QUALITATIVE PERFORMANCE EVALUATION OF WORKERS
ACROSS DIFFERENT CROWDSOURCING PLATFORMS

A Thesis in
Information Sciences and Technology

by

AISHWARYA TADIKONDA

© 2014 AISHWARYA TADIKONDA

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2014
The thesis of Aishwarya Tadikonda was reviewed and approved* by the following:

Dongwon Lee  
Associate Professor of Information Sciences and Technology  
Thesis Adviser

Madhu Reddy  
Associate Professor of Information Sciences and Technology

Erika Poole  
Assistant Professor of Information Sciences and Technology

Pete Forster  
Head of the Department of Information Sciences and Technology

*Signatures are on file in the Graduate School.
ABSTRACT

Crowdsourcing is a problem solving technique where macro projects typically done by a designated worker are split into micro tasks and are made available online to people called workers. Rewards are offered to the workers to complete tasks either in the form of cash or gift cards. Crowdsourcing has become a common platform for businesses including multi-national companies, academic institutions, research, etc. for problem-solving to increase productivity and improve cost efficiency. Typical jobs that are performed using these services are image tagging, data categorizations, data collection, data analysis, sentiment analysis and survey. The requester could customize the jobs as per requirements and needs. The first question that hovers on our mind when considering these services would be the quality of the responses. The tasks are available online to any worker who wishes to respond to the task unless the requester has set restrictions (i.e., demographics, skills, etc.). Certain platforms offer workers with higher trust and skill levels which directly affects the quality of the respective crowdsourcing platform.

With increase in the number of crowdsourcing platforms available today there has also been a steep escalation in the number of online workers. Crowdsourcing is an extremely easy method for earning extra income for anybody i.e. elders, college students, graduates and professionals across the globe, as these workers are not monitored physically. As a result there has also
been an increase in spam and less reliable workforce, making it impossible for quality responses. Certain services could be more reliable with higher skilled workers compared to others. Certain services have restrictions for workers to register like geographical location, skills, education and minimum age requirement; this helps improve the quality of a platform and also improve the probability of getting quality results.

Kosinski, Yoram, Bachrach, Kasneci, Van-Gael and Graepel in [2] talk about only evaluating the performance of workers on Amazon Mechanical Turk using an IQ questionnaire. Although there is lot of ongoing research on crowdsourcing and there has been a large amount of existing study done on the concept of crowdsourcing, the main focus has only been around the actual practice and concept of crowdsourcing and how workers in general performed on a particular crowdsourcing service. But no study has so far evaluated the performance of workers across several platforms.

This study in particular evaluates the skills and the performance levels of workers across platforms to understand if all platforms are similar; if all the platforms provide solutions of same quality; to understand if all the solutions obtained on all these platforms are comparable in terms of their productivity and quality; to evaluate if all the workers are the same with similar efficiency producing comparable results and; lastly do any external factors like skills, difficulty of tasks, geographical locations, rewards offered affect the efficiency and performance of these workers.
# TABLE OF CONTENTS

LIST OF FIGURES ..................................................................................................................... viii

LIST OF TABLES ......................................................................................................................... xi

ACKNOWLEDGEMENTS .............................................................................................................. xii

DEDICATION ............................................................................................................................... xiii

Chapter 1 Introduction ............................................................................................................... 01

Chapter 2 Research Question .................................................................................................. 06

Chapter 3 Crowdsourcing Platform and Channels ................................................................. 09
   3.1. Crowdsourcing Platform ................................................................................................. 09
   3.2. Crowdsourcing Channels ............................................................................................... 12
      3.2.1. Amazon Mechanical Turk ........................................................................................ 13
      3.2.2. ClixSense .................................................................................................................. 13
      3.2.3. JunoWallet ................................................................................................................. 14
      3.2.4. NeoBux ...................................................................................................................... 15

Chapter 4 Experiment .............................................................................................................. 17
   4.1. Dataset Design ................................................................................................................ 20
   4.2. Task Design .................................................................................................................... 26
      4.2.1. Types of Tasks ............................................................................................................ 27
      4.2.2. Types of Sub-Questions ............................................................................................ 31
4.3. Experimental Details ................................................................. 32
  4.3.1. Design Job............................................................................. 32
    4.3.1.1 Data............................................................................... 32
    4.3.1.2 Build Job ........................................................................ 33
    4.3.1.3 Preview .......................................................................... 35
  4.3.2. Manage Quality .................................................................... 35
    4.3.2.1 Test Questions ............................................................... 35
    4.3.2.2 Workers ........................................................................... 36
    4.3.2.3 Job Settings .................................................................... 38
  4.3.3. Get Results .......................................................................... 40
    4.3.3.1 Launch ............................................................................ 40
    4.3.3.2 Monitor ............................................................................ 41
    4.3.3.3 Results ............................................................................ 43

Chapter 5 Data Analysis ................................................................. 44
  5.1. One Sample t-test ................................................................. 44
  5.2. Paired t-test ............................................................................ 44

Chapter 6 Hypothesis ................................................................. 46
  6.1. Hypothesis 1 .......................................................................... 46
  6.2. Hypothesis 2 .......................................................................... 51
  6.3. Hypothesis 3 .......................................................................... 55
  6.4. Hypothesis 4 .......................................................................... 60
  6.5. Hypothesis 5 .......................................................................... 63
LIST OF FIGURES

Figure 2.1 Binary tree representation of data hierarchies ......................... 07

Figure 4.1. Hierarchy representation of Animal .................................... 23

Figure 4.2. Hierarchy representation of Plant ....................................... 23

Figure 4.3. Hierarchy representation of Person ...................................... 24

Figure 4.4. Hierarchy representation of Artifact .................................... 24

Figure 4.5. Hierarchy representation of Sport ....................................... 25

Figure 4.6. Hierarchy representation of Geological Formation ................. 25

Figure 4.7. Hierarchy representation of Fungus .................................... 26

Figure 4.8. A sample IMG2TAG question .......................................... 28

Figure 4.9. A sample TAG2IMG question .......................................... 30

Figure 4.10. Set of tasks in a typical CrowdFlower job ......................... 33

Figure 4.11. Title and instructions of a task in CrowdFlower ................. 34

Figure 4.12. CML Editor in CrowdFlower ....................................... 34

Figure 4.13. Max Judgments per contributor and IP ........................... 37
Figure 4.14. Channel Settings in CrowdFlower .......................... 38
Figure 4.15. Task Settings in CrowdFlower .......................... 39
Figure 4.16. General settings for a job in CrowdFlower ............... 40
Figure 4.17. Launch settings for a job in CrowdFlower ............... 41
Figure 4.18. Dashboard for the results of a job in CrowdFlower ........ 42
Figure 4.19. Advanced analytics for the results of a job in CrowdFlower .. 42
Figure 4.20. Reports of a job in CrowdFlower .......................... 43
Figure 6.1. Pie Chart representation for H1 vs H2 vs H3 vs H4 ............ 48
Figure 6.2. A sample of (H1, H2) task ........................................ 49
Figure 6.3. A sample of (H1, H3) task ........................................ 50
Figure 6.4. Pie Chart representation for H1- IMG2TAG vs TAG2IMG ..... 53
Figure 6.5. Pie Chart representation for H2- IMG2TAG vs TAG2IMG ..... 53
Figure 6.6. Pie Chart representation for H3- IMG2TAG vs TAG2IMG ..... 54
Figure 6.7. Pie Chart representation for H4- IMG2TAG vs TAG2IMG ..... 54
Figure 6.8. Pie Chart representation for ClixSense- H1 vs H2 vs H3 vs H4 ................................................................. 57
Figure 6.9. Pie Chart representation for NeoBux- H1 vs H2 vs H3 vs H4

Figure 6.10. Pie Chart representation for H1- ClixSense vs NeoBux

Figure 6.11. Pie Chart representation for H2- ClixSense vs NeoBux

Figure 6.12. Pie Chart representation for H3- ClixSense vs NeoBux

Figure 6.13. Pie Chart representation for H4- ClixSense vs NeoBux

Figure 6.14. Pie Chart representation for H1- Animal vs Plant vs Person vs Fungus vs Artifact vs GF vs Sport

Figure 6.15. Pie Chart representation for H(X)- Animal vs Plant vs Person vs Fungus vs Artifact vs GF vs Sport

Figure 6.16. Pie Chart representation for H1- Africa vs Asia vs Europe vs North America vs Australia vs South America

Figure 6.17. Pie Chart representation for H(X)- Africa vs Asia vs Europe vs North America vs Australia vs South America

Figure 6.18. Pie Chart representation for H1- Asia vs Europe vs North America

Figure 6.19. Pie Chart representation for H(X) - Asia vs Europe vs North America
### LIST OF TABLES

Table 3.1 List of available channels in CrowdFlower ................................ 16

Table 4.1. Consolidated sample data set for Animal category .................. 22

Table 6.1. One-Sample T: H1, H2, H3, H4…………………………………….. 47

Table 6.2. One-Sample T: H1-IMG2TAG, H2-IMG2TAG, H3-IMG2TAG, H4-IMG2TAG ................................................................. 51

Table 6.3. One-Sample T: H1-TAG2IMG, H2-TAG2IMG, H3-TAG2IMG, H4-TAG2IMG ................................................................. 52

Table 6.4. One-Sample T: H1-ClixSense, H2-ClixSense, H3-ClixSense, H4-ClixSense ................................................................. 56

Table 6.5. One-Sample T: H1-NeoBux, H2-NeoBux, H3-NeoBux, H4-NeoBux ................................................................. 56

Table 6.6. One-Sample T: Animal-H1, Animal-H(X), Plant-H1, Plant-H(X), Person-H1, … ................................................................. 61

Table 6.7. One-Sample T: Africa-H1, Africa-H(X), Asia-H1, Asia-H(X), Europe-H1, … ................................................................. 64
ACKNOWLEDGEMENTS

I thank my adviser Dr. Dongwon Lee for his continuous support during my Masters’. I thank the professors at the College of Information Sciences and Technology who helped me gain the necessary knowledge on research and a qualitative approach to problem solving. I express my gratitude to Dr. Madhu Reddy and Dr. Erika Poole for their willingness to be a part of my thesis committee and for their invaluable guidance and feedback.

Most importantly, I am indebted to my family and friends for putting with me through all these hard times and thoroughly appreciate their encouragement and help. I would like to in specific thank my sisters, Neelima and Prathima and my brother Buddy for the endless love, support and guidance.
DEDICATION

I would like to dedicate the thesis to my parents, Prem Kumar and Venkata Ramani, and my husband Santosh.
Introduction

The term crowdsourcing has been coined in 2005 by Jeff Howe and Mark Robinson but the actual practice began around twelve years ago. The formal integrating definition for crowdsourcing has been developed by Enrique Estellés-Arolas and Fernando González Ladrón-de-Guevara, after studying more than forty definitions of crowdsourcing in the scientific and popular literature. Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken [1]. Crowdsourcing is a method used for individual, industrial and commercial, and educational purposes where
tasks are split into several micro tasks. These tasks are available to workers registered with the various available crowdsourcing platforms across the globe for benefits i.e., rewards like money and gift cards. Workers have the flexibility to respond to the tasks at their convenience and also are able to choose the level of complexity of the tasks. Various crowdsourcing platforms function as different types of organizations ranging from non-profit to fortune 500 companies.

Crowdsourcing is a distributed problem-solving method and production model. Usually, the problems are broadcast to an unknown group of solvers and an open call is given for solution. Users, known as the crowd, typically form into online communities, and the crowd submits solutions [4]. The crowd participates in sorting through the solutions, finding the best ones. These best solutions are then owned by the entity that broadcasted the problem called the crowdsourcer and the winning individuals in the crowd are usually rewarded [4].

Crowdsourcing is a powerful concept that harnesses the power of a community, using advanced social media technology and knowledge of crowd behavior, to collect, evolve and rank ideas and contributions to reveal the strongest performers [4].

Crowdsourcing can shorten time to market for new products, help in uncovering ways to cut costs or improve service levels, and increase market success for new products or enhancements [4]. Crowdsourcing technology is
used by companies, organizations and governments to rapidly and economically understand market needs, opportunities and priorities [4].

Crowdsourcing relies on human workers to complete a job, but humans are prone to errors, which can make the results of crowdsourcing arbitrarily bad. The reason is two-fold. First, to obtain rewards, a malicious worker can submit random answers to all questions [5]. This can significantly degrade the quality of the results. Second, for a complex job, the worker may lack the required knowledge for handling it. As a result, an incorrect answer may be provided [5].

The benefits of crowdsourcing include having an opportunity to get tasks done in a short timeframe at lower budgets. Workers have the ability to compete against a higher talent unless specified by the requester, helps workers build their expertise and portfolio and most importantly workers have the ability to work at his/her convenience.

The main objective of this study is to evaluate the performances of workers across different crowdsourcing platforms. For this purpose, five hypotheses have been derived from the results based on the hierarchy of the questions i.e. H1 vs H2 vs H3 vs H4, the two types of tasks i.e. choose the correct image for the corresponding tag and choose the correct tag for the corresponding image, the two channels i.e. ClixSense vs. NeoBux, the seven different categories of images and their tags and finally the geographical location of the workers who participated in the study. Each of these hypothesis
have been discussed in chapter 10. The performance is analyzed by posting similar tasks on various platforms. For the study, CrowdFlower was the chosen crowdsourcing platform. The study focused on four different platforms. The tasks have been designed using CrowdFlower’s API and CrowdFlower Markup Language. Each task comprises of two sub-questions, a filtering question and a real question; the main purpose of the filtering question is to understand the workers intentions and their attention to detail on completing the task and to also discourage and filter spam.

CrowdFlower is used as the parent platform for conducting the experiment. Using CrowdFlower’s user friendly developer API and design tools several tasks of varying complexity have been launched on four channels that are part of their channel partner list. The flexibility of being able to get responses from workers across different channels by developing the survey using a single API instead of having to use individual APIs on each of these channels individually is the key reason to choose CrowdFlower. The platform was used to launch two types of tasks of four hierarchy level images and tags onto four separate channels to compare the workers performance on each of the four channels respectively. The two types of questions used in this study are

1. Choose the correct tag for the corresponding image.

2. Choose the correct image for the corresponding tag.
Each task was further split into two sub-questions, a filtering question and a real question. The purpose of having a filtering question was to analyze the workers attentiveness and interest while responding and most importantly to filter and discourage spam.

Using Minitab statistical software, the results from the survey have been evaluated. The results show that the hierarchy of the question affected the worker performance. But surprisingly hierarchy two (H2) tasks were the least accurately answered followed by hierarchy three (H3) and hierarchy four (H4). The reason might be that certain images and their tags in H3 were relatively less complex than H2. The workers responded to “choose the correct image for the corresponding tag (TAG2IMG)” type of questions better than “choose the correct tag for the corresponding image (IMG2TAG)” type of questions in H1, H2 and H4 whereas for H3, IMG2TAG were more accurate. Overall, ClixSense performed better than NeoBux. The maximum number of responses were from Asia and Europe and only a few responses from other continents like Africa, Australia and North America and South America. But the statistics show that workers from North America provided the maximum number of correct responses to the tasks. The last analysis showed that Plant category was the most correctly responded tag genre and Sport was the least correct answered followed by Fungus and Animal category respectively.
Research Question

The objective of the study is to analyze and evaluate the performance of workers across different crowdsourcing platforms. The parent platform that has been used for the study and to conduct the experiment is CrowdFlower. CrowdFlower allows launching tasks using their API and scripting tools on to their various channels i.e. platforms partnered with them. The flexibility of being able to get results from workers across different platforms by developing the survey using a single API instead of having to use APIs on individual channels is the reason CrowdFlower has been chosen.

Using CrowdFlower, similar tasks of varying complexity have been launched on four separate channels to compare the workers performance levels on each platform individually.

For the survey, ImageNet, an image database has been used to create the micro tasks. This database which is an organization of images on hierarchy basis has been represented in a binary tree format as seen in figure 2.1.
Two types of questions have been used for designing the jobs:

1. Choose the correct tag for the corresponding image
2. Choose the correct image for the corresponding tag

Each task has been further categorized into two questions i.e. the filtering question and a real question. The purpose of including a filtering question for each task was to discourage and filter spam.

For this purpose, using ImageNet four hierarchies i.e. (H1, H2, H3, and H4) have been used. Hundred instances of (H1, H2), (H1, H3), (H1, H4) each have been derived with two types of task format combinations i.e.

1. [ (H1, H2)$_{100}$, (H1, H3)$_{100}$, (H1, H4)$_{100}$ ] – IMG2TAG
2. [ (H1, H2)$_{100}$, (H1, H3)$_{100}$, (H1, H4)$_{100}$ ] – TAG2IMG
\[ (H1, H2)_{100} + (H1, H3)_{100} + (H1, H4)_{100} \] * Two types of tasks * Four channels = 2400 tasks. Two thousand four hundred tasks have been launched on CrowdFlower for the experiment.

Hierarchy one i.e. H1 consists of the easiest images with their respective tags. The images and tags belonging to this hierarchy like animal or plant are always used as filtering questions and H2, H3 and H4 for real evaluating questions. The difficulty of tasks increases as the hierarchy increases, with H4 being the most complex tasks used for the study.

The results of the survey have later been categorized into different scenarios for understanding how the crowd behaved and performed given different situations and under various factors.
Crowdsourcing Platform and Channels

Platforms are the medium that connect the job requester and the worker. CrowdFlower is a crowdsourcing platform. CrowdFlower has a network of channel partners. They partner with various websites and communities that manage the registration and payment of their users, members or contributors [17] like Amazon Mechanical Turk, ClixSense, NeoBux and JunoWallet.

3.1 Crowdsourcing Platform

CrowdFlower is one of the leading platforms in the enterprise crowdsourcing businesses started by a group of young talented entrepreneurs from San Francisco, California, USA in 2009. CrowdFlower has the highest on-demand workforce with over five million workers. It has contributors from 208 countries and territories, who have completed over one billion plus judgments so far.

CrowdFlower offers a wide range of solutions to its clients like

- Data collection and enhancement
- Improving the quality of data in a business through business intelligence
• Data categorization to organize data for a business and also to make it more navigable

• Sentiment Analysis to understand the real and internal meaning of how people think in situations of various circumstances. For example, with social media becoming a medium for celebrities and businesses to reach to their fan base and followers, sentiment analysis helps understand how general crowd respond to what is being posted on social media platforms like Twitter, Facebook, etc.

• Content Creation for various purposes like articles, sales and marketing purposes for an organization of high quality

• Content Moderation used for filtering data i.e., eliminating improper data

• Conducting surveys for multiple purposes targeted at large amount of crowd source for better understanding of the study or respective purpose; and

• Used for Relevance Tuning Search where biggest names in search and ecommerce rely on CrowdFlower's human-calibrated search relevance solution to train their algorithms to consistently provide the most relevant search results for their customers [17].
CrowdFlower provides the flexibility of creating our own micro tasking jobs with tools that can be used to customize tasks. CrowdFlower provides a developer API to design tasks and complex applications which can interact with and utilize all the advanced features that the platform offers in an automated way. CrowdFlower offers task design tools reaching to all the unique needs and challenges for different businesses. CML, CSS and JavaScript allow partners or clients to create jobs with specific needs and requirements. CrowdFlower assures high quality control by providing confidence score for every unit of work completed to track everything from contributor response velocity to answer distribution [17]. Requestors can also customize the target of workers for the job by geography, demographics and trust levels.

CrowdFlower has a network of channel partners that contribute to their five million contributors or crowd force. They partner with various websites and communities that manage the registration and payment of their users, members or contributors [17]. In exchange, they provide a steady stream of online micro tasks for users to complete. This platform handles the task distribution and quality assurance. CrowdFlower’s partners include micro task marketplaces, rewards sites, outsourcing companies, social game publishers, offer walls and social enterprises [17]. The channels partnered with CrowdFlower are ClixSense, CrowdGuru, instaGC, NeoBux, SwagBucks and many others.
Until the end of 2013, by the time when the tasks for this study were launched, CrowdFlower worked slightly different than what they do now. Requesters were able to select the channels of their choice to make specific jobs available to only contributors from those channels but now this flexibility has been modified leaving requestors no option of being able to choose a selected group but rather select CrowdFlower internal workforce i.e., all the five million contributors together. But, CrowdFlower tracks every detail possible including the channel from which the response is coming.

CrowdFlower’s customers consists of some giants from various industry sectors like Apple, Google, Instagram, Microsoft, Adobe, Walmart, Lowe’s, Ford, Toyota, Tripadvisor, Imgur, YouTube, Unilever, eBay, etc.

3.2 Crowdsourcing Channels

CrowdFlower is partner with various websites and communities that manage the registration and payment of their contributors [17]. In exchange, CrowdFlower provides a steady stream of online micro tasks for users to complete. The platform handles the task distribution and quality assurance [17]. For the study, four of the following channels have been chosen on the basis of their number of Facebook likes and the percentages of trusted & untrusted levels, refer table 3.1 for details about the channels available and their related information. The channels that have been chosen for this study are:
3.2.1. Amazon Mechanical Turk

Amazon Mechanical Turk is one of the oldest and robust crowdsourcing platforms with the highest requester rate as well as worker traffic. A lot of MNCs use Amazon Mechanical Turk which increases the amount of micro tasks of various genres available to the worker offering high incentives. Due to this reason, the results for tasks launched in AMT were slow with less than fifteen responses in the first 24 hours of posting the tasks; also the workers preferred providing responses to only the easiest set of jobs (H1, H2). The responses being slow, has pushed the tasks behind in the queue which led to re-launching the same tasks with the same incentive on Amazon Mechanical Turk’s website directly. The incentive offered being really less has made it difficult to gather responses even from AMT after re-launching the tasks.

3.2.2. ClixSense

ClixSense is an industry proven method that allows website publishers of every size or budget to direct targeted and unique traffic to their website [18]. ClixSense offers this opportunity to every web merchant whether a startup business with a minimal advertising budget or businesses already established but looking for additional web exposure [18]. ClixSense allows tasks to be open to potential workers even with low incentives. ClixSense has developed sophisticated online tracking and monitoring systems to ensure the best quality products for their customers.
ClixSense is among one of the many channels available on CrowdFlower that provide money as a reward rather than gift cards. The responses for the channel were quick and took less than an hour to complete all the six hundred tasks after launching regardless of the type of the question and the hierarchy of the task.

3.2.3. JunoWallet

JunoWallet is a social marketing rewards app that offers its workers rewards for their work in form of gift cards. It works exclusively with CPA (Cost-Per-Action) ad campaigns [19]. With highly engaged users spread all over the globe, they fill campaigns quickly and are always looking for great partners with diverse campaigns from all over the world [19].

JunoWallet is the only channel that didn’t get any responses even after a month of launching the tasks. For this reason, the jobs had to be re-launched using CrowdFlower’s internal workforce, meaning it was a combination of workers from all the available channels. The responses were quick but the channels were mostly ClixSense and NeoBux. So the responses received from the internal workforce were excluded.
3.2.4. NeoBux

NeoBux is a platform developed to provide paid-to-click services. NeoBux is a free worldwide service available in a multi-language environment [20]. Their service consists of allowing advertisers to reach thousands of potential customers by displaying their advertisement(s) on NeoBux and users to earn money by viewing those advertisements [20]. Their registered workers click on the advertiser's advertisement and view them within the amount of time specified by the advertiser. After viewing the advertisement, the user gets credited with a pre-determined amount of money in their NeoBux account [20]. It is a mobile-friendly channel meaning its workers have the flexibility of logging in from their smartphones to finish tasks and make money on the go.

The responses from NeoBux for the tasks posted were quick which took less than an hour to complete all of the six hundred tasks after launching regardless of the type of the task and the hierarchy of the task. Refer to table 3.1 for a list of all the channels on CrowdFlower and their details, like number of Facebook likes and trust levels and so on.
<table>
<thead>
<tr>
<th>Channels</th>
<th>Trust %</th>
<th>Top Countries</th>
<th>Members</th>
<th>Facebook Likes</th>
<th>Google+ Likes</th>
<th>Twitter Followers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Mechanical Turk</td>
<td>92</td>
<td>USA(59), IND(35), CHN(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clicks-FX</td>
<td>96</td>
<td>USA(33), SGP(16)</td>
<td>18290</td>
<td>515</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>ClixSense</td>
<td>84</td>
<td>USA(15), PHIL(8), IND(8)</td>
<td>3,442,085</td>
<td>97,823</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EntropiaPartners(SEcond Life)</td>
<td>92</td>
<td>USA(45), DEU(10), BGR(10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FresherWorld87</td>
<td>84</td>
<td>USA(82), GBR(12), DEU(4)</td>
<td>120131</td>
<td>474</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FusionCash</td>
<td>92</td>
<td>USA(97), CAN(2), PRI(0)</td>
<td>432135</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get-Paid.com</td>
<td>90</td>
<td>IND(15), PHL(15), ROU(9)</td>
<td>(nice website)</td>
<td>42k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JunoWallet</td>
<td>100</td>
<td>USA(81), RUS(19)</td>
<td></td>
<td></td>
<td>106k</td>
<td></td>
</tr>
<tr>
<td>NeoBux</td>
<td>79</td>
<td>USA(11), IND(8), ROU(5)</td>
<td>2,524,732</td>
<td>13,586</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perk.com</td>
<td>100</td>
<td>USA(64), IND(36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poin-web</td>
<td>Non-English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Points to Shop</td>
<td>83</td>
<td>USA(38), GBR(21), IND(9)</td>
<td></td>
<td>242k</td>
<td>6.9k</td>
<td></td>
</tr>
<tr>
<td>PrizeRebel</td>
<td>85</td>
<td>USA(58), CAN(20), GBR(4)</td>
<td></td>
<td></td>
<td>65k</td>
<td>529</td>
</tr>
<tr>
<td>StuffPoint</td>
<td>91</td>
<td>CHL(14), DNK(10), USA(9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swag Bucks(Prodege)</td>
<td>90</td>
<td>USA(42), CAN(39), GBR(14)</td>
<td>5 million</td>
<td>1596719</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tremor Games</td>
<td>82</td>
<td>USA(15), DEU(14), POL(9)</td>
<td></td>
<td>26,211</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZoomBucks</td>
<td>93</td>
<td>CAN(47), USA(26), VNM(6)</td>
<td></td>
<td></td>
<td>9835</td>
<td></td>
</tr>
<tr>
<td>instaGC</td>
<td>90</td>
<td>USA(81), CAN(7), MYS(7)</td>
<td>38,335</td>
<td>19,503</td>
<td>27414</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1. List of available channels in CrowdFlower
Experiment

The experiment has been conducted on CrowdFlower. It provides requestors and contributors the flexibility of customizing the settings for their micro tasks (image tagging, content writing, surveys, etc.) and also provides excellent workforce management (Geography, Channels and Skills and performance levels). For a user to post tasks, he/she will have to follow a series of steps that are discussed later in this chapter.

CrowdFlower has a user-friendly form builder available for its customers to easily create crowdsourcing jobs specific to their needs. Users can also take advantage of CrowdFlower Markup Language paired with custom CSS and JavaScript - all powered by Twitter's Bootstrap front-end framework [17]. When data is being uploaded onto the platform, it has to be only a Comma Separated value (csv).

For advanced customers who wish to build complex applications that interact and utilize all of CrowdFlower's most advanced features in an automated way, the platform has a user friendly developer's API. User can gain increased control over their job while still taking advantage of the built-in time-saving tools [17].
For the study, the following are the series of steps that have to be followed after uploading the spreadsheets containing the image URLs and its corresponding four options that includes the correct tag for IMG2TAG questions and the tags with four image URLs containing the correct image URL for TAG2IMG questions, as a comma separated value file.

After uploading the file, taking advantage of the CML scripting tool the jobs have been designed; followed by

1. Previewing the designed task to check if the task meets all requirements
2. Selecting the amount or incentive that the user wishes to assign for each task to the worker; for this study an incentive of one cent per each task has been assigned
3. Now, the requester would have to select the number of tasks that a worker would be able to respond. For the study, only one person per task has been allowed to be able to evaluate the performance of a larger number of individual workers and get larger number of unique responses.
4. Next, we would have to select the skill level of the worker who we wish to assign the task; CrowdFlower groups workers on three skill levels: Level 1, Level 2, and Level 3. Level 1 skill would be the workers with the highest skills and Level 3 workers with the least amount of skills. For the experiment, only Level 1 workers have been chosen.
5. Users are also able to set the time that they wish to assign to each worker in order to respond to a task. This study has allotted a time limit of two minutes per each task which includes answering both the filtering as well as the real questions that are included into each task.

6. The sixth step would be one the most important steps where the requester would have to select the channels that he/she wishes to assign their tasks and be able to use those channel workforce only. CrowdFlower is the default option i.e., using its internal workforce that is a combination of workers from all the channels that CrowdFlower is partner. The requester could either use the internal workforce or select the channels listed as per choice and requirement. For this study, the channels that have been used are Amazon Mechanical Turk, ClixSense, JunoWallet and NeoBux. The reason for choosing these four channels has been illustrated in table 3.1.

Also, for future research references, CrowdFlower has made changes to its platform starting 2014, the requester no longer have the choice of choosing channels but would rather have to choose CrowdFlower's internal workforce.

7. Lastly, the user would have to set restrictions for the workers. The two options available are – show results unanswered; show incorrectly answered and not fully responded results. Both these options were used
when setting up jobs because this study is based on understanding the workers' intelligence and IQ.

8. Finally, the platform would prompt the user to enter an email address to which CrowdFlower would send notifications when tasks are completed.

9. Just before launching the tasks the user would have to enter reference i.e., tags, types and channel names for identification purposes to avoid confusion later.

10. All the settings at this stage would have to be saved and are ready to be launched and available to online workers.

4.1 Dataset Design

ImageNet was developed by Stanford University. It is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. There are more than 100,000 synsets in WordNet, majority of them are nouns (80,000+) [21]. They aim to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated [21]. Currently it comprises an average of over five hundred images per node. The database is targeted mainly to researchers, educators and students.
This database has been the primary source in creating the dataset used for the study. The database is hierarchical i.e., (H1, H2, H3, H4, H5……) with H1 being the highest level hierarchy and so on. This study focuses only on the top four hierarchies i.e., (H1, H2, H3, and H4) and the notation is represented in figure 2.1.

Imagenet.org mainly comprises of nine categories considered as H1, among which this study has used only seven categories depending upon the availability of images in ImageNet under each category.

Considering the binary tree in figure 2.1 under every class, five image URLs or tags have been used depending on the type of question; under each of the seven H1 selected, five image URLs for each subnets/sub-hierarchies H2, H3, H4 have been selected. This has resulted in seventy five image URLs under each main category H1, includes five image URLs of H1 as shown in consolidated table 4.1. The seven categories used were \{Animal, Plant, Person, Fungus, Artifact, Sport, Geological Formation\}.

The database consists of image URLs that are downloadable along with the tags for each image URL respectively and individually. The three image combinations that are used for the study are: (H1, H2), (H1, H3), (H1, H4).
<table>
<thead>
<tr>
<th>Hierarchy Level</th>
<th>Category</th>
<th>ImageURL</th>
<th>Correct Option</th>
<th>Option 2</th>
<th>Option 3</th>
<th>Option 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Animal</td>
<td><img src="http://farm4.static.flickr.com/3178/2678394084_ce9235b1e.jpg" alt="Image" /></td>
<td>Animal</td>
<td>Plant</td>
<td>Person</td>
<td>Fungus</td>
</tr>
<tr>
<td>H2</td>
<td>Omnivore</td>
<td><img src="http://farm4.static.flickr.com/3178/2678394084_ce9235b1e.jpg" alt="Image" /></td>
<td>Omnivore</td>
<td>Hexapod</td>
<td>Biped</td>
<td>Larva</td>
</tr>
<tr>
<td>H2</td>
<td>Invertebrate</td>
<td><img src="http://farm1.static.flickr.com/108/276140204_62b5bd9935.jpg" alt="Image" /></td>
<td>Invertebrate</td>
<td>Omnivore</td>
<td>Herbivore</td>
<td>Insectivore</td>
</tr>
<tr>
<td>H3</td>
<td>Worm</td>
<td><img src="http://farm1.static.flickr.com/108/27611849203_1988e1e49d.jpg" alt="Image" /></td>
<td>Worm</td>
<td>Rotifer</td>
<td>Peanut worm</td>
<td>Woodborer</td>
</tr>
<tr>
<td>H3</td>
<td>Shellfish</td>
<td><img src="http://farm1.static.flickr.com/185/489000105_e32d34f286.jpg" alt="Image" /></td>
<td>Shellfish</td>
<td>Peanut worm</td>
<td>Lamp Shell</td>
<td>Rotifer</td>
</tr>
<tr>
<td>H3</td>
<td>Young Mammal</td>
<td><img src="http://farm4.static.flickr.com/3140/268210932_7e61b3a2a4.jpg" alt="Image" /></td>
<td>Young Mammal</td>
<td>Orphan</td>
<td>Spat</td>
<td>Hatchling</td>
</tr>
<tr>
<td>H3</td>
<td>Young Bird</td>
<td><img src="http://farm2.static.flickr.com/1051/104654226_0c37f7a3f0.jpg" alt="Image" /></td>
<td>Young Bird</td>
<td>Hatchling</td>
<td>Orphan</td>
<td>Young Mammal</td>
</tr>
<tr>
<td>H4</td>
<td>Arrowworm</td>
<td><img src="http://images.aad.gov.au/img.py/1fc2.jpg" alt="Image" /></td>
<td>Arrowworm</td>
<td>Woodworm</td>
<td>Helminth</td>
<td>Flatworm</td>
</tr>
<tr>
<td>H4</td>
<td>Flatworm</td>
<td><img src="http://farm1.static.flickr.com/185/426959025_6519d2e248.jpg" alt="Image" /></td>
<td>Flatworm</td>
<td>Ribbon worm</td>
<td>Woodworm</td>
<td>Helminth</td>
</tr>
<tr>
<td>H4</td>
<td>Gastropod</td>
<td><img src="http://farm4.static.flickr.com/3525/3826187781_ad6656cc6.jpg" alt="Image" /></td>
<td>Gastropod</td>
<td>Chiton</td>
<td>Scaphopod</td>
<td>Bivalve</td>
</tr>
<tr>
<td>H4</td>
<td>Bivalve</td>
<td><img src="http://farm4.static.flickr.com/3432/3845870244_c7d6294755.jpg" alt="Image" /></td>
<td>Bivalve</td>
<td>Gastropod</td>
<td>Chiton</td>
<td>Scaphopod</td>
</tr>
<tr>
<td>H4</td>
<td>Lamb</td>
<td><img src="http://farm2.static.flickr.com/1357/1404762032_25528827d3e5.jpg" alt="Image" /></td>
<td>Lamb</td>
<td>Baby</td>
<td>Calf</td>
<td>Cub</td>
</tr>
<tr>
<td>H4</td>
<td>Calf</td>
<td><img src="http://farm1.static.flickr.com/175/38076245_e07226fbdf.jpg" alt="Image" /></td>
<td>Calf</td>
<td>Kit</td>
<td>Lamb</td>
<td>Suckling</td>
</tr>
<tr>
<td>H4</td>
<td>Chick</td>
<td><img src="http://farm3.static.flickr.com/2169/2365292082_32176a2eba.jpg" alt="Image" /></td>
<td>Chick</td>
<td>Eaglet</td>
<td>Cygnet</td>
<td>Nestling</td>
</tr>
<tr>
<td>H4</td>
<td>Nestling</td>
<td><img src="http://farm3.static.flickr.com/2088/2446247663_94c9d2dfae.jpg" alt="Image" /></td>
<td>Nestling</td>
<td>Chick</td>
<td>Fledgling</td>
<td>Eaglet</td>
</tr>
</tbody>
</table>

Table 4.1. Consolidated sample data set for Animal category
A detailed overview of the binary tree representation for each of the seven categories are shown below:

Figure 4.1. Hierarchy representation of Animal

Figure 4.2. Hierarchy representation of Plant
Figure 4.3. Hierarchy representation of Person

Figure 4.4. Hierarchy representation of Artifact
Figure 4.5. Hierarchy representation of Sport

Figure 4.6. Hierarchy representation of Geological Formation
4.2 Task Design

The study used two types of questions for the evaluation of worker performances across the four channels. The tasks were designed such that maximum evaluation could be tapped when using different types of formats for the study.

Each task irrelevant of the type of question had been designed such that it consists of a filtering question and a real question. The filtering question and the real question always belonged to the same type of question. The filtering question is always a tag/image of H1 whereas the real question could be either H2, H3 or H4.
4.2.1. Types of Tasks

Choose the correct tag for the corresponding image:

For this type of question, an image was provided with four options of tags among which one would be the right tag. A total of seventy five image URLs under each category were chosen i.e., five image URLs under each hierarchy (H1, H2, H3 and H4) of that category, having an option set of four tags among which one would be the correct tag corresponding to the image and the remaining three being randomly picked tags from the same hierarchy of the same category. Figure 4.8 illustrates IMG2TAG type of question.
Figure 4.8. A sample IMG2TAG question.
Choose the correct image for the corresponding tag:

This type of question was similar to the first type except that the worker would be given a tag with four images to choose the correct image. A total of seventy five tags under each category were used i.e., five tags under each hierarchy (H1, H2, H3 and H4) of that category, having an option set of four image URLs of which one would be the correct image corresponding to the tag and the remaining three being randomly picked image URLs from the same hierarchy of the same category. Figure 4.9 illustrates TAG2IMG type of question.
Figure 4.9. A sample TAG2IMG question.
4.2.2 Types of Sub-Questions

Each individual task comprises of two questions-filtering question and real question.

Filtering Question: The filtering questions are always H1, the highest hierarchy that are relatively very simple and can be answered correct without high skills and little attention to detail would be sufficient. The purpose of the filtering question was mainly to understand how consciously the worker is providing his responses.

Real Question: The real questions were framed using hierarchies H2, H3 and H4. Typically the difficulty level for these questions ascends, with H4 requiring high level skills, intelligence, attention to detail and most importantly concentration and interest.

The three combinations of tasks used were: (H1, H2), (H1, H3) and (H1, H4) with their difficulty and complexity level increasing as the hierarchy gets more granular. Both the sub-questions i.e., the filtering and the real questions were always paired within the same type of question meaning an IMG2TAG type of question would be tagged with an H1 level question for filtering in the same type of question format and similarly TAG2IMG type of question was always tagged with an H1 level question for filtering in the same type of question format, continued with a real question in a similar format from H2, H3 or H4.
4.3 Experiment Details

4.3.1. Design Job

In this step, data was uploaded to CrowdFlower in the format that is suitable to process and post jobs. The framework of the job is designed and built using CML editor. A preview of the tasks is provided after each job design.

4.3.1.1. Data

The format of the uploaded data structure is: H1 based question followed by four options and H2, H3 or H4 question followed by four options. A sample screen shot of units in a job can be seen in figure 4.10. Six jobs per channel were created (3-IMG2TAG, 3-TAG2IMG). Each job has 100 data units. Each channel has been posted with (H1, H2), (H1, H3), (H1, H4) questions in both Img2Tag and Tag2Img categories. The channels chosen for this experiment were NeoBux, ClixSense, AMT, and JunoWallet. So, a total of 2400 tasks (4x2x3x100) were created for this experiment.
4.3.1.2. Build Job

This step is crucial because this is what the workers see when performing tasks. This represents the main design and task details of each job. Title and instructions for the job are added. Title and instructions are very important. They provide the required details to the workers to perform their task. The title should be very precise and short. Instructions should include details as to what the worker is expected to do to complete the task. Using the data, proper locations are assigned as to where they reside and add questions for the job to be performed. CrowdFlower provides their legacy custom CML editor which is similar to HTML and JavaScript syntactical notation. In the CML editor, we can design and theme out the questions of how they appear to the workers. In
figure 4.11 and figure 4.12, we can see screen shots of title, instructions and CML code. CML is CrowdFlower Markup Language. CML is made up of a set of helper tags, which makes defining forms to collect information from labor pools quick and painless.

Figure 4.11. Title and instructions of a task in CrowdFlower

Figure 4.12. CML Editor in CrowdFlower
4.3.1.3. Preview

After processing data, designing tasks and adding questions – a preview is generated of how the task would appear to the workers. Also, number of units per task is provided at this step.

Units per Task: This is the number of units that a worker must complete on a page before submitting their answers.

The number of units per task was chosen as one because the study was focused on getting just one response from one unique worker. Multiple responses from the same worker will result in ambiguous and irrelevant responses. The job is saved at regular intervals to avoid any data loss.

4.3.2 Manage Quality

After setting up the job and previewing to verify that the task is set up without any issues, test questions are setup, workers for the job are selected and job settings are assigned.

4.3.2.1 Test Questions

Test Questions are units with known answers that are regularly inserted throughout the job. They drastically improve results by:

- training workers – workers see why they got the test question wrong if they fail
• removing underperformers – workers are removed if they fail to answer too many test questions

In this experiment, we did not ask any test questions because every worker was allowed to work on only one task per job. Forcing unique workers to perform the tasks is key to this experiment.

4.3.2.2 Workers

Workers can be selected at this step. They can be selected based on their location, skills, behavior settings, channels.

• Geography- The options available are all countries, specific countries or exclude specific countries. For this experiment, workers from all countries were allowed to work on the tasks.

• Skills- Skill requirements can be assigned to allow workers that meet these requirements before working on the tasks. Level 1, level 2 and level 3 workers are available to select. Level 1 workers were selected so that we get accurate responses. Though not assigned in this experiment, finer tune settings are available to specify language skills required for the worker.

• Behavior Settings: These are used to assign the maximum judgments per worker and maximum judgments per IP.
  
  o Max Judgments per Worker: This is the total number of judgments that any one worker can complete.
- Max Judgments per IP: Max responses that can be received per IP address for a task.

In order to get unique responses, max judgments per worker is set to one and max judgments per IP is also set to one as seen in figure 4.13.

![Figure 4.13. Max Judgments per contributor and IP](image)

- Channels: This was most important setting for this experiment. CrowdFlower has over five million workers in its on-demand workforce, ready and waiting to complete the jobs. The channels chosen for this experiment were
  - Amazon Mechanical Turk
  - ClixSense
  - JunoWallet
  - NeoBux

The responses from these channels are analyzed in later chapters.
4.3.2.3. Job Settings

The number of units per task, payment per task, task expiration time, judgments per unit and other settings can be assigned here.

- Tasks: In figure 4.15, we can see that units per task, payment and task expiration time can be set up. One cent per task has been paid to the worker upon successful completion of the task.
Unites per Task: This is the number of units that a worker must complete on a page before submitting their answers.

Payments Cents per Task: This is how much a contributor gets paid for each task.

Task Expiration Time: This is the time in minutes before a task expires for the worker.

- **Judgments**: The judgments per unit is set to one. This is the number of workers who will complete each unit of data.
- **General**: Here we can assign tags to jobs for easy retrieval and organization. Also, we can add e-mail’s so that we can be notified by CrowdFlower once all the tasks in the job are successfully completed. This can be seen in figure 4.16.
API: Other CrowdFlower API settings can be changed here.

4.3.3. Get Results

Once all the job settings are assigned appropriately. The jobs are ready to be launched, monitored and get results.

4.3.3.1. Launch

The number of units of job to launch are set to 100. Account due and balance is shown here. If there is no account balance, credit can be added here. Figure 4.17 shows a job with 100 units which is ready to be launched.
4.3.3.2. Monitor

A dashboard is presented showing real time responses, times at which tasks were completed, number of responses received, pending and approved. Figure 4.18 shows a graphical representation of this data along with satisfaction rating of workers who performed these tasks.
Advanced Analytics give a detailed summary about workers, quality of responses, distributions etc.
4.3.3.3. Results

Apart from the dashboard and advanced analytics monitoring, after all the tasks in the job are completed, all the results in raw form can be downloaded for further analysis of results. CrowdFlower offers full, aggregated, source, worker data and Json format raw result files to download.

- Full - This is the full dataset of all workers judgments.
- Aggregated - CrowdFlower aggregates the workers judgments to give the highest confidence result for each unit.
- Source – Results along with source data included.
- Worker – Data related to workers who performed the tasks.
- Json- Results data in Json format.

![Figure 4.20. Reports of a job in CrowdFlower](image)
Data Analysis

Data gathered from the experiment is analyzed using two statistical methods which are One Sample t-test and Paired t-test.

5.1. One Sample t-test

The One sample t-test is used to calculate the mean for the samples individually which is the aggregation of all the H1 from the two channels including the two types i.e., the IMG2TAG and TAG2IMG from ClixSense as well as NeoBux and similarly for H2, H3 and H4. The analysis also provides the confidence interval which is the percentage range that the particular set of hierarchy data (H1, H2, H3 or H4) can be answered correctly. Higher the Confidence Interval (C.I.) better is the chance that the question has been answered correctly.

5.2. Paired t-test

The Paired t-test is used to measure the independent means of the three set of hierarchy questions i.e., {H1 H2}, {H1 H3}, {H1 H4}. This test provides the difference of the two means i.e. mean of H1 and mean of H2 (in the case of (H1,
H2)); these values define the accuracy of the workers performance by checking the number of workers that answered both the filtering question as well as the real question correctly. This method also provides the confidence interval for the $\mu$ difference, higher the confidence level better is the accuracy for the respective set of questions.

CrowdFlower organizes the survey by providing intensive details, the IP address of the worker; the country, region and the city from where the worker is providing their input; workers trust level; survey questions i.e., (H1 H2), (H1 H3) or (H1 H4) along with the four given options for the respective question and the response to it from the worker. Basing on this data, five hypothesis have been deduced for the study

- How the hierarchy affected the workers responses?
- How the type of the question affected the workers performance?
- How the two channels overall varied in their worker performance and accuracy?
- Worker performance based on the region they belonged.
- Worker performance based on the tag genre.
Hypothesis

After the analysis of the data downloaded from CrowdFlower, five hypotheses have been derived to further understand scenarios that helped to evaluate the workers performance. Both one sample t-test and paired t-test have been applied for all five hypotheses individually. The inputs for the tests were collected by combining hierarchies as per the requirement of the hypothesis.

6.1. Hypothesis 1

*Title:* How the hierarchy of the image/tag affect the performance of the worker?

*Input:* The data used for the hypothesis are H1, H2, H3 and H4. H1 consisting of an aggregated list of H1’s from NeoBux and ClixSense including IMG2TAG and TAG2IMG

\[
\begin{align*}
H1: \{(H1_{IMG2TAG} + H1_{TAG2IMG}) \text{ ClixSense} + (H1_{IMG2TAG} + H1_{TAG2IMG}) \text{ NeoBux}\} \\
H2: \{(H2_{IMG2TAG} + H2_{TAG2IMG}) \text{ ClixSense} + (H2_{IMG2TAG} + H2_{TAG2IMG}) \text{ NeoBux}\} \\
H3: \{(H3_{IMG2TAG} + H3_{TAG2IMG}) \text{ ClixSense} + (H3_{IMG2TAG} + H3_{TAG2IMG}) \text{ NeoBux}\} \\
H4: \{(H4_{IMG2TAG} + H4_{TAG2IMG}) \text{ ClixSense} + (H4_{IMG2TAG} + H4_{TAG2IMG}) \text{ NeoBux}\}
\end{align*}
\]
### Table 6.1. One-Sample T: H1, H2, H3, H4

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>1154</td>
<td>0.91421</td>
<td>0.28017</td>
<td>0.00825</td>
<td>(0.89803, 0.93039)</td>
</tr>
<tr>
<td>H2</td>
<td>402</td>
<td>0.5846</td>
<td>0.4934</td>
<td>0.0246</td>
<td>(0.5362, 0.6330)</td>
</tr>
<tr>
<td>H3</td>
<td>402</td>
<td>0.7015</td>
<td>0.4582</td>
<td>0.0229</td>
<td>(0.6566, 0.7464)</td>
</tr>
<tr>
<td>H4</td>
<td>350</td>
<td>0.6629</td>
<td>0.4734</td>
<td>0.0253</td>
<td>(0.6131, 0.7126)</td>
</tr>
</tbody>
</table>

**Output:** Referring to the data from the table 6.1, one-sample t-test shows the highest mean value for H1 with 91.4% accuracy rate as expected because H1 are the least difficult hierarchy questions used as filtering questions. The next hierarchy questions were H3 with 70% accuracy followed by H4 with 66% accuracy and H2 with 58% accuracy respectively. The reason for H4 being better responded than H2 was the less sample size meaning lesser number of workers were willing to answer the most difficult H4 tasks. But interestingly, H3 and H4 were more accurately answered than H2.

This resulted in lower confidence interval for the workers who attempted H2 hierarchy questions. The paired t-test shows evidence to the reason for the low H2 percentages as the sets (H1, H2) were the least correctly answered questions among the three sets i.e., (H1,H2), (H1,H3) and (H1,H4). The $\mu$ difference for (H1, H2) is 32% where the $\mu$ difference for (H1, H3) being 21% followed by 25% for (H1, H4). As mentioned earlier, higher the $\mu$ difference lesser is the accuracy.
One of the reasons for H2 being answered the least accurate can be understood from figure 6.2 and figure 6.3, it illustrates that certain H2 images and their tags were more difficult in nature to understand or visualize compared to the latter H3 task. Analysis shows almost forty percent of the H2 images and tags belonged to this category of complexity. Another reason for H4 being better answered than H2 was because of the relatively less sample size of H4 vs. H2 of seventeen percent. The sample size for H2 is 402 compared and 350 for H2. Lesser number of workers responded to H4 tasks as a result of their complexity.
Figure 6.2. A sample of (H1, H2) task
Figure 6.3. A sample of (H1, H3) task
6.2. Hypothesis 2

Title: How the type of the question affected the performance of the worker

Input: Two inputs i.e. IMG2TAG and TAG2IMG are used as the main criteria for this hypothesis, each further classified into four hierarchies i.e. H1, H2, H3 and H4. H1 for IMG2TAG consisting of an aggregated list of H1 from IMG2TAG from ClixSense and NeoBux; H1 for TAG2IMG consisting of an aggregated list of H1 from TAG2IMG for ClixSense and NeoBux and similarly for H2, H3 and H4.

IMG2TAG:

\{(H1_{ClixSense}+H1_{NeoBux})_{IMG2TAG}, (H2_{ClixSense}+H2_{NeoBux})_{IMG2TAG}, (H3_{ClixSense}+H3_{NeoBux})_{IMG2TAG}\}.

TAG2IMG:

\{(H1_{ClixSense}+H1_{NeoBux})_{TAG2IMG}, (H2_{ClixSense}+H2_{NeoBux})_{TAG2IMG}, (H3_{ClixSense}+H3_{NeoBux})_{TAG2IMG}\}.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE</th>
<th>Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-IMG2TAG</td>
<td>604</td>
<td>0.8907</td>
<td>0.3122</td>
<td>0.0127</td>
<td>0.8907</td>
<td>(0.8658, 0.9157)</td>
</tr>
<tr>
<td>H2-IMG2TAG</td>
<td>202</td>
<td>0.5743</td>
<td>0.4957</td>
<td>0.0349</td>
<td>0.5743</td>
<td>(0.5055, 0.6430)</td>
</tr>
<tr>
<td>H3-IMG2TAG</td>
<td>202</td>
<td>0.7574</td>
<td>0.4297</td>
<td>0.0302</td>
<td>0.7574</td>
<td>(0.6978, 0.8170)</td>
</tr>
<tr>
<td>H4-IMG2TAG</td>
<td>200</td>
<td>0.56</td>
<td>0.4976</td>
<td>0.0352</td>
<td>0.56</td>
<td>(0.4906, 0.6294)</td>
</tr>
</tbody>
</table>

Table 6.2. One-Sample T: H1-IMG2TAG, H2-IMG2TAG, H3-IMG2TAG, H4-IMG2TAG
<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE</th>
<th>Mean 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-TAG2IMG</td>
<td>550</td>
<td>0.94</td>
<td>0.2377</td>
<td>0.0101</td>
<td>(0.9201, 0.9599)</td>
</tr>
<tr>
<td>H2-TAG2IMG</td>
<td>200</td>
<td>0.595</td>
<td>0.4921</td>
<td>0.0348</td>
<td>(0.5264, 0.6636)</td>
</tr>
<tr>
<td>H3-TAG2IMG</td>
<td>200</td>
<td>0.645</td>
<td>0.4797</td>
<td>0.0339</td>
<td>(0.5781, 0.7119)</td>
</tr>
<tr>
<td>H4-TAG2IMG</td>
<td>150</td>
<td>0.8</td>
<td>0.4013</td>
<td>0.0328</td>
<td>(0.7352, 0.8648)</td>
</tr>
</tbody>
</table>

Table 6.3. One-Sample T: H1-TAG2IMG, H2-TAG2IMG, H3-TAG2IMG, H4-TAG2IMG

Output: The findings for the second hypothesis show that TAG2IMG type of questions have been answered more accurately than the IMG2TAG type of questions for (H1, H2) and (H1, H4) but, for (H1, H3) the IMG2TAG type of questions were performed better. The mean variations for (H1, H2) and (H1, H4) are 3% and 16% respectively whereas for (H1, H3), IMG2TAG type of questions were 11% more accurately answered than TAG2IMG type of questions.
Figure 6.4. Pie Chart representation for H1- IMG2TAG vs TAG2IMG

Figure 6.5. Pie Chart representation for H2- IMG2TAG vs TAG2IMG
Figure 6.6. Pie Chart representation for H3- IMG2TAG vs TAG2IMG

Figure 6.7. Pie Chart representation for H4- IMG2TAG vs TAG2IMG
General human psychology could have been the reason for IMG2TAG type of questions to be responded better in the case of (H1, H3) tasks. As the hierarchy increased the difficulty of the questions increased phenomenally making it difficult to look at a tag and guess what the image could be. Certain times images help humans understand and visualize how something could look like or resemble. Also, the reason TAG2IMG questions were better answered than IMG2TAG questions might be just how TAG2IMG type of questions were designed where the worker was able to see the four image options leading to less confusion; pictures most of the times help for better understanding.

6.3. Hypothesis 3

Title: How did the two channels vary on overall in terms of their worker accuracy i.e., ClixSense vs NeoBux?

Input: The input for the hierarchy are the two channels i.e. ClixSense and NeoBux, each further classified into four hierarchies i.e. H1, H2, H3 and H4 aggregating all combinations including the types of questions. H1 for ClixSense consisted of an aggregated list of H1 from ClixSense including IMG2TAG and TAG2IMG; H1 for NeoBux consisting of an aggregated list of H1 from NeoBux including IMG2TAG and TAG2IMG, similarly for H2, H3 and H4.
ClixSense:

[{H1IMG2TAG+H1TAG2IMG,H2IMG2TAG+H2TAG2IMG,H3IMG2TAG+H3TAG2IMG,H4IMG2TAG
 +H4TAG2IMG}ClixSense]

NeoBux:

[{H1IMG2TAG+H1TAG2IMG,H2IMG2TAG+H2TAG2IMG,H3IMG2TAG+H3TAG2IMG,H4IMG2TAG
 +H4TAG2IMG}NeoBux]

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE</th>
<th>Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-ClixSense</td>
<td>578</td>
<td>0.9291</td>
<td>0.2569</td>
<td>0.0107</td>
<td>(0.9081, 0.9501)</td>
<td></td>
</tr>
<tr>
<td>H2-ClixSense</td>
<td>202</td>
<td>0.6238</td>
<td>0.4856</td>
<td>0.0342</td>
<td>(0.5564, 0.6911)</td>
<td></td>
</tr>
<tr>
<td>H3-ClixSense</td>
<td>201</td>
<td>0.7313</td>
<td>0.4444</td>
<td>0.0313</td>
<td>(0.6695, 0.7931)</td>
<td></td>
</tr>
<tr>
<td>H4-ClixSense</td>
<td>175</td>
<td>0.6914</td>
<td>0.4632</td>
<td>0.035</td>
<td>(0.6223, 0.7605)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4. One-Sample T: H1-ClixSense, H2-ClixSense, H3-ClixSense, H4-ClixSense

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE</th>
<th>Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-NeoBux</td>
<td>576</td>
<td>0.8993</td>
<td>0.3012</td>
<td>0.0125</td>
<td>(0.8747, 0.9240)</td>
<td></td>
</tr>
<tr>
<td>H2-NeoBux</td>
<td>200</td>
<td>0.545</td>
<td>0.4992</td>
<td>0.0353</td>
<td>(0.4754, 0.6146)</td>
<td></td>
</tr>
<tr>
<td>H3-NeoBux</td>
<td>201</td>
<td>0.6716</td>
<td>0.4708</td>
<td>0.0332</td>
<td>(0.6062, 0.7371)</td>
<td></td>
</tr>
<tr>
<td>H4-NeoBux</td>
<td>175</td>
<td>0.6343</td>
<td>0.483</td>
<td>0.0365</td>
<td>(0.5622, 0.7063)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.5. One-Sample T: H1-NeoBux, H2-NeoBux, H3-NeoBux, H4-NeoBux

**Output:** The results for this hypothesis show that ClixSense worker performance was more accurate and genuine when compared to NeoBux. The individual percentages for ClixSense and NeoBux are illustrated in the pie charts.
Figure 6.8. Pie Chart representation for ClixSense- H1 vs H2 vs H3 vs H4

Figure 6.9. Pie Chart representation for NeoBux- H1 vs H2 vs H3 vs H4
Figure 6.10. Pie Chart representation for H1- ClixSense vs NeoBux

Figure 6.11. Pie Chart representation for H2- ClixSense vs NeoBux
Figure 6.12. Pie Chart representation for H3 - ClixSense vs NeoBux

Figure 6.13. Pie Chart representation for H4 - ClixSense vs NeoBux
6.4. Hypothesis 4

*Title:* How did the tag genre affect the performance of the workers?

*Input:* The data used were the seven categories from ImageNet i.e. {Animal, Plant, Person, Fungus, Artifact, Sport, Geological Formation}. A combination of each category from both the channels including TAG2IMG and IMG2TAG and also a list of all those category questions used in all the three sets i.e. (H1, H2), (H1, H3), (H1, H4) were used as the input for this hypothesis. For example, the notation for the category animal is

Animal:

{Animal_{IMG2TAG} + Animal_{TAG2IMG} + Animal_{(H1, H2)} + Animal_{(H1, H3)} + Animal_{(H1, H4)}}_{ClixSense} +

{Animal_{IMG2TAG} + Animal_{TAG2IMG} + Animal_{(H1, H2)} + Animal_{(H1, H3)} + Animal_{(H1, H4)}}_{NeoBux}
Output: As all the filtering questions showed approximately the same mean values, the analysis of this hypothesis focused only on the real questions i.e. H2, H3 and H4. The findings show that the most accurately answered tasks belonged to the category Plant and the least accurately answered tasks belonged to the category Sport.

The order of the category that were the most accurately performed in descending order is {Plant, Geological Formation, Artifact, Person, Animal, Fungus, and Sport}. 

Table 6.6. One-Sample T: Animal-H1, Animal-H(X), Plant-H1, Plant-H(X), Person-H1, ...
Figure 6.14. Pie Chart representation for H1-
Animal vs Plant vs Person vs Fungus vs Artifact vs GF vs Sport

Figure 6.15. Pie Chart representation for H(X)-
Animal vs Plant vs Person vs Fungus vs Artifact vs GF vs Sport
The reason for Plant to have the most correctly answered tasks was probably because its tags/images were relatively simple to understand whereas Animal and Fungus category tasks were among the least correctly answered tasks as this category’s tags used nomenclature representation and scientific/biological names which are usually hard to understand unless the worker responding to them has prior knowledge in animal science. Sport category was probably confused with Person as images from Sport category are mostly picture of people engaging in different sports which could be misleading and confusing when the option provided tags like person and a sport.

6.5. Hypothesis 5

Title: How the geographical location of the worker affected his/her performance?

Input: To understanding the geographical location of the crowd whose responses were the most accurate was obtained by aggregating all of the responses from the two channels together. The responses from different locations that the workers responded from were aggregated. As a lot of countries that participated in this study have a very small sample set, all the countries were merged to their respective continents. i.e. {Africa, Asia, Australia, Europe, South America and North America}. The six continents that participated in this study have a wide variation in the sample size. Since the number of people responding from Africa, Australia, and South America was
considerably less we had to consolidate the data for this hypothesis by analyzing only Asia, Europe and North America. The sample sizes of Asia and Europe was 184 and North America significantly less at 31.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFRICA-H1</td>
<td>23</td>
<td>0.913</td>
<td>0.288</td>
<td>0.060</td>
<td>(0.7885, 1.0376)</td>
</tr>
<tr>
<td>AFRICA-H(X)</td>
<td>23</td>
<td>0.6957</td>
<td>0.4705</td>
<td>0.0981</td>
<td>(0.4922, 0.8991)</td>
</tr>
<tr>
<td>ASIA-H1</td>
<td>517</td>
<td>0.9207</td>
<td>0.2705</td>
<td>0.0119</td>
<td>(0.8973, 0.9441)</td>
</tr>
<tr>
<td>ASIA-H(X)</td>
<td>517</td>
<td>0.6557</td>
<td>0.4756</td>
<td>0.0209</td>
<td>(0.6146, 0.6968)</td>
</tr>
<tr>
<td>EUROPE-H1</td>
<td>502</td>
<td>0.9004</td>
<td>0.2998</td>
<td>0.0134</td>
<td>(0.8741, 0.9267)</td>
</tr>
<tr>
<td>EUROPE-H(X)</td>
<td>502</td>
<td>0.6295</td>
<td>0.4834</td>
<td>0.0216</td>
<td>(0.5871, 0.6719)</td>
</tr>
<tr>
<td>NORTH AMERICA-H1</td>
<td>95</td>
<td>0.9474</td>
<td>0.2245</td>
<td>0.023</td>
<td>(0.9016, 0.9931)</td>
</tr>
<tr>
<td>NORTH AMERICA-H(X)</td>
<td>95</td>
<td>0.7053</td>
<td>0.4583</td>
<td>0.047</td>
<td>(0.6119, 0.7986)</td>
</tr>
<tr>
<td>AUSTRALIA-H1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(1.0000, 1.0000)</td>
</tr>
<tr>
<td>AUSTRALIA-H(X)</td>
<td>5</td>
<td>0.6</td>
<td>0.548</td>
<td>0.245</td>
<td>(-0.080, 1.280)</td>
</tr>
<tr>
<td>SOUTH AMERICA-H1</td>
<td>12</td>
<td>0.9167</td>
<td>0.2887</td>
<td>0.0833</td>
<td>(0.7333, 1.1001)</td>
</tr>
<tr>
<td>SOUTH AMERICA-H(X)</td>
<td>12</td>
<td>0.667</td>
<td>0.492</td>
<td>0.142</td>
<td>(0.354, 0.980)</td>
</tr>
</tbody>
</table>

Table 6.7. One-Sample T: Africa-H1, Africa-H(X), Asia-H1, Asia-H(X), Europe-H1, ...

Output: The statistical findings clearly show that North America performed best followed by Africa, South America, Asia, Europe and Australia. But with the wide variations in the sample sizes, when considering the top three continents with the highest sample size, North America performed better. Therefore workers from North America performed best.
Figure 6.16. Pie Chart representation for H1:
Africa vs Asia vs Europe vs North America vs Australia vs South America

Figure 6.17. Pie Chart representation for H(X):
Africa vs Asia vs Europe vs North America vs Australia vs South America
Figure 6.18. Pie Chart representation for H1- Asia vs Europe vs North America

Figure 6.19. Pie Chart representation for H(X) - Asia vs Europe vs North America
Discussion

Qualitative evaluation of the workers performance across the two channels of CrowdFlower, ClixSense and NeoBux, had shown results that justify the objective of the experiment and my research question. The findings from the first hypothesis show that the workers responded to hierarchy one (H1), hierarchy three (H3) and hierarchy four (H4) images and tags better than hierarchy two (H2). Workers have responded to “choose the correct image for the corresponding tag” type of task better than “choose the correct tag for the corresponding image”. The study also shows that by aggregating all the results for all hierarchy tags and images including the two task designs, ClixSense performed better than NeoBux. The study also provides information about how workers from different locations performed i.e. workers performance across different geographical locations. For this experiment, workers from North America performed best.

The results and the analysis therefore prove it is very important for a job requester to select a crowdsourcing platform with high trust levels and efficient worker base for higher quality solutions. Kosinski, Yoram, Bachrach, Kasneci, Van-Gael and Graepel in [2] talk about only evaluating the
performance of workers on Amazon Mechanical Turk using an IQ questionnaire. Although there is lot of ongoing research on crowdsourcing and there has been a large amount of study done around the concept of crowdsourcing, the main focus has only been around the actual practice and concept of crowdsourcing and how workers in general performed on a particular crowdsourcing service. But no study has so far evaluated the performance of workers across platforms in order to understand how important it is to choose the correct crowdsourcing platform for quality solutions.

So far, maximum number of research papers on crowdsourcing target the advantages of crowdsourcing and the essence of time and cost efficiency targeting the crowd, but this paper argues the quality of these workers in various situations like varying degrees in difficulty levels of tasks without restricting workers only from specific geographical locations and offering a low incentive. In the Crowd IQ paper [2], workers responses were tested based on an IQ test with varying incentives being offered but the authors weren’t able to see if these workers had similar or different skill levels as the quality of the worker might not always differ with the incentive offered but would vary with their actual skill level and knowledge. Future work could combine these two studies to see if the IQ of the worker varied with the same incentive offered and different levels of difficulty tasks.

Moreover, the crowd IQ [2] paper focuses on the workers intelligence and this study evaluates the workers quality on different crowdsourcing
platforms in order to provide an evidence how important it is to choose the right crowdsourcing service from the large number of services being offered today as well as to prove that not all workers from every platform provide the same quality of responses.
Limitations

Although the study evaluates the performances of workers across different crowdsourcing channels, it limits to only four channels. There can be a wider understanding by launching tasks on multiple platforms.

The study focuses only on launching similar tasks across different channels due to constraints but rather could post a similar set of tasks offering a higher incentive. This would broaden the study by understanding if workers perform better when offered higher incentives and also be able to see the variations in performance.

Another approach to this study which could further be extended could be, launching tasks on different platforms with different worker skill levels to study if that is causing any kind of variations to their responses. This particular study restricts to using only workers with level one skill.
Conclusion

This study was conducted to evaluate the performances of workers across different crowdsourcing platforms. It used resources from CrowdFlower and ImageNet: a hierarchy based image database. The experiment was conducted on CrowdFlower using their developer API and scripting tools to design the tasks. The tasks are designed such that every question consists of a filtering question and a real question. The real question is designed using hierarchies H2, H3 and H4 whereas the filtering question is designed using H1 hierarchy images/tags. The tasks are an equal number of two types of questions i.e., given an image choose the right tag representing the image and given a tag choose the correct image for the given tag. Each task, irrelevant of the type of question consists of two sub-questions, a filtering and a real question both belonging to the same type of question. After receiving responses for the tasks, two channels i.e. Amazon Mechanical Turk and JunoWallet had to be dropped from the study as the responses were null from JunoWallet and very limited from AMT because of the low one cent incentive offered, leaving the study focused on the remaining two channels- ClixSense and NeoBux. Five hypotheses have been derived based on the responses i.e. how the four hierarchies affected the
performance of the workers, how the two types of questions affected their performance, how the performance of the workers varied across the two channels, how the tag genre affected the performance of the workers and, how the performance based on the geographical location of the worker affected the study. Minitab, a statistical software package, has been used to analyze the data. Two statistical methods were used; one-sample t-test and paired t-test.

The analysis shows that the hierarchy of the question affected the worker performance. But surprisingly H2 tasks were the least accurately answered followed by H3 and H4. The reason might be that certain images/tags in H3 were relatively less difficult than H2 as illustrated in figures 6.2 and 6.3. The TAG2IMG type of questions were better answered than IMG2TAG type of questions in both (H1, H2) and (H1, H4) whereas for (H1, H3) IMG2TAG were more accurate. It is always better to see the name of an object or thing and guess how it would look visually from the available options. Overall, ClixSense performed better than NeoBux. The maximum numbers of responses were from Asia, Europe and North America and only a few responses from other continents like Africa, Australia and South America. The maximum number of correctly answered responses came from North America. The last analysis showed that Plant category was the most correctly responded tag genre and Sport was the least correct answered.

So far, there has never been a study conducted to show evidence comparing performance of crowd/workers across different crowdsourcing
platforms. This study shows evidence that worker performances across different crowdsourcing platforms differ based on various factors depending on the platforms reputation and trust level of their workers. This study proves that the performance of workers across different platforms varies in terms of skills, geographical location, design of questions and channels. Henceforth the paper and the study provides strong evidence that when choosing a crowdsourcing platform, one has to be aware that the crowd IQ varies and the quality of their responses depends strongly on the channel that one chose.
Bibliography


Appendix

1. AMT: Amazon Mechanical Turk.
2. Channel: A website in which a contributor can access a task.
3. CML: CrowdFlower Markup Language. A design tool.
4. Dataset: The entire set of data for a project.
5. Filtering Question: Question used in every task to filter spam responses designed using hierarchy 1 images/tags.
11. IMG2TAG: Image to tag, choose the correct tag for the corresponding image.
13. Unit: One row in a job set.
14. Real Question: Tasks designed to evaluate worker performance using hierarchy 2, hierarchy 3 and hierarchy 4 images or tags.
15. TAG2IMG: Tag to image, choose the correct image for the corresponding tag.
16. Task: A subset of data from the whole dataset.
17. Test Questions: Training data created in the platform.

18. Worker: Person completing assignments online.