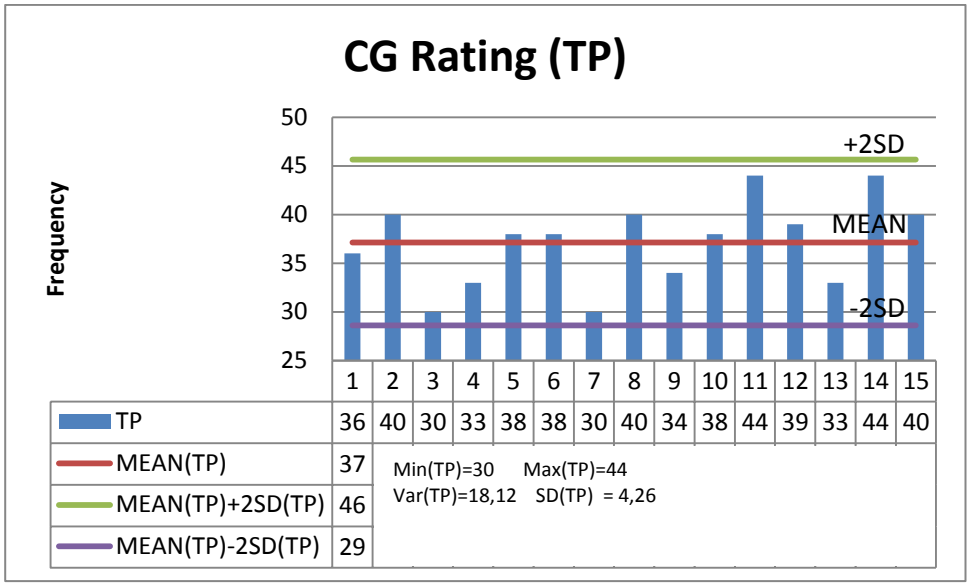


Figure 33– Observed frequency of TP occurrences in CG rating

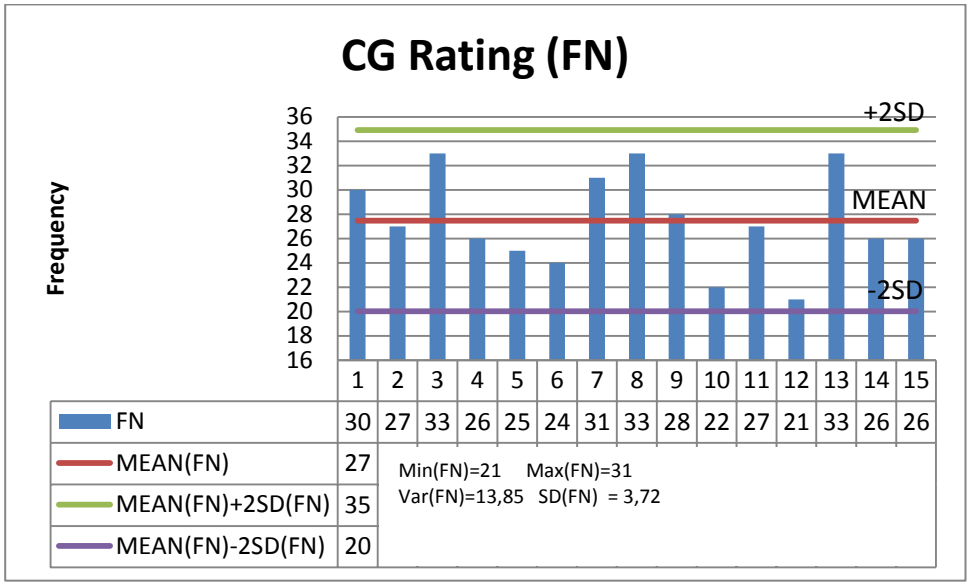


Figure 34 - Observed frequency of FN occurrences in CG rating

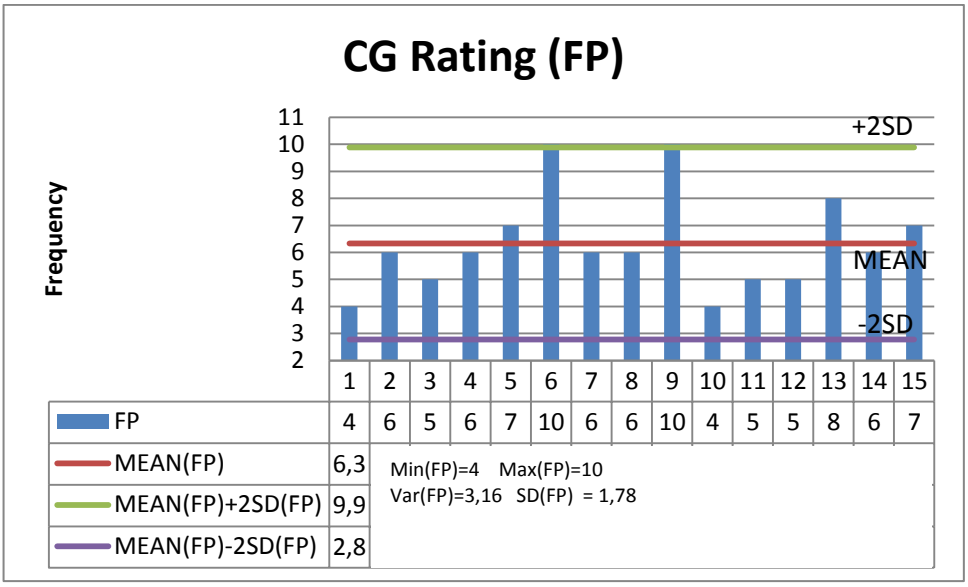


Figure 35 - Observed frequency of FP occurrences in CG rating

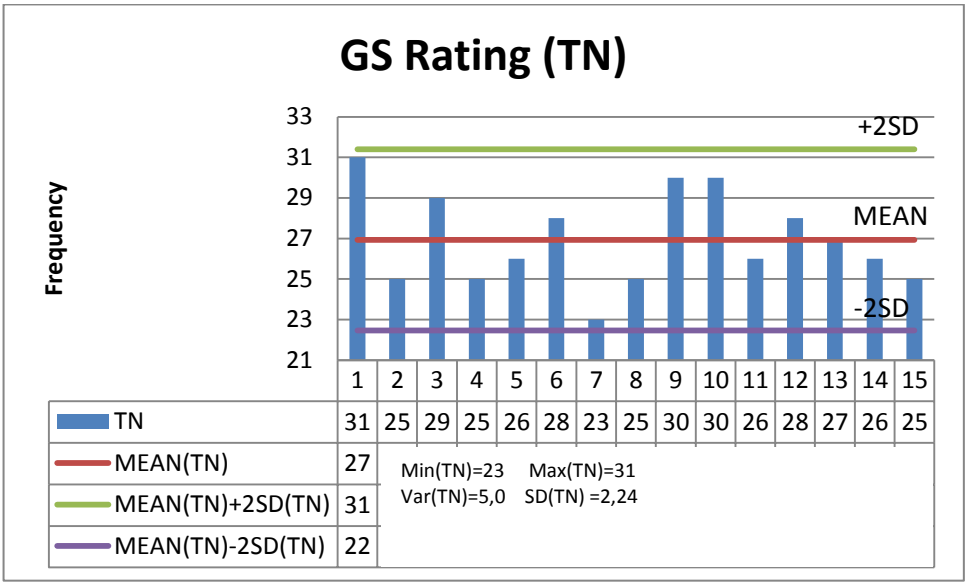


Figure 36- Observed frequency of TN occurrences in GS rating

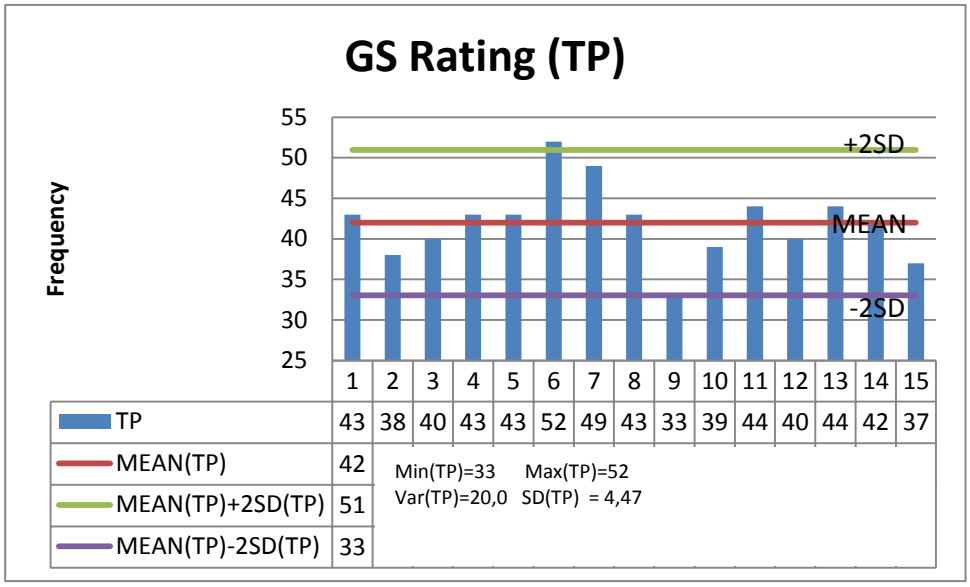


Figure 37- Observed frequency of TP occurrences in GS rating

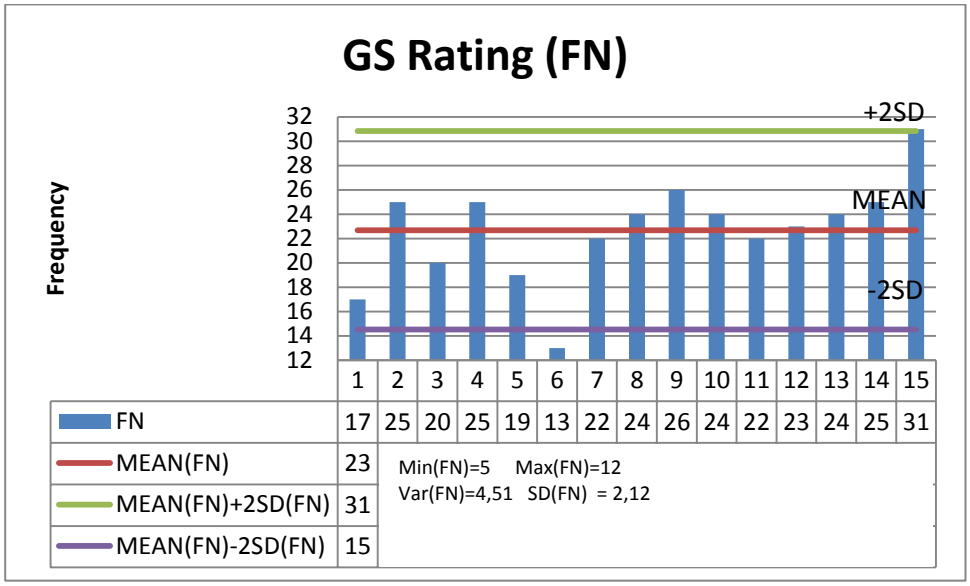


Figure 38- Observed frequency of FN occurrences in GS rating

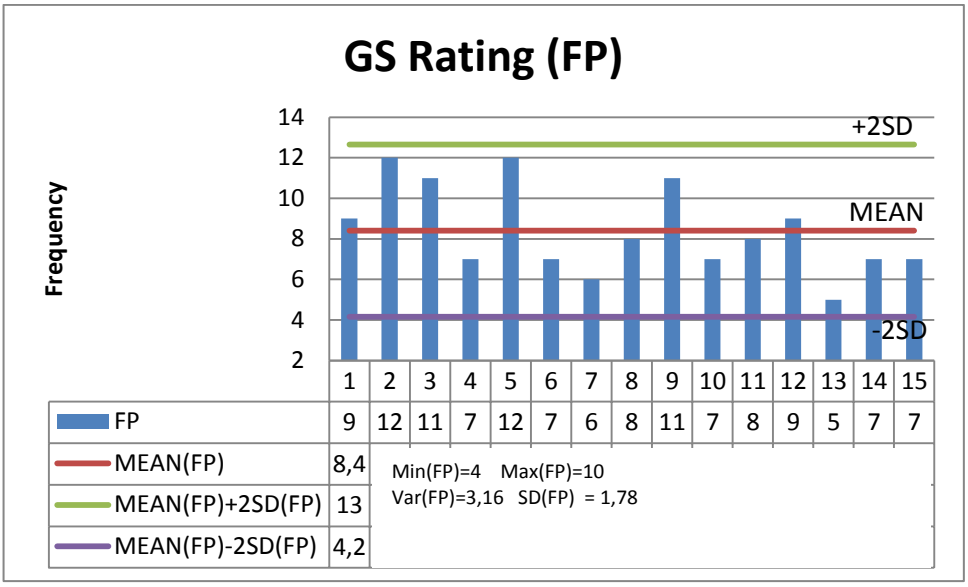


Figure 39- Observed frequency of FP occurrences in GS rating

APPENDIX B – SUPPLEMENTARY MATERIAL FOR ACTION RESEARCH 2

Observations of quality assurance outcomes through time

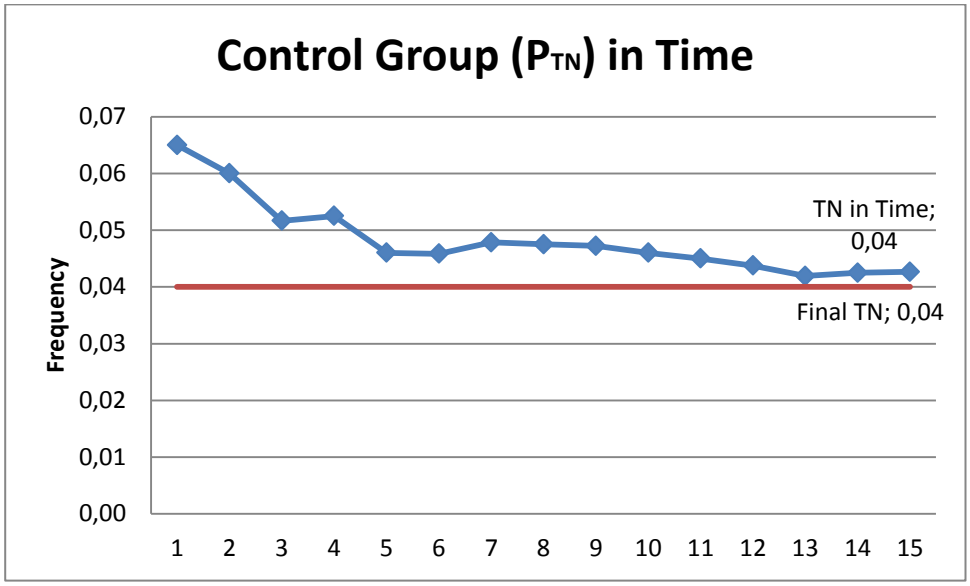


Figure 40– Observations of P_{TN} for Control Group through time

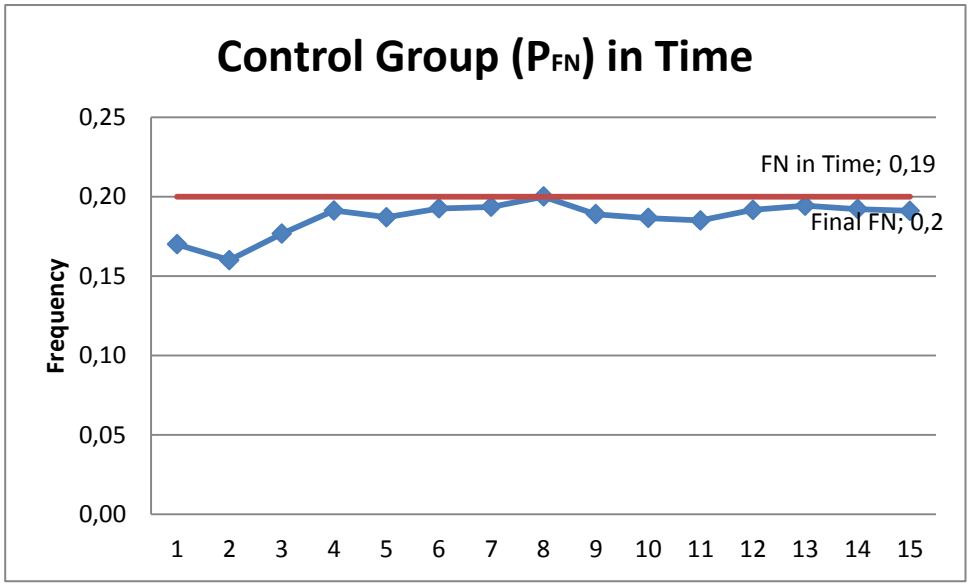


Figure 41– Observations of P_{FN} for Control Group through time

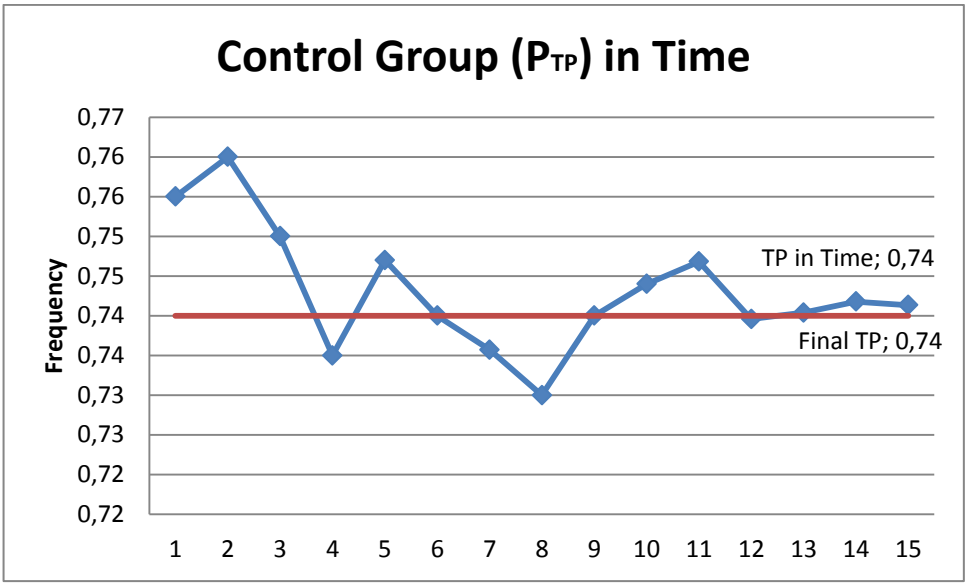


Figure 42– Observations of P_{TP} for Control Group through time

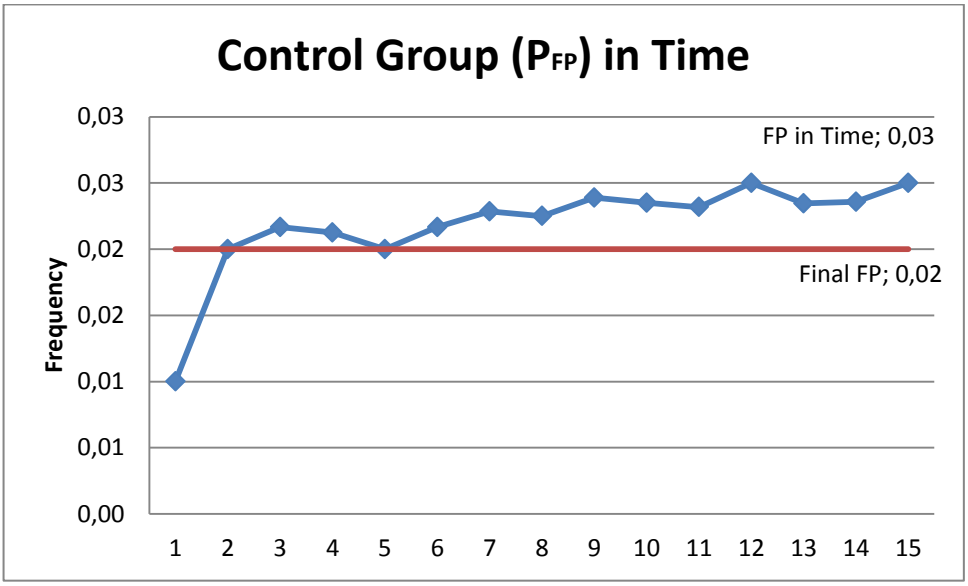


Figure 43– Observations of P_{FP} for Control Group through time

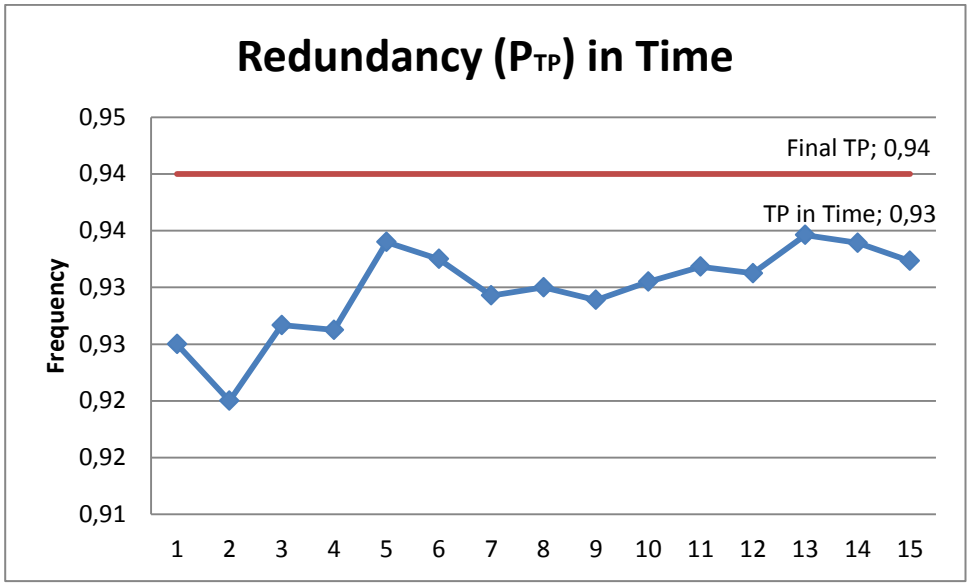


Figure 44– Observations of P_{TP} for Redundancy through time

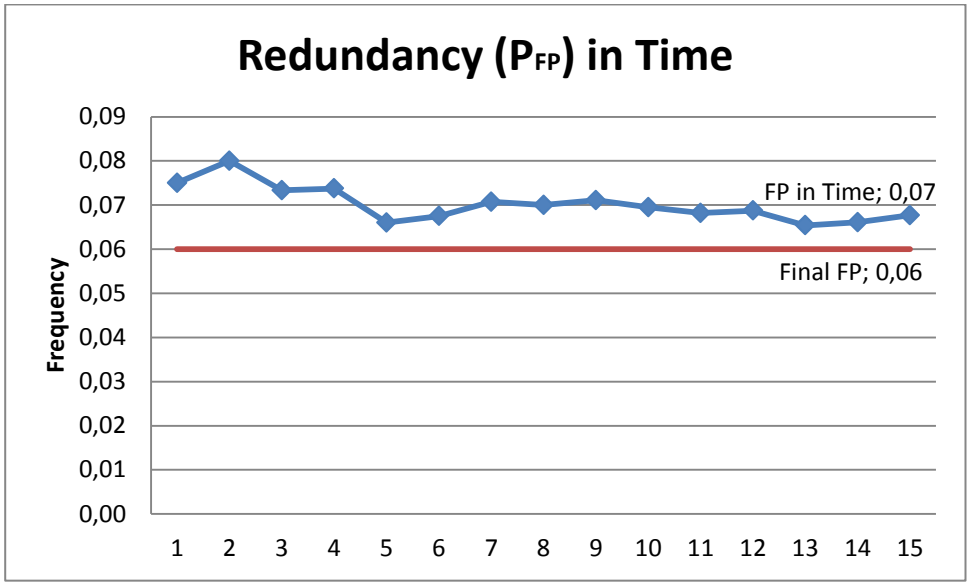


Figure 45– Observations of P_{FP} for Redundancy through time

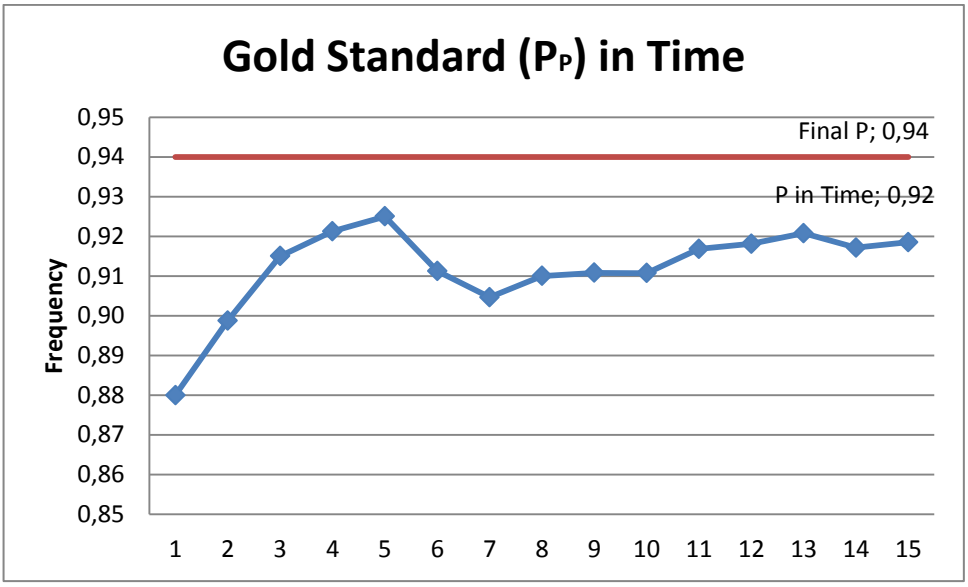


Figure 46– Observations of P_P for Gold Standard through time

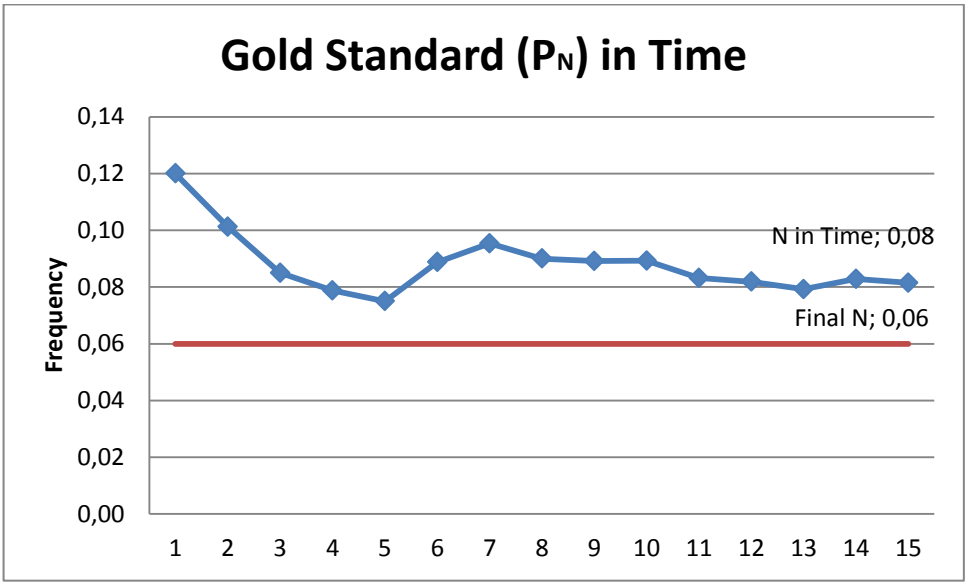


Figure 47– Observations of P_N for Gold Standard through time

Observations of quality assurance outcome frequencies

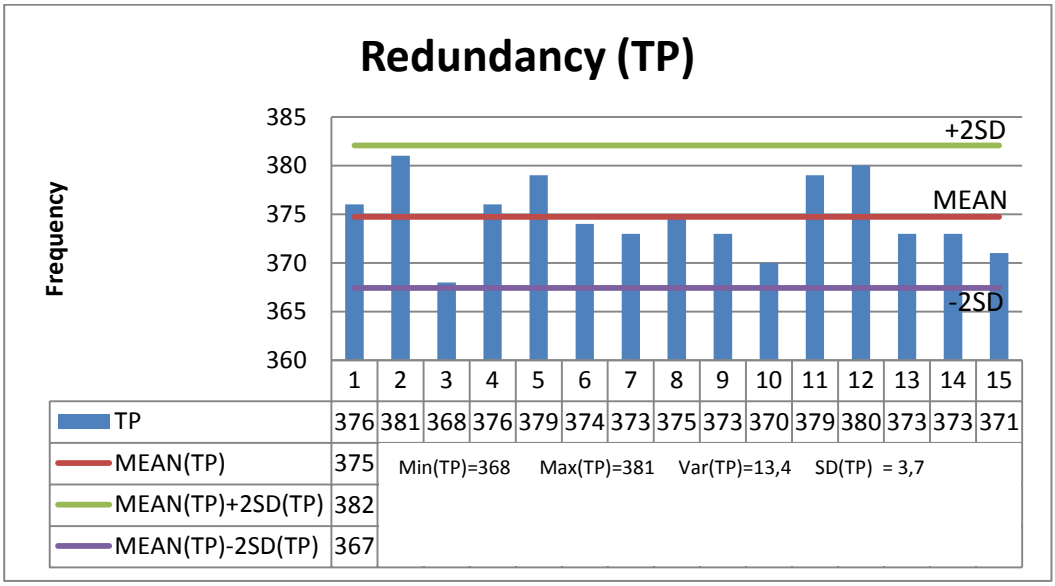


Figure 48– Observed frequency of TP occurrences in redundancy

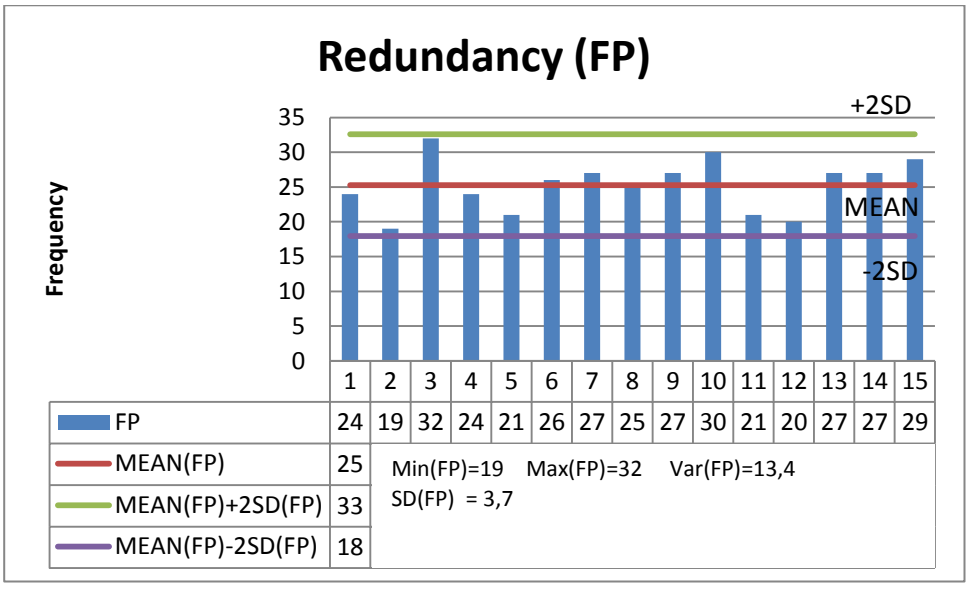


Figure 49– Observed frequency of FP occurrences in redundancy

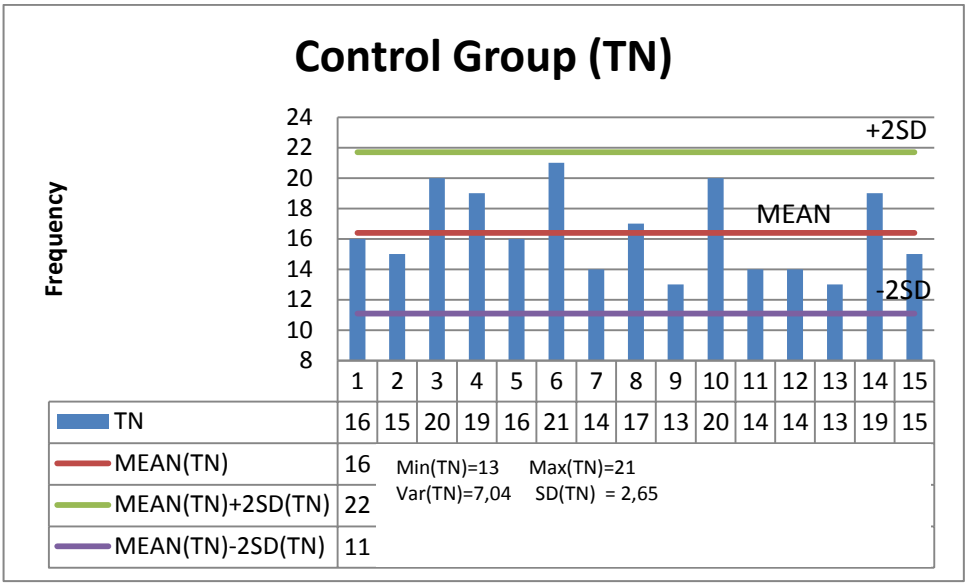


Figure 50– Observed frequency of TN occurrences in control group

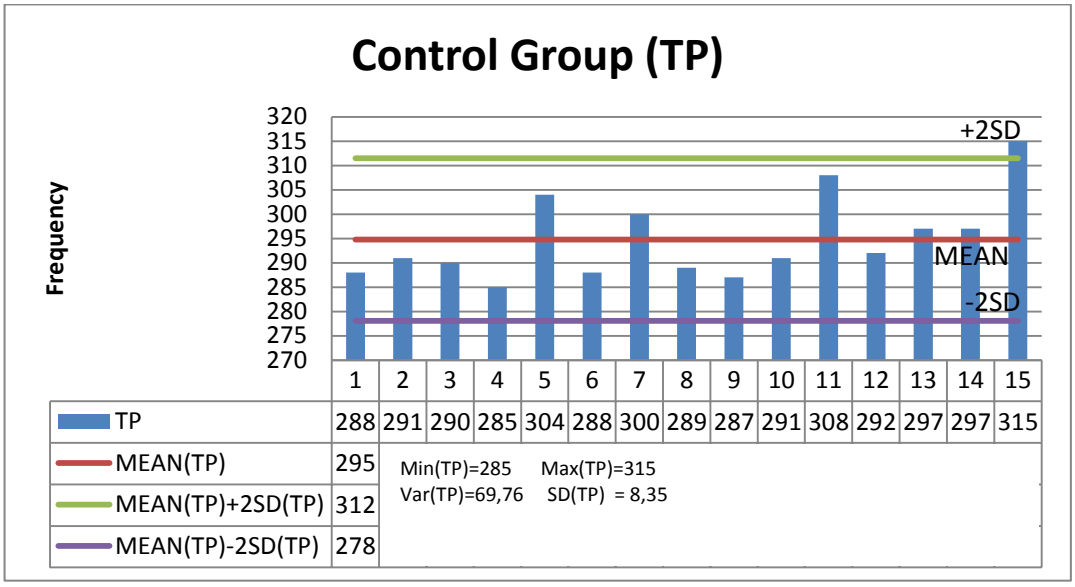


Figure 51 – Observed frequency of TP occurrences in control group

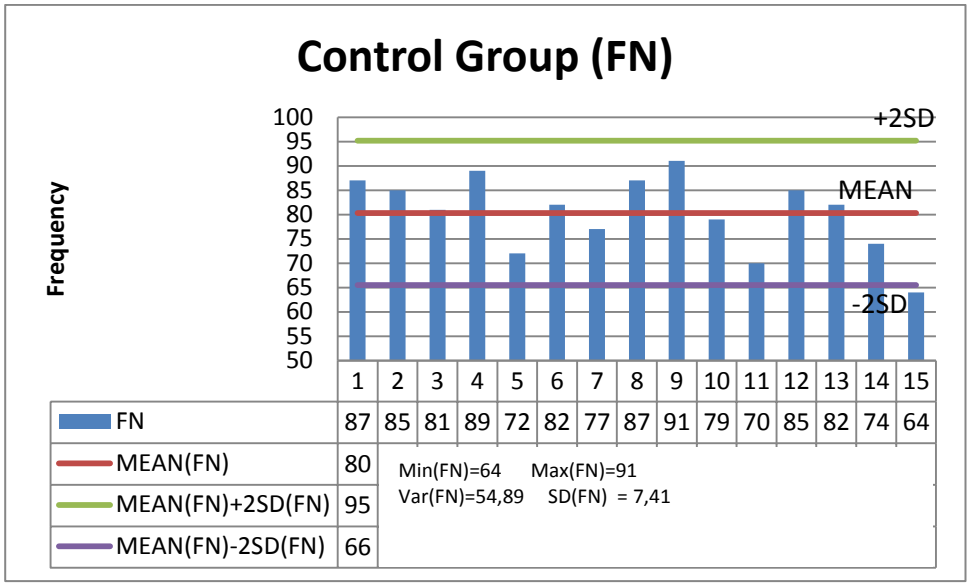


Figure 52– Observed frequency of FN occurrences in control group

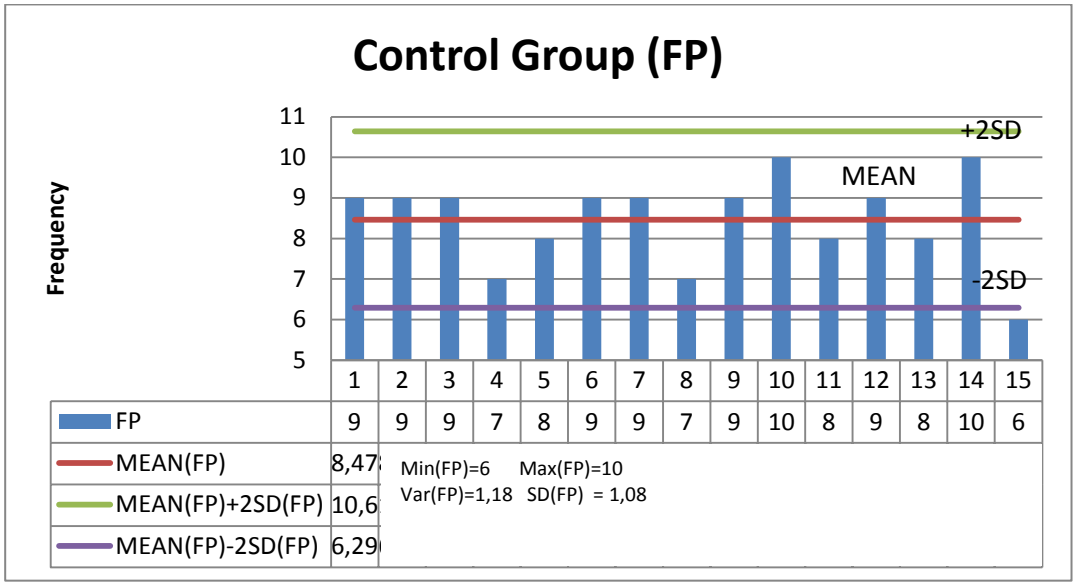


Figure 53– Observed frequency of FP occurrences in control group

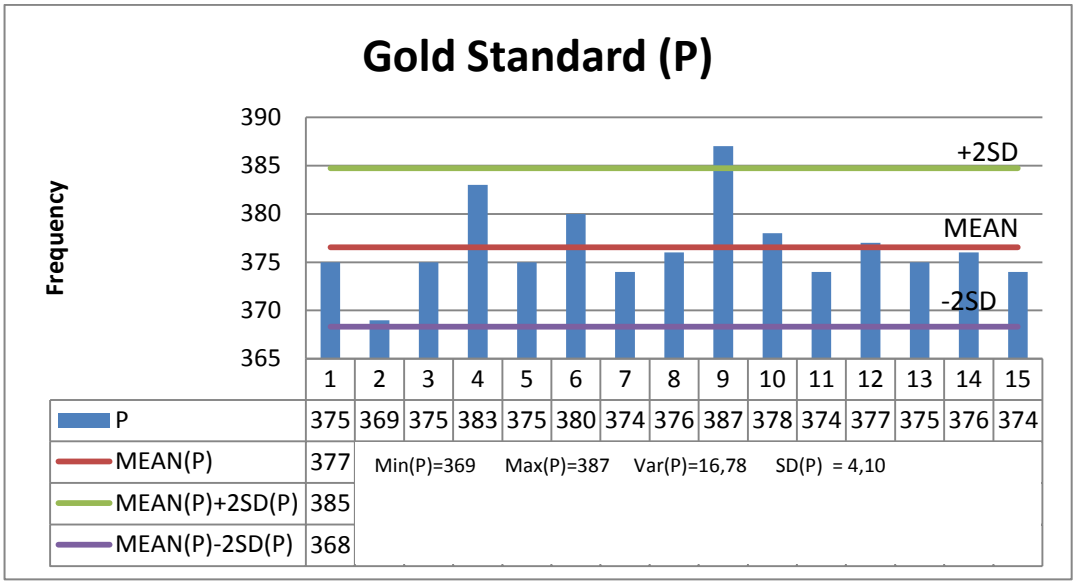


Figure 54– Observed frequency of P occurrences in gold standard

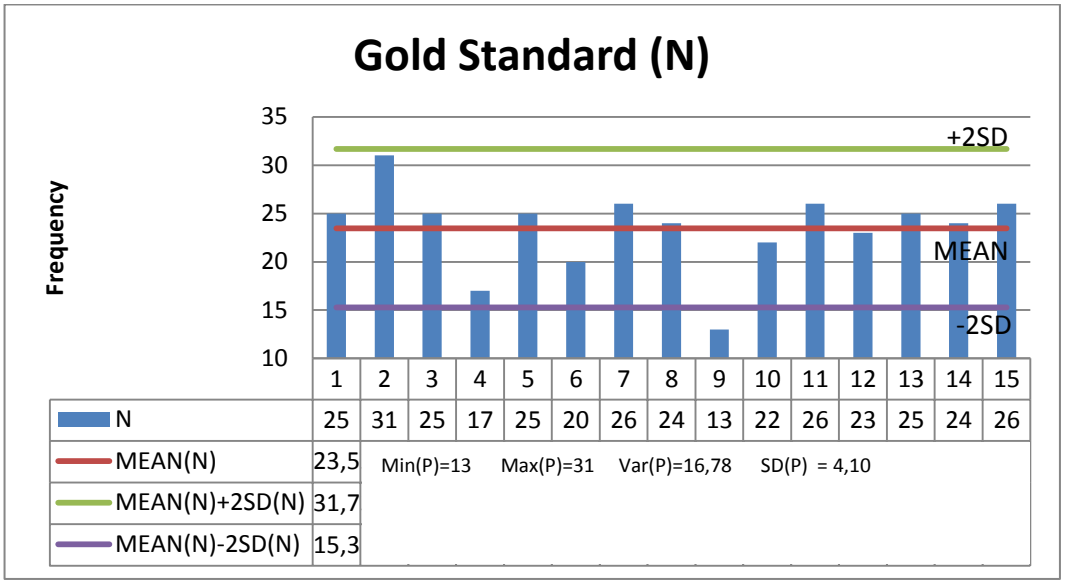


Figure 55– Observed frequency of N occurrences in control group

APPENDIX C – SUPPLEMENTARY MATERIAL FOR ACTION RESEARCH 3

Observations of quality assurance outcomes through time

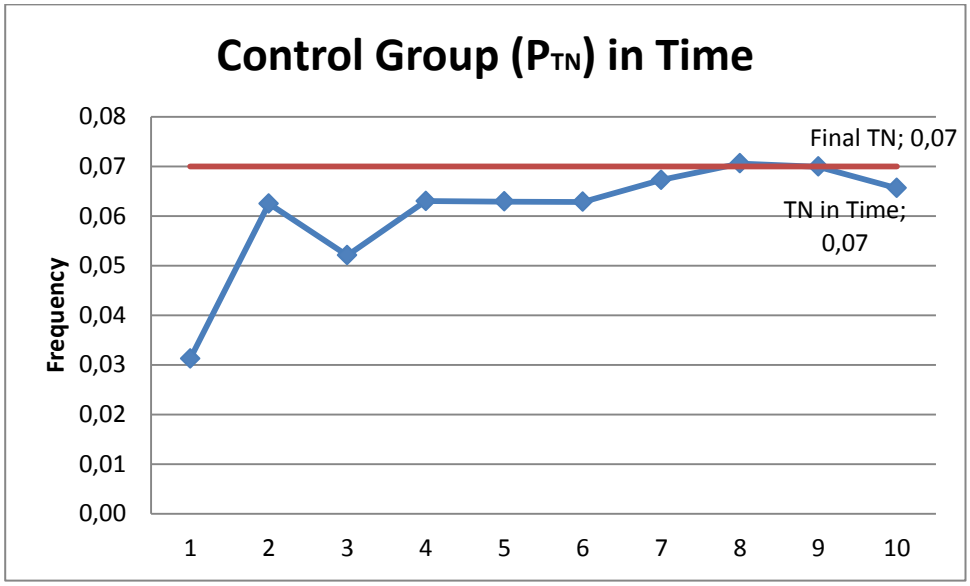


Figure 56– Observations of P_{TN} for Control Group through time

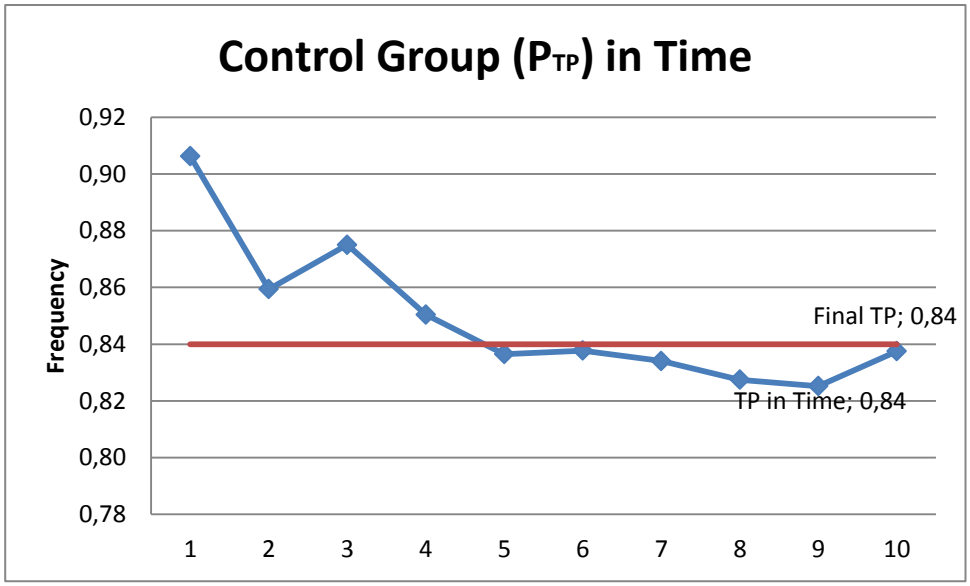


Figure 57– Observations of P_{TP} for Control Group through time

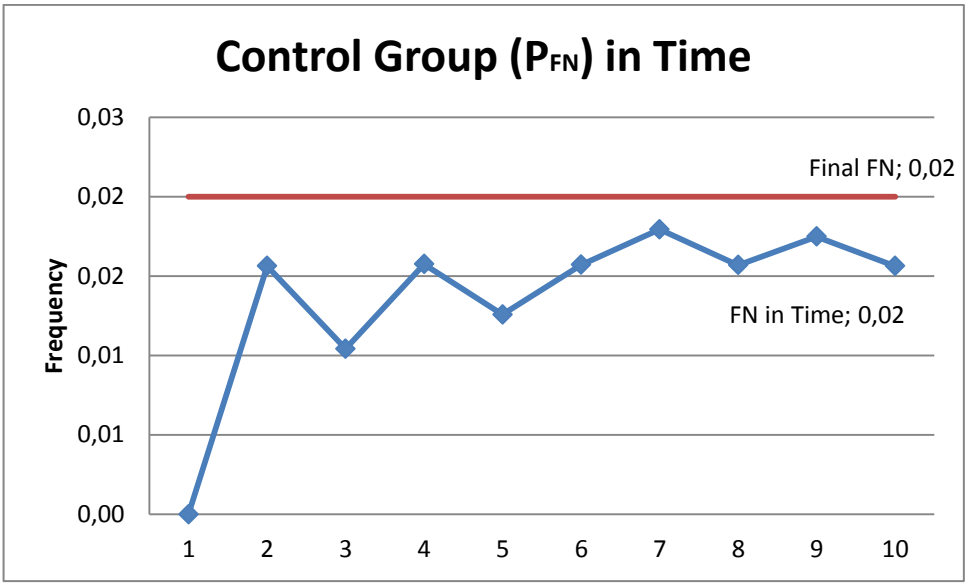


Figure 58– Observations of P_{FN} for Control Group through time

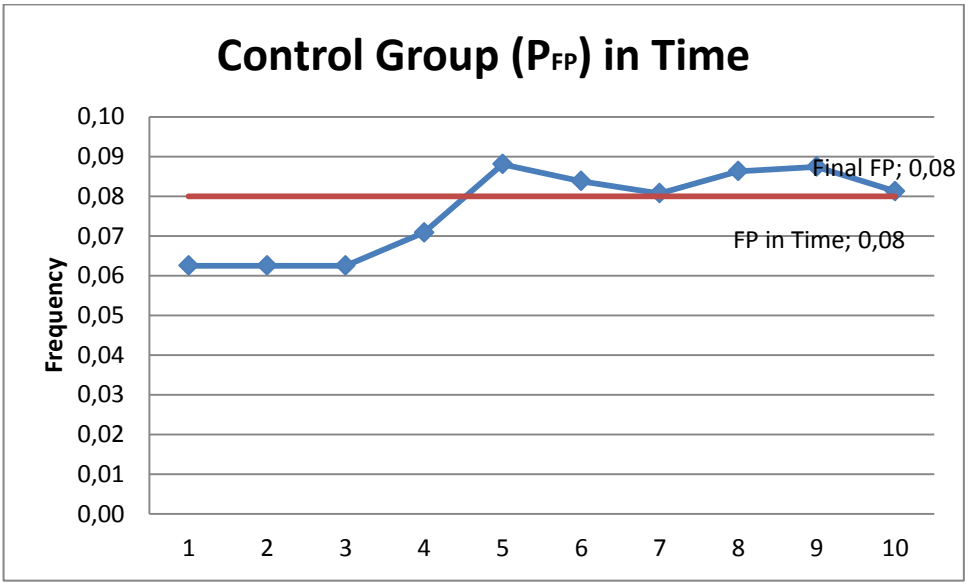


Figure 59– Observations of P_{FP} for Control Group through time

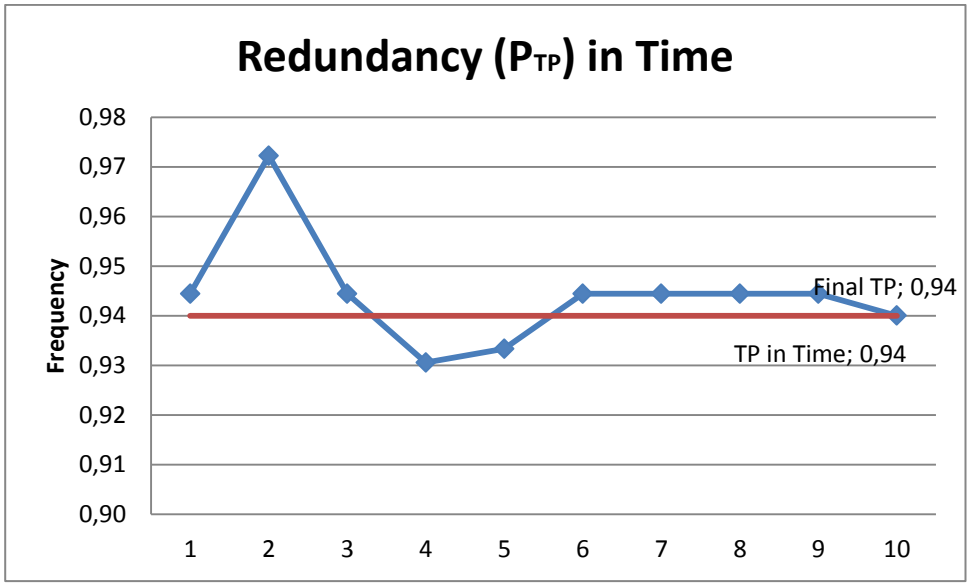


Figure 60– Observations of P_{TP} for Redundancy through time

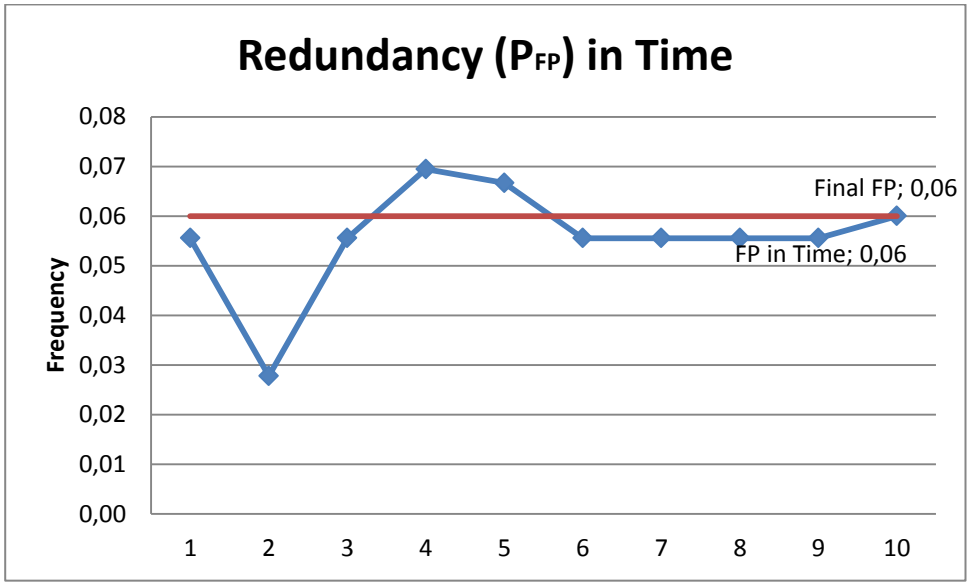


Figure 61– Observations of P_{FP} for Redundancy through time

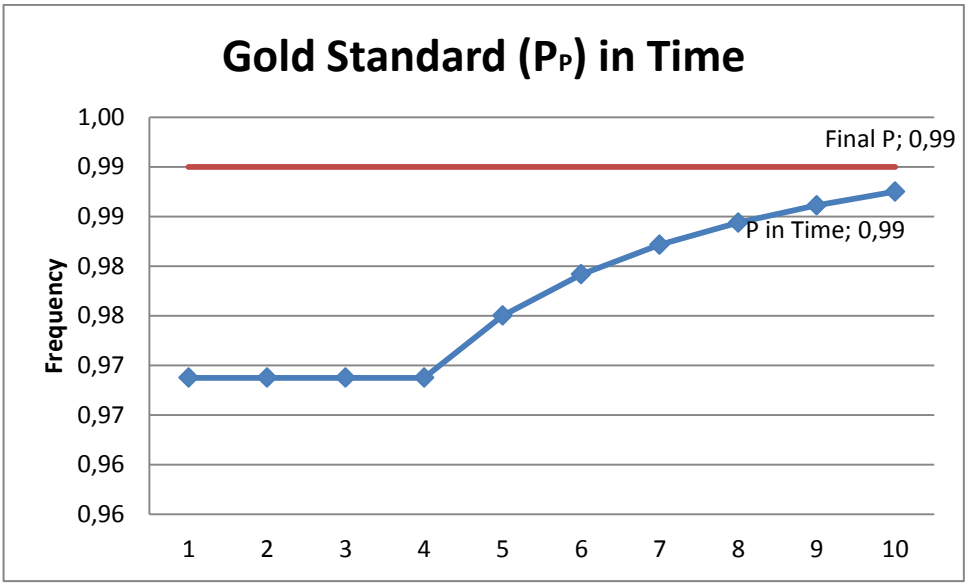


Figure 62– Observations of P_P for Gold Standard through time

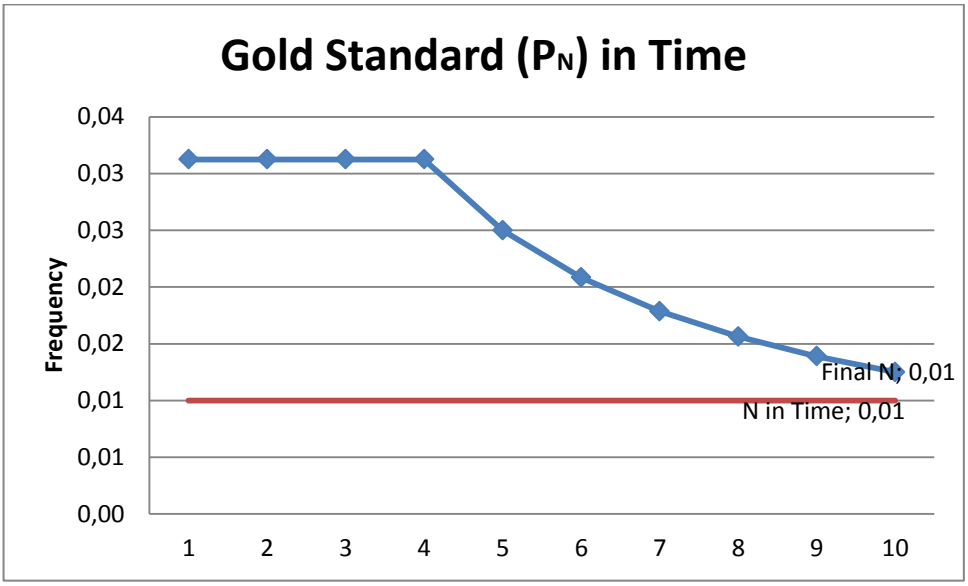


Figure 63– Observations of P_N for Gold Standard through time

Observations of quality assurance outcome frequencies

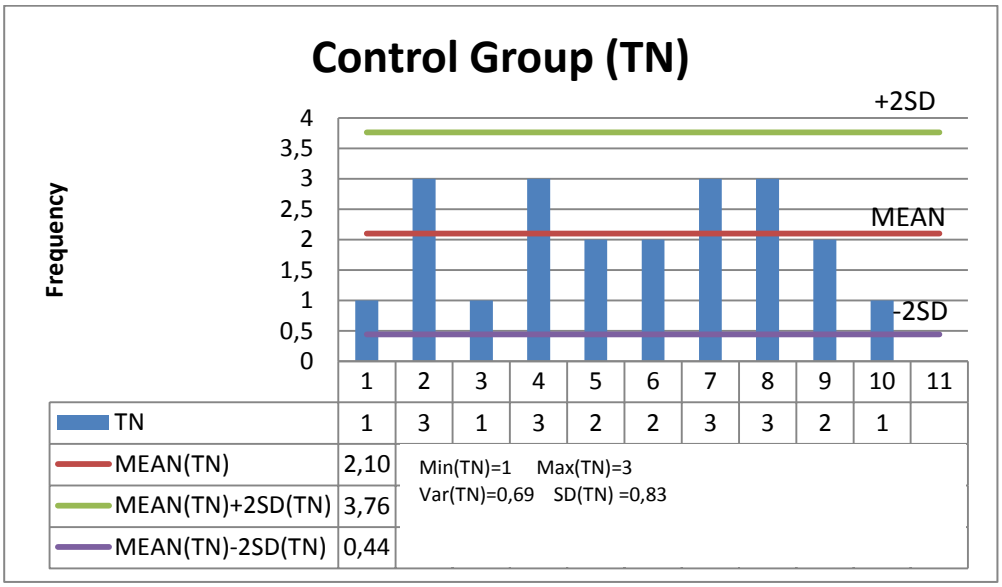


Figure 64– Observed frequency of TN occurrences in control group

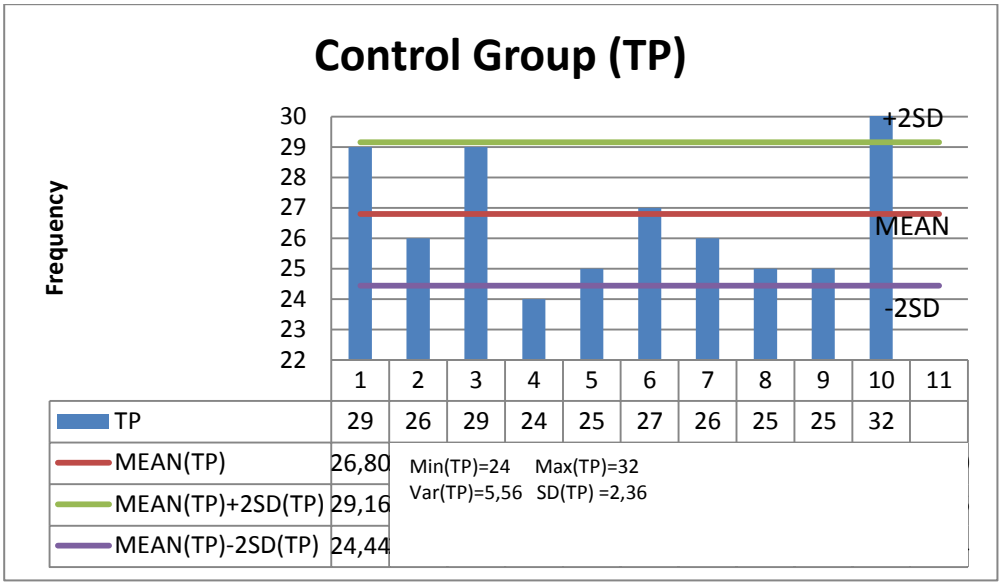


Figure 65– Observed frequency of TP occurrences in control group

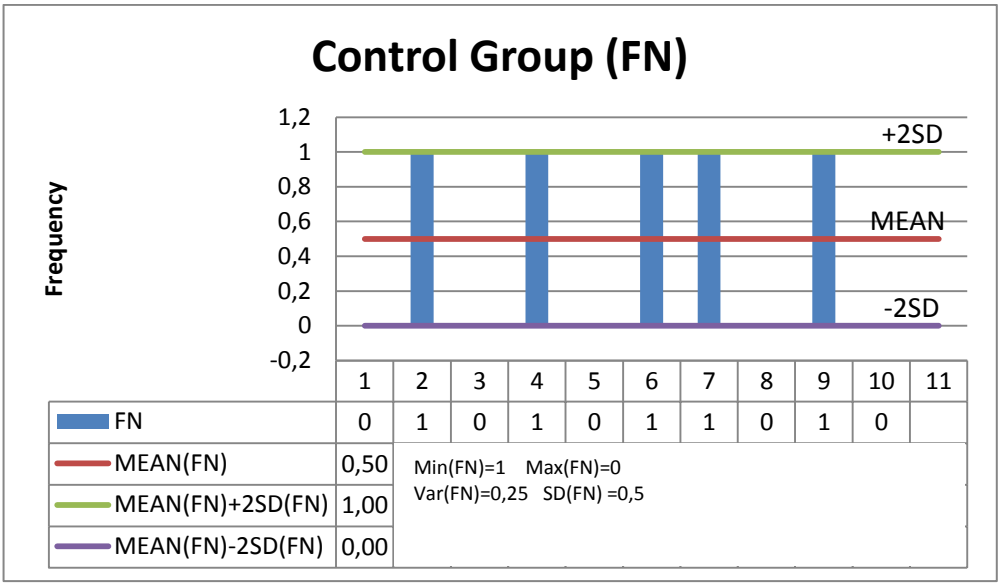


Figure 66– Observed frequency of FN occurrences in control group

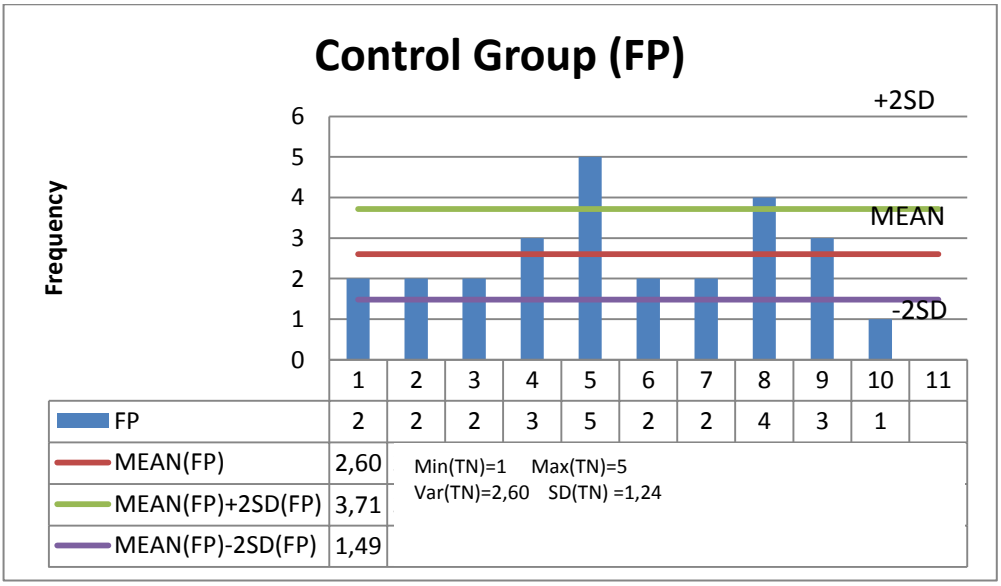


Figure 67– Observed frequency of FP occurrences in control group

APPENDIX D – INTERVIEW SCRIPTS

Table 20- Interview questions and answers

Stakeholder	Question and Answer
	1) Can you describe your role in this organization and project?
Vice President of METU	I am the vice president of METU. My responsibility areas cover university research projects and IT projects coordinated by the computer center.
Director of Computer Center	I am the acting director of the university computer center.
Project Manager	I am the manager of the project Integrated Information Systems, in which we utilized crowdsourcing.
Software Engineer 1	I work as a software engineer in the project Integrated Information Systems. I was responsible of developing the software solution for crowdsourcing the phonebook registry update tasks.
Software Engineer 2	I work as a software developer in Integrated Information Systems project. Generally my work focuses on establishing the digital object repository.
	2) Were you aware of the crowdsourcing concept before this project?
Vice President of METU	Yes, I was aware.
Director of Computer Center	Yes, even though I had no experience with crowdsourcing, I was aware of the concept.
Project Manager	I was aware.
Software Engineer 1	I had limited knowledge about crowdsourcing. Though I have never researched on the subject.
Software Engineer 2	I knew the basics of crowdsourcing but I have never been in a project using crowdsourcing.
	3) Have you ever participated in crowdsourcing before this project?
Vice President of METU	We never used crowdsourcing before as a part of a university project.
Director of Computer Center	No. As far as I know, this is the first application of crowdsourcing in public universities of Turkey.
Project Manager	No. I have never used crowdsourcing before this project.
Software Engineer 1	No. This was my first experience of crowdsourcing usage.
Software	No.

Engineer 2	
	4) In your opinion, what are the advantages of using crowdsourcing over traditional methods?
Vice President of METU	We achieved time and cost efficiencies. Additionally the accuracy of the results was better than we would have expected in a traditional data cleaning project.
Director of Computer Center	We applied crowdsourcing in solutions of two different types of problems. One of them required information which is only available in target groups. So by using crowdsourcing we were able to extract that information. Generally speaking, crowdsourcing is faster and costs less.
Project Manager	We used crowdsourcing in two different work packages. First one was about updating the personnel contact information of a large organization. Second one was about data comparison and cleaning. In the first work package advantages are as follows. We were not able to assign this job to a specific personnel or an administrative unit. Crowdsourcing draw attention and lead to fast completion of the job. Actually, this was not a job which can be done by one person. So it enabled us to solve the problem. In the second application, the job required significant effort and again we were not able to find a resource to assign that job. In my opinion we have a problem with the scalability of our workforce. By using crowdsourcing we could complete the task faster and we could achieve results with satisfactory quality. Traditional jobs have larger overheads; finding the resource, assigning the job and controlling the outputs… On the other hand, after we built the crowdsourcing system, the rest was easy. We were able to estimate how long the job was going to take and how much it was going to cost.
Software Engineer 1	Work is assigned to the worker in traditional settings. However crowdsourcing somehow gains attention and individuals participated. It acted like a motivating factor in doing the job collectively. In this project crowdsourcing enabled us to capture the knowledge which resides in the groups.
Software Engineer 2	Before deciding to use crowdsourcing in the project we analyzed other ways to perform this job. We estimated that the job would require 10 part time research assistants to spend 3 months in order to complete the job. Therefore, crowdsourcing had advantages in terms of cost and time. I can also say that the end result turned out to have very good quality.
	5) What are the disadvantages of crowdsourcing when compared to traditional methods?
Vice President of METU	Crowdsourcing may have a limited applicability.
Director of Computer Center	A major problem is reliability. We could not trust the worker outputs. So we assigned the tasks to many users at the same time. We used additional quality control methods which increase the quality costs

	significantly.
Project Manager	Establishing the crowdsourcing system requires additional work. Furthermore, the job must be transformed into a simpler version so that anyone can do it. But crowdsourcing fitted well to the problems we face. I had my doubts that the job would finish on time and within our budget estimations. Surprisingly, the job finished complying with our estimations.
Software Engineer 1	As a developer I had difficulty in grasping the underlying mechanisms of crowdsourcing, namely, quality assurance and work aggregation techniques…
Software Engineer 2	We had difficulties of persuading the stakeholders about the effectiveness of crowdsourcing. Since the crowd workers are not our personnel we have limited control over them. Thus, we cannot provide training or directly control their work. The people who were going to use the final product produced by the crowd had questions about the quality of the end result.
	6) Would you use crowdsourcing again in future projects of your organization?
Vice President of METU	We would. However most of our data consists of words and sentences in Turkish. Therefore, it would be difficult to find a Turkish speaking crowd. Nevertheless, if we encounter problems in which crowdsourcing is applicable, we would use crowdsourcing.
Director of Computer Center	Yes. Now we have experience with crowdsourcing and we saw that it is useful and feasible. The only problem is to achieve sustainability. I do not think there will be many problems which can be solved with crowdsourcing in the future. For example, we used crowdsourcing for data cleaning. We must focus on building systems which do not cause problems with the data in the first place, rather than focusing on solving these problems when they occur.
Project Manager	I will not have any doubts to use crowdsourcing in the future, since now I know that we can estimate and plan crowdsourcing. Nevertheless, it depends on the situation. The problem must be solvable by crowdsourcing.
Software Engineer 1	Yes. Surely.
Software Engineer 2	Yes.
	7) Do you think using cost of quality estimation methods enabled crowdsourcing to be a recognized way of problem solving for your organization?
Vice President of METU	I requested weekly updates about the crowdsourcing progress. I confirm that estimations of the project team were accurate.
Director of Computer Center	(Director's reply to Question 6 also covers this answer.)

Project Manager	Yes. I definitely think that crowdsourcing can be a valid way of problem solving in our organization, for the types of problems which can be solved with crowdsourcing.
Software Engineer 1	Yes. I completely trust that crowdsourcing can be recognized as a problem solving method in the organization.
Software Engineer 2	I think so.
	8) Would you use the same cost of quality estimation methods?
Vice President of METU	Yes.
Director of Computer Center	We did not experience any problems with effort estimation in these projects. Plans were accurate. We would use the same technique again.
Project Manager	Definitely.
Software Engineer 1	I think this estimation method is beneficial to the managers.
Software Engineer 2	Surely I would.

CURRICULUM VITAE

PERSONAL INFORMATION

Deniz İren
Ankara, 1981
deniziren@gmail.com
www.deniziren.com

EDUCATION

2008 – 2014 Middle East Technical University, (PhD)
Informatics Institute, Information Systems
2005 – 2008 Middle East Technical University, (MSc)
Informatics Institute, Software Management
1999 – 2004 Başkent University (BS)
Faculty of Engineering, Department of Computer Engineering

WORK EXPERIENCE

2012 – 2014 Middle East Technical University, Computer Center, Program Manager
2010 – 2012 Middle East Technical University, Computer Center, Research Assistant
2008 – 2009 Freelance – Project Management and Software Quality Process Consulting
2007 – 2008 Mikrokom Corp. – Software Project Coordinator
2005 – 2007 Mikrokom Corp. – Software Engineer

PUBLICATIONS

- Iren, D., & Bilgen, S. (2014). Cost of Quality in Crowdsourcing. Journal of Human Computation (In review).
- Iren, D., & Bilgen, S. (2014). Cost Models of Quality Assurance in Crowdsourcing. In IEEE (In publishing).
- Iren, D., Kul, G., & Bilgen, S. (2014). Utilization of synergetic human-machine clouds: a big data cleaning case. In Proceedings of the 1st International Workshop on CrowdSourcing in Software Engineering (pp. 15–18).

- Chouseinoglou, O., Iren, D., Karagöz, N. A., & Bilgen, S. (2013). AiOLoS: A model for assessing organizational learning in software development organizations. Information and Software Technology, 55(11), 1904–1924.
- Aysolmaz, B., Iren, D., & Demirörs, O. (2013). An Effort Prediction Model based on BPM Measures for Process Automation. In Enterprise, Business-Process and Information Systems Modeling (pp. 154–167). Springer Berlin Heidelberg.
- Iren, D., & Bilgen, S. (2013). Validating cost of quality models in subjective non-deterministic microtask crowdsourcing. URL: http://expertjudgment.com/publications/METU_II_TR_2013_22.pdf
- Iren, D., & Bilgen, S. (2013). Cost models of crowdsourcing quality assurance mechanisms. URL: http://expertjudgment.com/publications/METU_II_TR_2013_21.pdf
- Kul, G., & Iren, Y. D. (2013). Wireless network forensics: sources of digital evidence. Global Journal on Technology, 1.
- Chouseinoglou, O., Karagöz, N. A., Iren, D., Özen, G., & Bilgen, S. (2013). AiOLoS: Yazılım Geliştiren Organizasyonlarda Örgütsel Öğrenmeyi Değerlendirme Modeli. In UYMS.
- Iren, D. (2013). ODTÜ Bütünleşik Bilgi Sistemi Projesi Deneyimleri. In Akademik Bilişim. URL: http://ab.org.tr/ab13/bildiri/298.odt
- Iren, D., & Bilgen, S. (2012). Methodology for Managing Crowdsourcing in Organizational Projects. In Modeling and Analysis of Novel Mechanisms in Future Internet Applications. Würzburg.
- Iren, D., & Dalci, M. (2011). Çevik Yazılım Geliştirme Süreçlerinde Kullanılabilirlik Çalışmalarının Yeri. In UYMS.
- Sengün, B., Iren, D., Kasacı, D., Ocak, N., Karataş, R., & Yalçın, Y. (2011). ODTÜ Bütünleşik Bilgi Sistemi Servis Tabanlı Süreç Yönetim Platformu. In UYMS.

RESEARCH INTERESTS

Crowdsourcing, human computation, agile software development, human-computer interaction.

WORKING FIELDS OF INTEREST

Software engineering, software quality, organizational quality and maturity models, governance frameworks, project management office, CMMI, ITIL, Cobit, information security, knowledge management, project management, human-computer interaction.

CERTIFICATIONS

2009 - Project Management Professional (PMP); PMI

2011 - PMI-Agile Certified Professional (PMI-Agile); PMI

2011 - Professional Scrum Master I (PSM-I); Scrum.org

2008 – Certified Project Manager (PY); Ankara Project Management Association

OTHER INFORMATION

2014 – Nature Travellers' Association Member

2012 – Rock lizards (Kertenkeleler) Adventure Team Captain

2008 - PMI Member

2010 - Usability Professionals Association Founding Member

2008 - 2009 Administrative Board Member of Project Management Association

2003 - Xasiork Short Story Contest Honorable Mention Award

2000 - Ankara Photography Artists Association (AFSAD) Member

1999 - 1'st World Air Games Referee

INTERESTS

Outdoor sports, adventure racing, orienteering, rock climbing, mountaineering, mountain biking, triathlon, long distance running, photography.

BLOGS

Personal and Professional:	www.deniziren.com
Adventure Racing:	www.kertenkeleler.com
Photography:	summitproject.wordpress.com