Bachelor Degree Project

UNIVERSITY OF SKÖVDE

VIABILITY OF HUMAN INTELLIGENCE TASKS AS A METHOD FOR PASSWORD CATEGORIZATION

Bachelor Degree Project in Information Technology
With a Specialisation in Network and System
Administration G2E, 22.5 HP
IT610G, Spring term 2018
2018-06-17

Christopher Palm
b15chrpa@student.his.se

Supervisor: Joakim Kävrestad
Examiner: Marcus Nohlberg

Abstract

This study investigates the viability of using Human Intelligence Tasks (HIT) in password categorization. To achieve this, this study constructs and performs a HIT experiment on the online crowdsourcing platform Amazon Mechanical Turks. The study performs the experiment on the site Amazon Mechanical Turks, and gathers data in the form of answers from the workers. A a mixed quantitative and qualitative analysis of the data is performed to investigate on the workers ability to derive the categories of passwords from different categories and difficulties. The study results indicate that HIT workers seem to be unable to reliable categorize more complex passwords correctly, compared to more common and simple passwords. With this result, the study concludes that the quality and reliability of HIT password categorization is lower than would be required to make HIT a valid method for password categorization. The study ends with a discussion on how and why this may be the case and briefly discuss on how the HIT task might be changed in future development to increase its viability.

Keywords: Human Intelligence Tasks, HIT, Password categorization

1 Introduction

Research into how people create passwords, and what type of content they structure these from, can assist a multitude of areas. Such as IT security and IT forensics. In IT security, password categorization can give indications on which are the most common types of categories used by users, which could show where to put effort in creating guidelines and assistance for users to increase the strength of their passwords in these categories. In IT forensics it could be used to indicate common strategies and show where to place most effort in developing tools that aims to attack passwords in a more effective way by specializing the attack to the most likely categories (Kävrestad, Eriksson, & Nohlberg, 2018).

There exist several theoretical methods to perform password categorization. From theoretical future computational methods of AI and neural networks, to more contemporary computational methods of programs and scripts, that are customized for password categorization, to manual categorization of passwords by a person.

To categorize password manually is today seen as the most realistic method, however it is slow and requires going through thousands of passwords and categorize them into a password model. This can then be used to determined which strategy was used for the creation of a password, and which strategies may be most prevalent in different user groups (Kävrestad et al, 2018).

In recent times, Human Intelligence Tasks (HIT) platforms have been launched. HITs are small tasks performed by humans that require a low amount of training or time to perform. HIT has been used in areas such as image and video classification (Deng, Dong, Socher, 2009), and in research its gaining usage in social research, such as behavioural research (Crump, McDonnel, Gureckis, 2013). HIT offers an intriguing new choice of method to perform password categorization with human workers. This study believe that categorizing passwords holds similarities to image, video, and text classification. These are areas that computer systems are improving in, but still have a hard time performing (Zhu, Yan, Bao, Yang, Xu, 2014). HIT workers however, are human, and seems able to perform relatively well (Snow, O'Connor, Jurafsky, and Ng, 2008). These HIT crowdsourcing platforms can theoretically be used to perform password categorisation.

While manual password categorization is limited to the speed of the researcher or worker performing the categorization, and a small number of active workers at a time, HIT offers the ability to utilize a large pool of workers from all over the world. If HIT can provide good quality and reliability in its categorizations, it could theoretically be used to perform a larger amount of password categorizations quicker than manual categorization, at a relatively cheap price. This study seeks to test if HIT is a viable method by testing its quality and cost in performing password categorization.

2 Background

This section will provide basic information that should increase knowledge and aid in understanding the different aspects of the project.

2.1 Password categorization

To categorize passwords into different categories depending on how they are created, and what content is used, would require going through data samples in form of existing passwords. These can then be checked to the categories in a password model. Through this categorization, it can then be determined which strategies are used for each password, and which strategies are most prevalent. The obvious way to perform this categorization is for the researcher to manually go through passwords and match them to categories. However, there exists other methods to perform this task.

2.2 Methods to perform password categorization

This section will provide a quick rundown of the different methods that were identified to be possible methods of which to perform password categorization, and will examine some positive and negative aspects of each.

2.2.1 AI.

Utilizing Artificial Intelligence (AI) is a theoretical application of processing this kind of data. Modern forms of this are advanced versions of today's neural networks or AI. These could theoretically be configured and trained into being able to categorize passwords according to models. To create an AI, or more realistically, a neural network that is able to perform this kind of work are assumed to be unavailable and too costly to develop at this time.

2.2.2 Computers/Programming.

While the AI or neural networks methods are mostly theoretical future possibilities, computers, programs, or scripting, exists as more contemporary possibilities, that can utilize computational power to perform tasks. These computers and programs could theoretically be structured and configured with specific instructions to be specialized in password categorization. In theory, this method could be used in password categorization since it resembles other current work in object classification and categorization, where these systems are trained to identify objects (Bermeitinger, Christoforaki,Donig, Handschuh, 2017).

However, this also provides the problem of the cost to develop these systems, since no known program or systems currently exists that are designed to categorize passwords. Creation of such programs and systems would require advanced development, and in more advanced and complex systems, training of these systems. A simpler version would be scripts or simple programs that follow rigid coding and instructions to be able to classify password where elements in the text string (password) would be matched to instructions and structures which would make the program categorize a password to a specific strategy.

The problem with these systems are that they would need to be configured using set instructions in how a password is categorized into each category. For example, a script might be coded to classify passwords with the category Words by searching for passwords that mainly are built from using dictionary words. While this script would be able to categorize passwords that are constructed using dictionary words, it wouldn't be able to categorize the password if the word were misspelled, leading to a misscategorization.

2.2.3 Manually by researcher.

While using programs or systems to categorize passwords would theoretically be fast, some data and tasks are structured or contains data that computer systems and programs are unfit or unsuited for: these may at times require customization and development that make the endeavour less, or not at all, cost effective to develop (Franklin, Kossmann, Kraska, Ramesh, Xin, 2011). Password categorization is an area that is assumed to be hard to develop these systems for. The alternative to using these systems is to use the existing method of using trained people to perform categorization.

The method of using trained people for password categorization is in this study referred to as manual categorization of passwords. In this method, the person or persons that wish to perform a password categorization project are the ones that will be categorizing the passwords. This method ensures that the person that perform the categorisation process presumably has, or will gain, a good understanding on how to properly categorize passwords.

The quality of manual categorization is therefore assumed to be considered of high quality and reliability, since the person presumably has a good understanding on how to perform the categorization through research or training in the area. The person is also assumed to have a willingness to categorize the passwords and take the time needed to properly perform this task. These aspects might therefore lead to a higher quality of password categorization than utilizing outside workers not as familiar to password categorization.

The negative aspect of this method is that it relies upon a single, or limited amount of people, which limits the number of passwords that can be classified during a set span of time to the number of people involved in the project. It also takes away the time of the person performing the project or research from other areas this person could be doing. Furthermore, the work is of a repetitive nature, and when performing a large-scale categorization project may take a significant time. It also relies upon the decisions of a single or limited amount of people to decide which category a password belongs to, which includes the risk of biased results depending on the background and knowledge of the person.

2.2.4 Crowdsourcing through HIT.

As an alternative to having a researcher manually categorizing passwords, or directly hire and train workers specifically for the task which might be expensive, there exists the option of crowdsourcing the task online through Human Intelligence Tasks. While manual categorisation by the researcher are limited to the speed, working hours, knowledge, and demographic of an individual researcher, HIT could allow for password categorization research to utilize a workforce pool from around the world (Ross, Irani, Silberman, Zaldivar, Tomlinson, 2016). This would theoretically allow for a method that contains the quality and reliability of using humans for the categorization, while freeing up the researcher’s time, and might be of a lower cost that to specifically hire people to perform the task. This also might provide the benefit of utilizing the knowledge of a wider range of workers from different demographics instead of relying on the knowledge by a single researcher. By utilizing these HIT tasks, projects could theoretically go through a large number of passwords for a relative cheap cost.

2.3 Human Intelligence Tasks

Data unfit for, or hard to structure for computational automation and processing, requires humans which can easily process some kinds of data that computers and programs have a hard time to process, such as picture or video categorization and classification (Zhu et al, 2014). To this end Human Intelligence Tasks can be utilized to go through and process this kind of data. HITs are often structured as small tasks that are quick and easy for humans to understand and perform. HITs can be found online on platforms that creates the infrastructure necessary for these tasks (Franklin et al, 2011). These platforms are often seen as a sort of crowdsourcing platforms (Difallah, Catasta, Demartini, Ipeirotis, Gudré-Mauroux, 2011) where the site creates an environment that connect people (workers) willing to perform small, short timed tasks for a small monetary reward, with work providers (requesters) that provide these tasks.

HIT’s comes in many different forms. Popular examples of HITs are tasks such as surveys and psychological research (Casler, Bickel, Hackett, 2013), transcribing conversations in audio and video onto text, translations of texts that computers find hard to grammatically translate properly, and of special interest to this study, the categorizing and sorting of pictures and video based on elements in that media (Deng et al, 2009).

Utilizing HIT in password categorization could theoretically be a method that uses the advantages of human manual categorization, while limiting some of the drawback of only using a limited pool of workers and cost of hiring dedicated personnel. However, due to the nature of HITs, as short and relative quick tasks, they require quick and easily understandable training material for the workers in HIT (Crump 2013). This training material would have to be constructed and implemented in such a way that it can provide sufficient understanding for a HIT worker to be able to be provided with a password and be able to categorize which strategy it is constructed from.

In theory this method should provide similar results to that of the manual categorization, with the benefit of having access to a large pool of workers that enables the researcher to perform other tasks.

2.4 Amazon Mechanical Turks

One of the platform for HITs are the website Amazon Mechanical Turks, which is often called mturk or AMT. The name is a reference to a (fake) mechanical chess machine created in the 18th century. The machine was actually controlled by a person (*What is Amazon Mechanical Turks n.d.*) and only giving the illusion of automation.

The work providers on AMT, called Requesters, will create tasks of different natures and complexity depending on the type of data and aim of the overall job. The Requester can offer different monetary rewards for completions of these HIT assignments depending on how complicated and time consuming the task they wished performed are (Amazon Mechanical Turk Pricing n.d.). Each HIT contains small tasks, with larger jobs often being broken up into several HITs which all forms a HIT batch. A HIT batch may be, for example, to classify 1000 images, where each image is considered its own HIT assignment that together forms a batch. The requesters can also decide on how many assignments are done on each HIT, with each assignment being done by an unique worker. The requester then publishes the HIT task on the online platform which places it into the pool of available tasks. The workers can then pick through and decide on which HIT to work on from this pool.

2.5 HIT workers

Compared to other conventional methods of performing studies, and data gathering in research fields, online HIT platforms has the advantage of being able to utilize a workforce from all over the globe instead of local resources (Casler et al. 2013). While newer research into worker demographics were not found, earlier studies, for example Ipeirotis (2010), suggest that most workers are from either the US, or India, and that the majority reasons for performing HITs differ depending on where the worker is from. Workers from India were more likely to indicate that HIT is a significant portion of their monetary income, than compared to workers from the US which indicated more towards treating it as a leisure activity or receiving pocket money.

2.6 Quality of HIT

Past studies of HIT workers quality and reliability that were found seems mostly focused on social and psychological aspects and results of utilizing HIT in research. For these areas Paolacci and Chandler (2014), argues that past research indicates that HIT worker data quality is reliable in this area, compared to conventional methods for gathering data. For data that relies solely on objectively correct answers, Aker, El-Haj, Albakour, and Kruschwitz (2012) indicates that good quality can be achieved. For area's a bit more close to the type of HIT this study will use, Snow et al (2008) tests annotation labels of text from HIT workers, and compare them to expert labels, and concluded that a high quality of the results can be achieved. Image and data classification were a type of HIT task which were observed to be utilized on the AMT platform, however, no direct research on this type of HIT and expected result quality from these HITs were found by this study.

Common areas that are of similar nature of this study, are areas such as image, video (Deng et al, 2009), and data classification, language annotations (Snow et al, 2008). From these areas, it's believed that HIT could be used for password categorization.

2.7 Password categories

Password strategies dictates how a password is constructed and what it contains. A password will be structured differently depending on the strategies that's utilized in its creation. Different strategies can be grouped up into different categories depending on their structure and content. This study bases its categories on the password model created by Kävrestad et al, (2018). The following sections will provide a quick description of each category used in this study:

2.7.1 Biographical

Passwords that falls into the biographical category are passwords that in some form contains words or other elements that contains personal information about the creator of the password. This could be names of the person, family, or pets of the creator. The password could also contain personal information in form of street address, interests, or events involving the user, such as dates, or locations.

2.7.2 Phrases

A password falls into this category if it contains words that either form a standard language sentence or contains a collection of words.

2.7.3 Words

Passwords under this category are constructed using a simple word for a password. The word can be from different languages.

2.7.4 Words in Words

This category contains passwords that utilize a strategy of placing words inside other words which breaks up the first word into two parts that are then placed in front of and behind the second word. An example would be using the words ‘Experiment, Car’ in the form ‘ExperiCarment’.

2.7.5 Leet Speak

This is the usage of replacing letters in a word or sentence with numerical or special characters that in some form are similar to the replaced letter. Examples would be replacing ‘i’ or ‘L’ with 1 or |, ‘b’ with 6, etc. This can be used in different scales, from replacing single letters in a word, to every character in a phrase.

2.7.6 Mnemonic

Passwords under this category are based upon a phrase, and then converting this phrase into a password that seemingly contains a random string of characters. The common strategy that's provided by Kuo, Romanosky, & Cranor (2006), is to use the first letter in each word. This password can then further be strengthened by using different structures, capitalization, and characters for each first letter. For example, utilizing leet speak for some of the letters. The main structure of mnemonic passwords used by this study is the large capitalization of letters of regular words in a phrase, and lower capitalization of articles and conjunctions etc, in that phrase.

2.7.7 Pattern

Passwords under this category can seem similar to passwords in the alphanumeric category, since at first glance they may seem to be randomly chosen. However, these passwords are created by using some kind of pattern, either physical or otherwise, to create and structure the password. For example, the string ‘qwerty’ is created by using a straight line on the keyboard.

2.7.8 Alphanumeric (Random)

This category indicates that a password is of a random nature and does not use strategies from other categories. Instead, these are created through randomly choosing different characters. These are often considered the most secure since they do not hold any information that can be linked, and therefore guessed, to the person creating it, and is not built from words, which can be attacked using dictionary attacks.

In this study, this category is a gathering of the ‘alphanumeric’, ‘alphabetical’, ‘numerical’, and ‘special characters’ categories, in the model. These are grouped into a single category in this study for simplicity since categorizing whether a password only utilizes small alphabetical characters, or only use special characters, etc, is not of interest, and therefore the categories are grouped into a category that includes all passwords that does not fall into any of the other categories.

2.8 Common strategies

Riley (2006) performed a study of common practices that were used when creating passwords, while this study by does not define as many categories as this project does, drawing parallels between those categories and this projects categories, the results indicates that most users will utilize some form of biographical data to create passwords, while replacing characters with special characters (Leetspeak) were uncommon. The website Geeknoob.com (2012) provides a list of 1000 of some of the most common passwords. Going through this list shows that the most common strategies of the passwords are to utilize words, names, and basic patterns, with some phrases, and utilization of basic leetspeak. This indicates that these are the most common strategies used for the majority of user passwords are of a simple nature.

2.9 Related Works

This study uses the password model developed by Kävrestad et al, (2018) to create and define categories for passwords that are used to create material for the study, and act as a basis for the strategies used to create passwords.

3 Problem background

This section will provide the aim the study and how it seeks to achieve this aim, some delimitations in the study, and what it seeks to achieve in the password categorization research area.

3.1 Aim

Using HIT for password categorization is, according to the author’s knowledge, a method that has not been tried. However, before HIT can be used for this purpose we would need to know if the method provides reliable results of a good enough quality to make its output reliable. Therefore, the aim of this study is to evaluate if utilizing HIT as a method to categorize passwords seems to be viable for future research and projects in regard to password categorization, and furthermore look into what factors, challenges, and considerations needs to be thought of when performing password categorization with HIT. For this evaluation to be performed, two main factors were identified to be of importance, the quality of the workers categorization of passwords, and the cost of performing such a categorization project in HIT.

3.2 Research Questions

To achieve the aim of the project, two research questions were created. The first research question is focused on the quality and reliability of the answers provided by workers in HIT password categorization. The second question looks at the costs surrounding this method of using HIT to give an idea of expected costs of performing password categorisation through HIT crowdsourcing.

The research questions of the project are:

1. How well are HIT workers able to derive the strategy used in the creation of a password?
2. What are the costs of utilizing HIT in password categorization?

3.3 Ethics

While earlier research in HIT and AMT indicates that majority of workers does not regard HITs as a main income source, more recent studies indicates a shift, and now shows that a larger portion of workers treat HITs as a monetary income source instead of way to spend free time or gain a small amount of pocket money (Hara, Adams, Milland, Savage, Callison-Burch, Bigham, 2008)

This brings up the question of the ethics and morals of utilizing HITs for data processing that could otherwise be placed on an employee or other such workers or researcher. This is a topic that's outside the scope of the study and won’t be explored and is instead left for other research to debate on.

3.4 Delimitation

This section will take up some delimitations in the study.

Rewards

How different sized rewards on a HIT impacts the quality of answers, and the draw of workers to the task, are aspects in HIT that have been researched in other areas, but may have a different impact on password categorization HIT results. These were aspects that were originally intended to be explored in this research, however this study will not be able to look into these effects in more detail, mainly due to time constraints, since this is research that would build upon this study's finding, and therefore the effects of these aspects are only theorized in this report.

Study cost
Due to costs of performing a real-world experiment on HIT, the study will be limited in the amount of real world HIT testing that can be done. A best effort will be done to create a task and instructions-set for the HIT by looking into similar areas that utilize HIT and how these are constructed to develop a rudimentary HIT task for password categorization. 3rd party feedback and a pilot test will be used for small scale refinements of the task. To create a fully developed and tested structure and instructions-set for this type of HIT would require several iterations and real-world testing which would be beyond this studies scope and ability. It will instead aim to create an indication of what might be the viability of this method and if and how further development into a properly developed task might affect this viability.

Removed categories
In the model from Kävrestad et al, (2018) there exists two main categories, Machine generated, and user created passwords, with the user subcategories of biographical and neutral passwords. While some passwords could often be assumed to be created by a person, for example passwords with names etc, others are close to impossible to determine since they can just contain a word or text string a machine could have been configured to generate. Furthermore, a random string of characters that a person created to have a strong password are hard to distinguish from a random computer created passwords.

With this issue also comes the issue of the two user sub-categories. Biographical passwords were found to be possible to be derived to have been used in the creation of the passwords since these will contain information that relates to people, times, or places. However, neutral passwords do not contain this link to the user.

Due to this, the main categories of machine vs user created, as well as neutral passwords, were removed from the categorization task since it was determined that it was not possible to reliably classify passwords into these. For the biographical category, it was decided that it should remain to be able to classify passwords that contains information that indicated links to people, places, dates, etc.

3.5 Expected results

It's hoped that through this study, the method of using HIT for password categorization can be explored to see if it seems to be a method that can provide reliable results of good quality, and as such, act as a viable method alongside manual password categorization. Through this study, its hoped that future research and projects can use the insight and experience gained from this study to better develop and perform password categorisation through HIT.

4 Method

This chapter will take up the method design and planned methods for performing the experiment necessary to answer the research questions. Each section will first talk about what it seeks to test, and then detail on how this is then planned to be analysed.

4.1 Mixed design

For this study, a quantitative experiment was chosen for data gathering. While using a real-world setting does not provide the strict control of all variables that would be preferable in an experiment method according to Wohlin, Runeson, Höst, Ohlsson, Regnell, & Wesslen, (2012) and 'Robson & McCartan' (2016), its deemed to be the closest appropriate method for this study to gather data.

For the analysis and discussion in later parts, a mixed method of both quantitative and qualitative elements will be used. The experiment will be conducted on the AMT platform and aim to be structured and perform as close to how a real project would be performed.

The data provided from the experiment in form of worker answers to different passwords can then be analysed both quantitatively and qualitatively. The study will first look into the result quantitatively to see how many of the answers provided were correct and how many were wrong. This analysis will then be used to provide a basis for a qualitative analysis and discussion on which categories are most problematic for workers, and if HIT seems to be viable method for password categorization projects.

Past research by Riley (2006) shows that a person is more likely to create passwords that contains personal information, and an observation of the 1000 most common passwords from Geeknoob (2012), shows that most passwords are of a simple structure. This creates an assumption for real word password categorization projects, that most passwords would contain a majority of passwords of a simple and biographical nature. However, since this study is more aimed to test if HIT is a viable method for password categorization, a focus will be placed on using more difficult passwords to test the viability if more difficult passwords would be encountered. This is assumed to have a negative effect on the overall number of correct answers and the quantitative analysis. Therefore, more focus and weight will be placed on the qualitative analysis of the results to see if workers are able to categorize edge cases of passwords difficulty.

4.2 Quality

For this section, the project seeks to provide a basis for answering the first research question; "*How well are HIT workers able to derive the strategy used in the creation of a password?*"

With the aim of the study being to evaluate if HIT is a viable alternative method for password categorization, the quality of the answers of workers in password categorization is of importance. To be able to measure the quality of the HIT workers password categorizations, the project seeks to structure the experiment in such a way as to provide the possibility of a quantitative evaluation of the answers provided by workers.

The passwords, while taking inspiration, and seeking to emulate real passwords, will be created with specific, predefined, categories in mind. Therefore, each password will have a specific set of categories that are defined as being the correct answer, and therefore expected to be the provided answer from the HIT workers in the quantitative analysis of the answers. Answers outside these categories, or lacking required categories, will be considered to be incorrect.

Of particular note to the study is the ability on AMT of having multiple workers providing answers to the same assignment, this can be utilized to reach a consensus of which category a password belongs to. This tactic would theoretically allow for higher reliability of the majority answers since it would no longer rely on any one worker providing fully accurate answers, but instead relies on the answer that's chosen by most workers (Kamar & Horvits. 2012).

This study assumes that utilizing a form of a consensus method will be a requirement for HIT in password categorization. Therefore, this will be included as a method for determining if HIT can provide the correct category for a password. For this experiment however, its not known how many workers per password are needed to reach a reliable consensus. It's also theorized that different categories might require a higher number of answers to form a reliable consensus. This is an aspect that will be looked into and discussed but is an aspect that needs to be developed and looked into specifically in future research is results indicates that HIT could be a viable method. The pilot experiment that will be conducted during the study will be used to give an indication on which number of worker assignments per password will be used in this study.

4.3 Cost

To be able to answer the second research question; "*What are the costs of utilizing HIT in password categorization?"* the project will examine the following two metrics:

1. Time it takes to create HIT task.
2. Monetary cost of implementing a HIT task on AMT.

To answer the question of monetary cost of implementing password categorization HITs on AMT, this study will take on a mainly qualitative discussion monetary cost. Data from the experiment will be gathered, as well as pricing information from AMT to provide a discussion on the cost of this type of HIT, from the cost of creating the material needed for this type of HIT, to possible future development of the material, and what other aspects that might affect cost. Data on price information on HIT, as well as information from previous studies and research on how rewards might affect HIT answers, will be gathered during the experiment to act as a basis from which to derive a discussion of different strategies of implementing HIT. For example, use of multiple workers per password for consensus, and theorize on how these might be used or changed to gain or remove certain elements in this type of HIT.

An aspect that is deemed as of importance for the study to look into, is cost per password. The study will utilize research as well as results from the experiment to theorize on different aspects and strategies of creating the HIT, including number of passwords per task and reward for these tasks, and how they may affect this metric.

4.4 Passwords

The most obvious way to provide passwords for the study to perform the testing on is to use real world passwords. This could be argued to be the preferable way to test the quality of the HIT answers since real world data would be used and tested against. Such passwords could easily be gathered by collecting passwords from leaked password databases that are accessible on the internet. These real-world passwords could then be chosen and used in the experiment.

However, to use real world passwords could in of itself create a validity threat to the results. This is due to the fact that the study does not know the objectively correct creation strategy that were used when creating the passwords. This creates the possibility of what this study might categorize as the correct creation strategy of a password may be different to how a worker might categorise it as. This would further also only allow for testing of different password structures to those of the found passwords. Geeknoob (2012) provides a list of the 1000 most common passwords. This list shows that the most common passwords are of a relative simple structure, and theoretically, if real world passwords were used, the risk is that most passwords used in the experiment may be of a relative simple structure, which this study believes would make the categorization relative easy to perform.

Therefore, to be able to objectively know which categories were used in the creation of a password, and to test the ability of HIT workers to categorize different password structures in each category, the passwords used in the experiment will not be provided from leaked databases of real world passwords, and are instead created specifically for this study with specific categories, and different structures, in mind for each password.

4.5 Validity threats

Section provides validity threats identified to be applicable to this experiment.

Wohlin et al. (2012) provides the following validity threats that were identified to be of relevance to this study.

- **Single study group.** The workers can't be chosen, the workers that perform the test are those that chose to perform the HIT, therefore we can't guarantee that the workers are a true representative group of workers on the site.

- **History.** Different days and times of the year may provide different number of workers. Since drawing power to the HIT tasks are not directly tested and measured, the study performed the experiment to overlap with a weekend since it was speculated that more workers would have free time to perform tasks at that point in time.

- **Testing**. Due to the nature of the site the experiment is conducted on, workers that are involved in the planed pilot test may also be involved in the main experiment. This might affect how they approach the main experiment. Since this is something that may exist in real world application of this type of HIT, this validity threat is not actively countered.

- **Low statistical power**. To counter this threat, several passwords will be created for each category. Then 20 workers will be allowed to perform categorization on each password, to enable the study to perform an analysis of the overall quality of answers, that can be expected of workers for that password and categorization. The threat is still prevalent however due to the low number of passwords used per category, however this will remain since adding more passwords per category would significantly increase costs.

- Robson and McCartan (2016) also provides the threat of ***Ambiguity about casual direction*** which may include how any monetary reward change might affect several variables in the experiment, such as how many people are drawn, quality of the people drawn, and if they perform better. It's also unknown if monetary compensation affects the time spent on a password and how that may affect quality. This validity threat is not explored fully. The study looks at the quality of the data provided by the workers, but does not explicitly test how the monetary reward might affect other variables, a rudimentary analysis will be performed by checking the different in draw of workers between the pilot test and main experiment. Instead, deeper study into this aspect will be left for potentially future studies.

5 Development process of the HIT experiment

This chapter will first, in the section ‘Development overview’, give an overarching introduction into the different stages taken in the process of developing the projects HIT task. this is then followed by the ‘Development process’ section where each part is brought up and provides more in-depth information about the different actions and decisions that were made in each one. The final section ‘Experiment process’, will go through the process of performing the experiment.

5.1 Development overview

Development of the task were broken up into several specific parts. Each part focused into different aspects of the development of the HIT experiment. While the process is showed linearly in Figure 5-1 Development Overview, this is only to represent when most of the core development in an area were done, and changes were made in an earlier part’s areas during the whole process if a later part of the process indicated that changes needed to be made.

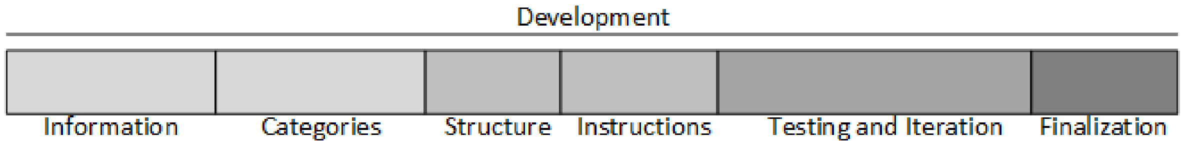


FIGURE 5-1 DEVELOPMENT OVERVIEW

The **Information** part of the process aims to gather basic data on different aspects of HIT and how HIT functioned on the AMT platform. This information will then to be used in future parts and decision making.

In the **Categories** part, the study will go through the Kävrestad et al (2018) model and decide on which categories will be used in the experiment and how these are to be structured. Then the decision on how to create and structure passwords for each category will be decided.

The **Structure** part of the process aims to choose, and develop, the structure of the HIT task that will be used in the experiment, by utilizing information gathered in the Information part on how different research areas and real-world projects structure their HITs.

The **Instructions** part of the process aims to create a set of instructions for each category and on how to identify if a password belongs in this category. These instructions will then be used by the workers to categorise passwords.

In **Testing and Iteration,** the study will perform some testing of, and make changes to, the HIT structure and instructions created in the previous parts, by using feedback from 3rd party individuals, and results from a planned pilot experiment.

In **Finalization** the HIT task should be complete with no changes needed for structure or instructions. The HIT task will be finalized, and preparation will be made for performing the experiment.

5.2 Development process

This section will go through the development process in each part more in depth, and take up some decisions made in each part.

5.2.1 Information

This study first began by gathering some basic information, and observe, on how HIT was performed on the AMT platform, so as to gain an understanding on how to structure the experiment and its HIT task. This included aspects such as how AMT specifically handles HIT, how different areas handle and structure their HIT tasks, and how cost and fees are handled. Then information was gathered on how these different aspects affects each other, such as how assignment rewards affects quality and amount of workers drawn to a HIT etc. This information was then used in future parts to serve as a basis for decision making in how to structure the HIT task, as well as what rewards and costs that were to be used. Kävrestad et al (2018) were also used to gather information on how to identify and categorize different categories of password strategies.

5.2.2 Categories

To decide on what categories were to be used for the study, the password model from Kävrestad et al (2018) were used for the creation of the different categories of password strategies. Most of these remained the same as the model, however some changes were made. The categories "machine" and "neutral" were, as brought up earlier in this study, removed since these were deemed impossible or extremely hard to derive from a password. Furthermore, the model utilizes subcategories of what types of characters are used in a password, these were not deemed necessary to be categorized in the experiment. This were due to only the overarching creation strategy of a password being of interest for this experiment, while knowing if a password only contains a specific type of characters is of little interest. If this would be an aspect of interest in future projects, its a categorization that can easily be done faster and more cheaply by computers.

Passwords tested per category

For the experiment, it was decided that 10 passwords would be tested in each category for a total of 80 passwords. This were mainly due to budgetary concerns since adding more passwords per category would considerable increase the cost of the experiment.

Each password will therefore have a "main" category. This category will be considered the correct answer to the password during quantitative analysis. Some password may contain secondary categories that will be required in the answer to mark the answer as correct. These passwords are mainly those that uses strategies that requires underlying strategies to exist. An example is the Leet Speak category, which requires the password to use some form of words or phrases.

Passwords structure consideration

With one aim of the study being to test the ability of HIT workers to categorize passwords with structures that's 'harder' to categorize than common passwords, the study uses passwords that would be assumed to be harder to categorize correctly than the norm for the majority of passwords in the real world.

Each category will therefore have a range of difficulty through different structures and methods depending on the password. How the difficulty is increased for a password depends on the category. Some passwords will simply have an increase the difficulty of the strategy structure, while some may be created with a specific aspect in mind. Passwords created for the category “Words” and “Phrases” for example, will take on an increased difficulty by having some words and phrases chosen to be of a different language than English. This is to test if workers can still recognize if the password is a word or a phrase. Other passwords in these categories will be created from uncommon words.

A general method to increase difficulty will be the capitalization of letters in different areas and amounts to see if this makes the workers unable to realise that the passwords formes words or corresponds to the specific strategy used to create it. An added challenge to some passwords in categories that uses words in some form, will be the use of small amounts of random letters and numbers in front of, or behind, the words.

In other categories, the passwords structure will be changed, or the strategy specific aspect increased. Examples of these are the Mnemonic and Leet Speak categories. In Mnemonic passwords, the basic strategy is to capitalize some letters, here the challenge is increased by deviating from the basic strategy by using more or fewer capitalized letters, or simply adding Leet Speak. In the Leet Speak category, the challenge is increased mostly by a more heavily reliance on the categories main aspect, in this case an increase in amounts of letters that are changed to numbers or special characters.

Mnemonic

Mnemonical passwords were created manly by following what Kuo (2006) calls a basic transformation, which is to select the first letter in the words. The common structure of large capitalization of letters of regular words, and lower capitalization of articles and conjunctions etc, in that phrase, were the basis of these passwords. Changes were made in structure with different passwords to see how well workers could derive the mnemonic category depending on different capitalisation structuring, if the workers relied on the basic method, and if problems occurred with mnemonic passwords that doesn't follow this structure.

Characters and Alphanumeric categories

The Alphanumerical categories were first used to be chosen not only for completely random passwords, but for all passwords that contained some part in its structure built up of random characters or numbers that were not used by the words or other main structure that were of another category. This were however seen as redundant since the category that were primarily used to create the passwords are of interest. It was therefore decided that the Alphanumerical categorization were to be only used to classify completely random passwords, and allow other categories to contain some random characters. To better convey this, the category was renamed in the HIT task to “Random”.

Phrases
Three phrases were created to utilize some form of non english element. Two phrases were made in spanish and german respectively, and a third utilized the word ananas, which is a common word in other languages for Pineapple. The Spanish phrase further used odd capitalization to further increase difficulty.

5.2.3 Structure

For choosing the overall structure of the HIT task that will be used in the experiment, the study looked into how how other areas structured their HITs. The areas looked into most, and that were most influential, were the area of image classification, where workers were asked to identify contents in images. This area was believed to be the most similar to what this study aimed to test in its experiment. Therefore the structure of the HIT used were modeled after those HITs.

The general structure of the HIT was divided into three main parts. The first section which contains the instructions, explains to the worker on how to perform the task and identify the elements that's tested in the tested object. The second section is simply the object that's tested, this object was identified to often be placed directly above the third section, which provides the worker with a selection of choices to select as answers.

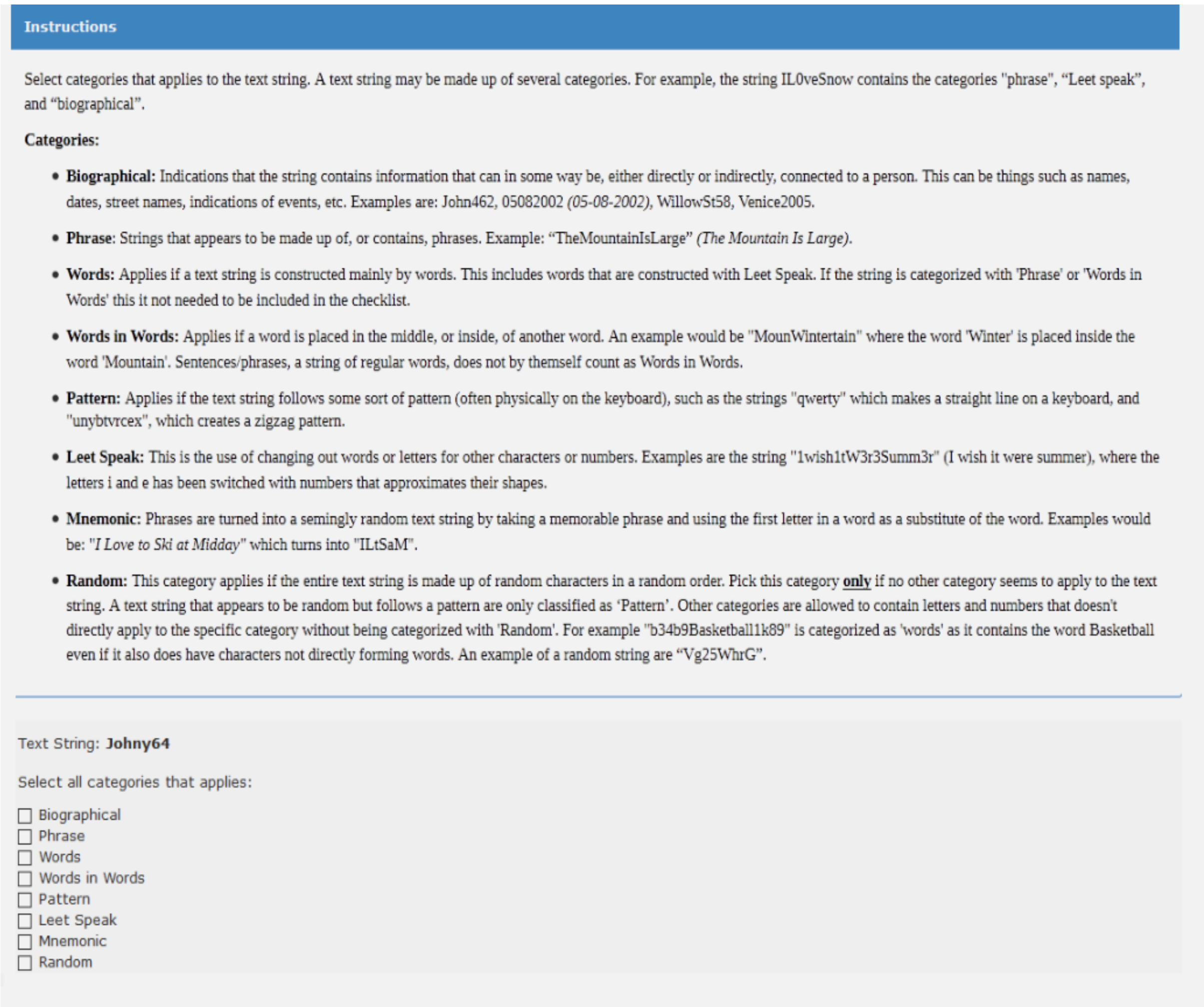
It was decided that the HIT task would use a checkbox system for its answers. These were presented in a line under the password, where each of the strategies were listed. The worker would then be prompted by the task to choose all the categories that applied to the password. Multiple answers can be checked, since some passwords contains several answers. The simplicity of the structure for listing all categories as separate were done since the study believed that simplicity in the structure would be the most efficient structure.

No checks were made by the task to see if any conflicting categories were chosen, for example if Random were checked together with other categories, or if there appeared to be categories missing, for example if only Leetspeak were checked, which needs a secondary category of either, Words, Phrase, or Words in Words to be used.

For the creation of the HIT task object itself on AMT, its estimated that roughly 5h were directly spent in creating it, mainly these were planning of the design, and due to lack of knowledge of the language used in its creation, leading to several versions being scrapped and redesigned.

The time it takes for workers to complete a task are of interest in calculating rewards for assignments involving multiple passwords. Theoretically you would want as many passwords per assignment as possible to lower the cost per password. However, adding too many passwords to an assignment increases the time it takes to perform the task, which might lower workers incentive to perform the tasks since the reward for the time spent may be seen as too low. To better gather and analyse the time it takes for workers to complete each task, the number of passwords per assignment will be limited to one (1) to be able to calculate a rough average of time it takes to perform the task.

The structure chosen for the HIT task in this experiment can be found in figure 5-2 HIT Structure.



Instructions

Select categories that applies to the text string. A text string may be made up of several categories. For example, the string IL0veSnow contains the categories "phrase", "Leet speak", and "biographical".

Categories:

- **Biographical:** Indications that the string contains information that can in some way be, either directly or indirectly, connected to a person. This can be things such as names, dates, street names, indications of events, etc. Examples are: John462, 05082002 *(05-08-2002)*, WillowSt58, Venice2005.
- **Phrase**: Strings that appears to be made up of, or contains, phrases. Example: "TheMountainIsLarge" *(The Mountain Is Large)*.
- **Words:** Applies if a text string is constructed mainly by words. This includes words that are constructed with Leet Speak. If the string is categorized with 'Phrase' or 'Words in Words' this it not needed to be included in the checklist.
- **Words in Words:** Applies if a word is placed in the middle, or inside, of another word. An example would be "MounWintertain" where the word 'Winter' is placed inside the word 'Mountain'. Sentences/phrases, a string of regular words, does not by themself count as Words in Words.
- **Pattern:** Applies if the text string follows some sort of pattern (often physically on the keyboard), such as the strings "qwerty" which makes a straight line on a keyboard, and "unybtvrcex", which creates a zigzag pattern.
- **Leet Speak:** This is the use of changing out words or letters for other characters or numbers. Examples are the string "1wish1tW3r3Summ3r" (I wish it were summer), where the letters i and e has been switched with numbers that approximates their shapes.
- **Mnemonic:** Phrases are turned into a semingly random text string by taking a memorable phrase and using the first letter in a word as a substitute of the word. Examples would be: "*I Love to Ski at Midday*" which turns into "ILtSaM".
- **Random:** This category applies if the entire text string is made up of random characters in a random order. Pick this category **only** if no other category seems to apply to the text string. A text string that appears to be random but follows a pattern are only classified as 'Pattern'. Other categories are allowed to contain letters and numbers that doesn't directly apply to the specific category without being categorized with 'Random'. For example "b34b9Basketball1k89" is categorized as 'words' as it contains the word Basketball even if it also does have characters not directly forming words. An example of a random string are "Vg25WhrG".

Text String: **Johny64**

Select all categories that applies:

☐ Biographical
☐ Phrase
☐ Words
☐ Words in Words
☐ Pattern
☐ Leet Speak
☐ Mnemonic
☐ Random

FIGURE 5-2 HIT STRUCTURE

5.2.4 Instructions

The instructions were created by utilizing the knowledge gained by this study from its works with the Kävrestad et al (2018) model and its descriptions of its categories. The aim of the instructions were to be as descriptive of the categories as possible while remaining short, direct, and easy to understand. Examples would also be provided to show simple example structures of passwords in each category. Early iterations did not contain the Mnemonic category since it was yet not known with which strategy the mnemonic passwords were to be created from.

The instructions were fine tuned by periodically requesting feedback from third party individuals. After changes to the instruction material from this process were complete, the instructions were used in a pilot experiment on the AMT to see if the instructions were functional for HIT.

For the creation of the HIT task and instructions its estimated that around 11h were directly spend on different aspects of the instructions, including early drafts and refinements. Several refinements were done after being provided third party feedback, as well as changes made after the pilot experiment.

The first set of instructions made can be viewed in Figure 5-3 Early Instructions.

Select categories that applies to the text string. A text string may be made up of several categories. For example, the string ILoveSnow contains the categories "phrase", "words", and “biographical”.

Main categories:

- **Biographical:** Indications that the string contains information that can be connected to a person, place, or time. This can be things such as names, years, street names, etc. Examples are: John462, 05082002, WillowSt52.

- **Phrase**: Strings that appears to be made up of, or contains, phrases. Example: “TheMountainIsLarge”.

- **Words:** Applies if a text strings contains words. This includes words that are constructed with Leet Speak.

- **Words in Words:** Applies if a word is placed in another word. To find these types it's recommended that if a word is found, checking if the few letters before and after the word seems to make up another word. An example would be "MounWintertain" where the word 'Winter' is placed inside the word 'Mountain'.

- **Pattern:** Applies if the text string follows some sort of pattern, such as the string "qwerty" which makes a straight line on a keyboard. Another example would be "unybtvrcex", which creates a zigzag pattern.

- **Leet Speak:** This is the use of changing out words or letters for other characters or numbers. Examples are the string "1wish1tW3r3Summ3r" (I wish it were summer), where the letters i and e has been switched with numbers that approximates their shapes.

- **Random:** Pick this category only if no other category seems to apply to the text string. Other categories are allowed to contain letters and numbers that doesn't directly apply to the category. For example "Basketball1k89" are categorized as 'words' as it contains the word Basketball. An example of a random string are “Vg25WhrG”

FIGURE 5-3 EARLY INSTRUCTIONS

5.2.5 Testing and iteration

Testing and iteration changes were made continuously throughout the creation of the structure and instructions. The process was done using third party individuals for the first versions of instructions, and then when it were deemed that the HIT task were mainly finished, it were used in a pilot experiment on the AMT platform. This were done to get indications on possible complications of the instructions, and to see if planned administrative configurations such as the reward, would provide enough workers to conduct the experiment.

The pilot test was conducted during the same week days period that the experiment itself were planned to be conducted on. The test was set to run for 3 days and to provide a 0.03$ monetary reward. The test would utilize about 50 passwords that could each be performed by 50 unique workers. The variable of 50 assignments per password were set to get an idea of where to put the number of workers per password for the experiment.

The test concluded with 1,268 out of 2,500 password strings being categorized, with 57 unique workers. Out of the 57 workers, 28 categorized under 10 passwords, 5 categorized between 10-20, 2 between 21-30, 2 between 31-40, 4 between 41-49, and 16 hit the cap of 50. On average, a password was categorized by 25 workers.

The pilot test indicated that most workers would either perform a relative high amount of password categorizations or only perform a few. During observations of the data, it was decided that the number of around 20 assignments per password would be used to give an indication of whether or not the workers could classify a password through a consensus answer.

From third party feedback, and results of the pilot experiments, some changes were made to the instructions. For the Words category, it was observed that it would frequently cause problems in conformity with workers answers since some workers would or would not mark down “Words” when categorizing a password as Biographical, Pattern, or Words in Words. To seek to conform as many answers from the workers in these categories as possible, its instructions were modified with the stipulation to not mark it down if the password contained the categories “Words in words”, or “Phrase”. Some categories also had their descriptions extended and changes made to its examples.

The random category instruction was made more clear that other categoriess were allowed to contain random elements, and the category was only to be marked if the whole password were random.

The iterations made after the pilot experiment and feedback were incorporated and resulted in the instructions that were used in the final experiment. The final instructions can be found in Figure 5-4 Final Instructions.

Select categories that applies to the text string. A text string may be made up of several categories. For example, the string IL0veSnow contains the categories "phrase", “Leet speak”, and “biographical”.

Categories:

- **Biographical:** Indications that the string contains information that can in some way be, either directly or indirectly, connected to a person. This can be things such as names, dates, street names, indications of events, etc. Examples are: John462, 05082002 *(05-08-2002),* WillowSt58, Venice2005.

- **Phrase**: Strings that appears to be made up of, or contains, phrases. Example: “TheMountainIsLarge” *(The Mountain Is Large).*

- **Words:** Applies if a text string is constructed mainly by words. This includes words that are constructed with Leet Speak. If the string is categorized with 'Phrase' or 'Words in Words' this it not needed to be included in the checklist.

- **Words in Words:** Applies if a word is placed in the middle, or inside, of another word. An example would be "MounWintertain" where the word 'Winter' is placed inside the word 'Mountain'. Sentences/phrases, a string of regular words, does not by themself count as Words in Words.

- **Pattern:** Applies if the text string follows some sort of pattern (often physically on the keyboard), such as the strings "qwerty" which makes a straight line on a keyboard, and "unybtvrcex", which creates a zigzag pattern.

- **Leet Speak:** This is the use of changing out words or letters for other characters or numbers. Examples are the string "1wish1tW3r3Summ3r" (I wish it were summer), where the letters i and e has been switched with numbers that approximates their shapes.

- **Mnemonic:** Phrases are turned into a semingly random text string by taking a memorable phrase and using the first letter in a word as a substitute of the word. Examples would be: "*I Love to Ski at Midday"* which turns into "ILtSaM".

- **Random:** This category applies if the entire text string is made up of random characters in a random order. Pick this category **only** if no other category seems to apply to the text string. A text string that appears to be random but follows a pattern are only classified as ‘Pattern’. Other categories are allowed to contain letters and numbers that doesn't directly apply to the specific category without being categorized with 'Random'. For example "b34b9Basketball1k89" is categorized as 'words' as it contains the word Basketball even if it also does have characters not directly forming words. An example of a random string are “Vg25WhrG”.

FIGURE 5-4 FINAL INSTRUCTIONS

5.2.6 Finalization

With the pilot test showing fewer responses to the task than would be preferable, especially with the planned increase from 50 passwords to 80 passwords, as well as due to time constraints, the choices were made to increase the reward per assignment from 0.03$ to 0.05$. With this, the assumption was that enough workers would be drawn to the HIT to complete it in the allotted time frame.

Observations of the answers indicated that around 20 answers would be sufficient to indicate some form of consensus of an answer to a password, it was decided to place the workers per assignment value to 20 workers per password. This is theorized to be enough to form a consensus on an answer.

5.3 Experiment process

When the instructions and structure were finalized, and administrative settings, such as reward etc, were set after observations from the pilot experiment, the experiment were constructed and performed on the AMT platform during a weekend. This time period was chosen since it's assumed that this may be the time period where most workers has free time to perform HIT. The experiment ran for approximately 5 days, an extension on the planed 3. This were done since a small number of passwords were hard to get all its assignments completed. Its theorized that this is since some workers may just see that there is a low amount of HIT assignments available and prefer not to complete such a low number of assignments.

The administrative settings were set to one password tested per assignment, with a 0.05$ reward per assignment, with 20 unique assignments per password. AMT masters qualification were required for workers to performs the HIT. AMT masters are workers that has fulfilled a requirement of a history of high quality work on AMT (What are masters, n.d.).

For comparison of how many passwords a manual categorisation method could produce in the time it took to create the HIT task, a timer was set, and passwords were categorized through a list of passwords that were randomly takes from a file of leaked passwords. The passwords and answers were both in a spreadsheet and answer were written down beside the password. An advantage that this structure hade was evident quite quickly, the lack of "overhead" between categorizations. In HIT, workers have to submit each password when they have marked down the answer and wait for the next HIT to load. Such overhead does not exist for manual categorization, leading to an increased ability of processing passwords. The manual categorization was timed for 30 minutes.

6 Data

This section will provide the data gained from the experiment and that will be used in section 7 to perform analysis.

6.1 HIT Task creation

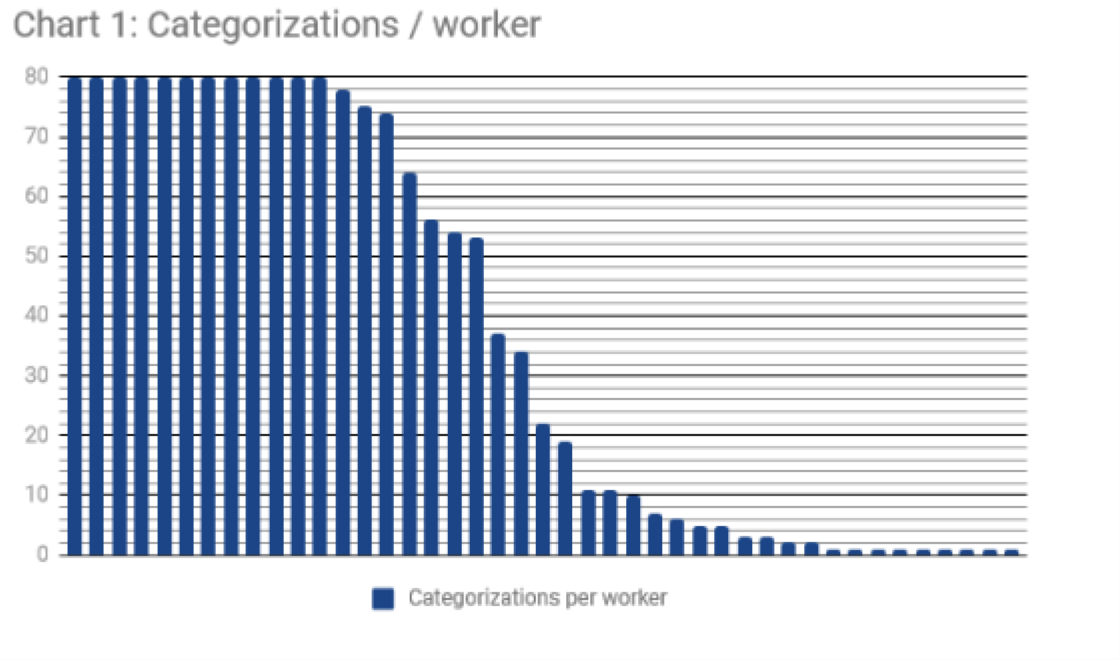
For the creation of the HIT task and its instructions its estimated that around 16h were directly spend on different aspects of the instructions and creation of the HIT tasks structure and code.

6.2 Manual password categorization

For the manual password categorization test, 147 passwords were categorized manually in 30 minutes, resulting an average of 12.24 seconds per password. For the time it took to create the HIT task, manual categorization could theoretically categorize roughly 4700 passwords.

6.3 Categorizations per Worker

The number of unique workers that worked on the HIT in the experiment were 43. Chart 1: *Categorizations / worker* show how many passwords were categorized by each worker



6.4 Worker answers

The results of the experiment in form of number of correct answers is provided in Table 6-1 *Answer Data*:

Category	Accepted answers	Accepted %/total	Wrong answers	Wrong %/total	Highest correct	Highest %/total	Missing	Missing %
Phrases	125	62.50%	75	37.50%	8	80.00%	10	5.00%
Biographic	90	45.00%	110	55.00%	5	50.00%	18	9.00%

al								
Words In Words	75	37.50%	125	62.50%	6	60.00%	31	15.50%
Words	68	34.00%	132	66.00%	5	50.00%	5	2.50%
Mnemonic	56	28.00%	144	72.00%	1	10.00%	0	0.00%
Pattern	65	32.50%	155	77.50%	2	20.00%	0	0.00%
Leet Speak	80	40.00%	120	60.00%	4	40.00%	27	13.50%
Random	131	65.50%	49	24.50%	9	90.00%	0	0.00%
Total	690	43.13%	910	56.88%	40	50.00%	91	5.69%

FIGURE 6-1 ANSWER DATA

6.5 Workers work time per password

While the experiment allowed for 2h to pass for a worker to complete an assignment, this were mainly due to keeping a wide margin of error, and only forcefully ending a worker's assignment if its considered that the worker has essentially stopped working on the assignment. This data is assumed to not be all that reliable since this study is unable to track why a worker may take longer time for a password compared to others. However, it is of interest for the discussion of HIT as a method, and how to structure these tasks.

For the analysis, an assumption will be made that no password should realistically take more than 200 seconds to categorize if the worker is focused on the task. Therefore, any data over 200 seconds will be disregarded since its then assumed that the worker performed other tasks that interfered with the time to complete an assignment or simply stopped working for a moment.

With this delimitation, the data from the experiment showed that 87% of all assignments were completed within the 200 seconds timeframe, with 75%, and 70.5% being completed within 100, and 60 seconds respectively. The average time spent on an assignment that falls within this delimitation, is roughly 28 seconds.

6.6 Monetary cost of implementing HIT

This section will bring up the monetary cost factors surrounding HIT that were found to be relevant and included when performing password categorization through HIT.

It was found that the cost depends on two main components:

- Monetary reward to workers
- AMT fees based on worker reward + extra fees.

The percentage fees of AMT that are applicable to the task are as following:

- **20% fee on reward** - This fee is always applied upon the reward provided to workers.
- **5% Master Qualification** - This fee is applied if the HIT is marked to require workers fulfils AMT master qualification.
- **20% on HITs with 10+ workers per assignments** - This additional fee applies if a HIT assignment is set to be performed by 10 or more workers.

7 Data Analysis and Results

This section will analyse the data gathered during the study and experiment. Including worker answers, as well as provide some information about cost.

7.1 How worker answers were analysed

When going through the data from the experiment, it was observed that, at times, workers would provide a wrong answer, but one that at times seemed to indicate that the worker might have understood how the password were created, but missed to provide the fully correct answers, or misunderstood the structure of the answers the task wished to receive. Most evident of this were answers in the LeetSpeak category, were some workers would simple answer “Leetspeak” but not provide other categoriess that were needed for a correct answer. These are presented in the data results as ‘Accepted answers’. These won't be counted in the analysis that looks into the chosen correct answers through consensus, but will be used in parts of the analysis and discussion to discuss if HIT seems to be a viable method, and what parts of categories workers seems to have the most problem with.

For seeing if the workers can provide the correct answer for a password through a form of consensus, the study selects the most provided answer as being considered the answer provided by HIT. For this selection, a wholly qualitative analysis is done on the answers by only allowing the main intended categories to be included in the answer. This data is represented in the data results as ‘Highest Correct’.

7.1.1 Phrase

The accepted answers provided for this category were lower than expected at only 62.5% of worker answers being included into an accepted answer. However, for the highest provided answer, 8 of the 10 passwords were correctly categorized. Most problems in this category seems to be from phrases that does not fall into normal English phrases.

The Spanish phrase were correctly categorized through highest answer, however it had far more answer combinations than other phrases. The Ananas phrase seems to have been mistaken for the name “Anna” since a large portion of the answers were for a Biographical categorization. Neither the German or Ananas phrases were correctly categorized through highest provided answer.

A common problem with phrases seemed to be that workers would mark down answers with Words in Words or just words. Workers failing to mark down the main category is a recurring theme in answers. For Words in Words, it's theorized by this study that this might be due to workers understanding the Words in Words category slightly wrong, and not limiting the category to words that are placed inside other words, and therefore breaking up another word.

7.1.2 Biographical

Biographical passwords also received lower accepted answers than expected, with only 45% of answers being of an accepted categorization, and only half of the passwords were correctly categorized by highest answer. However, common Western names and words were often correctly answered, while non-western, uncommon words, or words with odd capitalization, resulted in more incorrect answers and answer combinations. The Biographical phrase that was used was also only categorized as a phrase in 40% of the answers, this might however be due to previously mentioned problem with the HIT task lacking good enough structure and instructions.

7.1.3 Names

Workers seemed to have an easier time recognizing western style names, compared to non-western style names. This seems to be supported by the demographic of AMT which is mostly US based, with a large sub-demographic of Indian workers. Since most HITs on AMT seems to be English based (none were seen to be made from another language), it reinforces the theory that western/English based names will be easiest for workers to recognize.

Exception to this could be names or words in other languages that holds similarities to western words. One of the Words category passwords (Sekunde21), were wrongly categorized by a large margin (60%), where workers categorized it as Biographical. This seems to be due to Sekunde possible being a nickname for the South-Asian/Persian name Sikandar. This indicates that by utilizing HIT, problems might occur with consensus and conformity of answers for passwords that share similarities in English and other, mainly Indian, languages. This is a both a source of problems, and source of advantage, by utilizing HIT with different demographics of workers. The possibility does exist to limit the workers to a specific demographic, however this adds extra costs

7.1.4 Dates

Passwords that only contains a numerical date seems to be a problem for HIT workers. Both shortened (dd/mm/yy) and longer (dd/mm/yyyy) numerical dates resulted in a high categorization as Random. A factor that may also be of interest here might be differences in accustomization, where the US use mm/dd/yyyy. Its theorized that workers might not always read the first numbers as days but months, and not recognize the password as a date.

7.1.5 Words in Words

The answers provided from this category were at the same time below, and above, what was expected. With only 37.5% of the provided answers being accepted. However, by using the highest answer as the provided answer, the passwords were correctly categorized 60% of the times. This seems to indicate that a higher division exists in this category, but that this is spread out towards several incorrect combinations of answers instead of a small number of incorrect combinations, and that by utilizing a consensus method for answers can provide more reliable results.

7.1.6 Words

The category contained wo passwords (Centurion, Lunchroom), which were structured with a simpler structure (Common words, common capitalization), that is assumed to be the structure the vast majority of passwords would be created from (Geeknoob 2012). These passwords acted as a form of control to see how well simple words were categorized. The password Centurion were categorized as a word by 80% of workers, while Lunchroom were categorized by 45%. However, an aspect of the password Lunchroom that wasn't thought of at the time it were created as a control, were that it could be considered a phrase (Lunch Room), which 30% of workers did.

The results of the experiment showed that with the change to different structures in and around words, the overall accepted answers in the category were 34%. This indicates that while workers have an easy time categorizing a simple word, if other aspects and methods are involved, HIT workers will have a harder time to accurately categorize the password.

7.1.7 Mnemonic

The mnemonic category was expected to be hard for HIT workers to categorize. Only 28% of answers in this category were acceptable, and with only one of the 10 passwords were correctly categorized through highest answer. When the use of the common basic mnemonic strategy of capitalizing the letters were dropped, or lowered in regularity, the number of correct answers for these passwords were low to non-existent. The category that the mnemonic passwords seems to be mistaken for most of the time is Random, which is expected.

This indicates that while workers might in some cases recognise the basic mnemonic strategy, any more advanced strategy, or contains a structure more complex than the basic one, will result in low correct categorizations of these passwords. That workers relied on capitalization were further strengthened by having a Random category password that used a structure that looked similar to basic mnemonic passwords, which resulted in small amount of mnemonic answers for this password.

7.1.8 Pattern

Passwords that followed a pattern were categorized with acceptable answers 32.5% of the time, but with only two of the ten passwords being correctly categorized using highest answer, and the category Random being prevalent in answers for these passwords. This indicates that HIT workers are not able to categorize pattern based passwords.

Relative easy patterns, such as a straight line on the keyboard, numerical patterns, or patterns close to each other on the keyboard had higher correct answers than those that followed more complex patterns.

To categorize these patterns, a worker would have to either be able to see the pattern in the password, or use aid or tools to see/notice a pattern. In the instructions, it was hinted that the keyboard could be used for this to look for patterns, but this would require the worker to take the time to trace the passwords path on the keyboard. This is something that is assumed to not be done in most cases. Its theorized by this study that this categories answers might be able to be improved by the inclusion in graphical elements to the task, such as an automatically generated picture that shows the path the password takes on the keyboard to improve answers in this category.

7.1.9 Leet Speak

The passwords created for this category utilized heavy use of leetspeak. While other categories had words that might contain a small amount of leetspeak, this category used password that were heavily modified with this content. Answers by workers in this category were deemed accepted 40% of the time, with the most numerous answer being correct 40% of the time. This indicates that workers are to some extent, able to recognize words and phrases that uses heavy use of leetspeak, but not to a satisfactory extent.

7.1.10 Random

The random category was the category with the best result with 72% accepted answers. However, this is to be expected, especially since several passwords from other categories were often miscategorized as belonging in this category. Some passwords were constructed to show similar patterns to other categories, and this may have lowered the number of correct answer a bit.

An interesting aspect that were discovered from the random password answers, were answers that included categories that seems obvious that they don’t fit with the password. Going through the answers of other categories shows similar answers that doesn't seem to fit. Through a quick observation of the data and answers, its believed that there may be 1-3 answers for each password that seems obviously wrong and questionably, and does not relate to the answer norms. Its theorized that these may be junk answers, from workers that doesn't understand how to perform password categorization, and are from the group of workers that only performs a few categorizations before stopping. Another theory is that some of these answers are from workers that performs the work with little regards for quality, with the hope that the answer is accepted and paid by the requester. These are however, the types of answers that use consensus/majority-answer methods aims to counter.

7.2 HIT Cost

This study has gathered some data on what costs are involved with performing password categorization with HIT on the platform ATM, and what the cost per password is when performing this method. The data shows that the cost to have a HIT task performed is mainly derived from the reward that is placed upon an assignment. The cost per password is then dependent on how many passwords are categorized per assignment. This section will perform some discussion on how these different aspects may interact, using past research material into these areas, and what considerations were found to be of importance when establishing the reward and cost.

The most important cost for password categorisation projects in HIT is the cost per password, which is found by the study to be affected by three main factors:

- Cost per assignment
- Passwords per assignment
- Unique workers per assignment

The cost per assignment is based on the reward that is placed on an assignment since all applicable AMT fees are based on the reward. The fees are the base 20% fee, >9 workers/assignment fee of 20%, and 5% for using AMT masters. This study assumes that for password categorization, more than 9 unique workers per password seems to be needed to reliable categorize passwords, as well as using AMT masters since past studies shows that requiring workers to be masters may increase quality (Peer, Vosgerau, Acquisti, 2013). This creates an administrative AMT fee of 45% on the reward placed on an assignment.

When setting the reward, consideration should be placed on the complexity of the task, as well as expected time it will take to perform the task. The data from the experiment indicates that on average, a worker takes about 28 seconds to categorize a password. Manual categorization of passwords takes about 12 seconds on average. This study believes that this difference is mainly from difference in knowledge and familiarization of password categorization, as well as different workers having different speeds. However, one aspect that is believed to be a considerable factor is that only one password is used per assignment, while the manual categorization utilized a list of passwords. The HIT worker therefore has to spend time submitting and reloading the site between each password. This is believed to contribute to a higher time. By utilizing several passwords per assignment, its believed that workers will be able to perform categorization faster without the overhead.

Further consideration should be the number of workers that the project wants to be drawn to the task. A higher reward increases worker draw but does not necessarily influence quality (Mason & Watts, 2009; Buhrmester, Kwang, Gosling, 2011; Crumpet al, 2013).

A large increase in the cost per password is from the number of unique workers per assignment. This value depends on what number of workers are deemed to be needed for each assignment to form a reliable consensus of the answers. This study's HIT task is structed in such a way that is deemed non-optimal and requires a higher number than 9 workers per assignment to form a somewhat usable consensus. This adds not only the extra 20% fee, which by itself is a large increase in cost, but each unique worker for an assignment multiplies the cost of the assignment by the number of unique workers. To go from 1 to 10 workers/assignment therefore increases the cost per password by 10 if only one password is used per assignment. This can however be offset by using more passwords in each assignment, which decreases

The formula for a projects cost is:
Project Cost = (((Total Passwords ÷ Passwords/Assignments) · Workers/Assignments) · (reward + fee))

The formula for cost per password is:
Cost/Password = ((((Total Passwords ÷ Passwords/Assignments) ·Workers/Assignments) · (reward + fee))) ÷ Total Passwords)

8 Discussion

This section will provide discussion about the results of the analyzation and data in previous sections.

8.1 HIT’s viability as a method for password categorization

This section will discuss if HIT seems to be a viable method for password categorization.

8.1.1 HIT task structure

During the study, it was discovered that the chosen structure of the HIT task which were based on image classification HITs, might not be a suitable structure for password categorization HITs. The developed structure of having a checklist of all categories that each can be either the primary category, or act as secondary categories, might not give enough structure for the workers to understand how these categories interact.

Answers by workers included categories that does not have any relations or combinations to each other, and several answers seems to indicate that the worker understood the categories the password fell under, yet they didn’t select all categories that applied to the password. This is believed to significantly affect the overall quality of the answers negatively, especially for consensus forming.

Its theorized that a HIT task that has more structure in the answer selection portion of the task may be more suitable. For example, the task might first ask for the primary category, followed by a list of content categories that only shows the categories that are applicable to the main category. This might result in higher reliability since it limits the possibility of mistakes from the workers due to misunderstanding of instructions, as well as might assist in consensus since the number of combinations of unique answers workers can provide would be limited.

Viability

The study used harder structures and content for its passwords than is assumed to be the norm in a real-world setting. This was to test the quality of the workers answers when presented with a more challenging password. The results indicate that while workers are mostly able to categorize common, or simple structured, passwords with a good reliability, their ability to categorize a password correctly drops when presented with passwords that are structured in more complex ways, or are built from uncommon content and languages. In more complex categories, such as passwords that utilize uncommon mnemonic structures, or a high use of Leet Speak, the number of correct answers drops quickly.

Furthermore, the cost of utilizing HIT for password categorization is considered too high. This is mainly due to the high number of workers needed per assignment, to gather enough answers to gain a reliable consensus with the current HIT task.

However, utilizing consensus to provide the HIT worker group answer seems to be the right method to increase reliability, since some passwords had an overall low accepted answer rate, but were still able to be correctly categorized through the highest provided answer. With further development of the HIT tasks structure, which could lower the number of workers needed per assignment, could bring costs down significantly. Furthermore, recently there has been efforts made to create tools and support systems (Castano, Ferrara, Montanelli, 2016) to assist in this area. These systems aim to automate these types of consensus methods, and accepting or disregarding answers by workers outside the norm of the group answers. These systems might be of interest in future development and research into HIT password categorization.

8.2 Ethical aspects

Primarily, the ethical issue of HIT password categorization is the use of HIT itself. Research indicates that a larger portion of workers consider HIT their main source of income (Hara et al, 2008), and this coupled with the low amount of money provided for the task raises concerns. If projects in this area were to be performed, this study believes that the reward placed on the assignments should be thought through with regards to workers earning compared to the work provided, and aim for a fair compromise between low cost and worker earnings.

8.3 Societal relevance

By performing password categorization projects, we can learn which categories and strategies that users most often utilize to create their passwords. This could bring insight into an area important for security. In IT forensics, Kävrestad et al (2018) argues that password categorization could be helpful to know the common categories used by users, as well as provides ideas for its use in looking into the common categories in specific groups of users, so to be able to develop better tools that target specific categories.

8.4 Scientific relevance

Research into areas linked to passwords and security is always of interest. By researching into these areas, we can better understand how users create passwords, and structure our works with strengthening these areas better. This study faced some limitations due to budget concerns, especially with the structure of its HIT task. It's hoped that by learning from the development process of this study, future work into this area of research can learn from its insights and develop better methods and structures that could aid in increasing the viability of HIT as a method.

By indicating that workers in HIT having a hard time with complex passwords compared to more common ones, the study provides a basis for comparison to other methods of password categorization. It can be used for example to compare if manual categorization suffers from similar limitations.

8.5 Conclusion

During this study, the viability of HIT as a method for password categorization has been tested by developing material in the form of a HIT task, that were then used in a real-world experiment. The study tested passwords that were considered harder than the norm, and the results indicate that HIT as a method can't perform password categorization reliable if it encounters passwords that is of a harder challenge than common passwords. Therefore, this study concludes that the quality and reliability of HIT as password categorization method is not good enough to be a viable method for password categorization until the reliability can be increased.

8.6 Future Work

This study performed an experiment to see how well HIT workers can derive password categories from a provided password. An extension of this work would then be to test how well a trained manual worker, (researcher, or hired worker specially trained), can categorize passwords, and then compare the results and expected costs to this method, so to identify if the problem with HIT password categorization is mainly from the problems with the structure, or if they inherently are unable to categorize as reliable categorize hard passwords as compared to a manual worker.

The structure of the HIT task used in the study is believed to be flawed by aiming for too much simplicity in the answer structure. Future research should look into proper development of a HIT task that's more structured and effective for HIT in password categorization than the one created for the experiment.

REFERENCES

Aker, M., El-Haj, M., Albakour, M., Kruschwitz, U., (2012). Assessing crowdsourcing quality through objective tasks. In LREC, pages 1456–1461. Citeseer. Retrieved 22-03-2018 from https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.352.2947

Amazon Mechanical Turk Pricing, (n.d.), Amazon Mechanical Turk Pricing. Mturk.com, Retrieved 17-04-2018 from https://requester.mturk.com/pricing

Bermeitinger, B., Christoforaki, M., Donig, S., Handschuh, S., (2017) Object Classification in Images of Neoclassical Artifacts Using Deep Learning, Published in Digital Humanities, Retrieved 13-05-2018 from https://arxiv.org/abs/1710.04943v1

Buhrmester. M., Kwang. T., Gosling. D. S., (2011), Amazon’s Mechanical Turk, A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science, p. 3-5, doi:10.1177/1745691610393980

Casler, K., Bickel, L., Hackett, E., (2013), Separate but equal? A comparison of participants and data gathered via AMazon’s MTurk, social media, and face-to-face behavioral testing. Elsevier, Computers in Human Behavior, p. 2156-2160, doi:10.1016/j.chb.2013.05.009

Castano, S., Ferrara, A., Montanelli, S., (2016) Designing Crowdsourcing Tasks with Consensus Constraints, In 2016 International Conference on Collaboration Technologies and Systems, doi:10.1109/CTS.2016.0035

Crump. J. C. M., McDonnel. J. V., Gureckis. T. M., (2013), Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS ONE, doi:10.1371/journal.pone.0057410

Deng, J., Dong, W., Socher, R., (2009), ImageNet: A large scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, pages 248-255, IEEE, doi:10.1109/CVPR.2009.5206848

Difallah, E. D., Catasta. M., Demartini. G., Ipeirotis. G. P., Gudré-Mauroux. P., (2015), The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. WWW’15, p. 238-247. doi:10.1145/2736277.2741685

Franklin M. J., Kossmann D. , Kraska, T., Ramesh, S., Xin, R,. (2011) CrowdDB: Answering Queries with Crowdsourcing. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD ’11, pages 61–72, New York, NY, USA, 2011. ACM, doi:10.1145/1989323.1989331

Geeknoob, (2012), 1000 Most Common Passwords – May be yours one of them?, Retrieved 13-05-2018 from https://www.geeknoob.com/1000-most-common-passwords.html

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., Bigham, J, P., (2018), A Data-Driven Analysis of Workers Earnings on Amazon Mechanical Turk, In CHI '18 Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems Paper No. 449, doi:10.1145/3173574.3174023

Ipeirotis, G, P., (2010). Analyzing the Amazon Mechanical Turk marketplace. XRDS: Crossroads, The ACM Magazine for Students, Volume 17, 16-21. doi:10.1145/1869086.1869094

Kamar, E., Horvitz, E., (2012), Incentives for truthful reporting in crowdsourcing. In AAMAS '12 Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, Volume 3, Pages 1329-1330 Retrieved 12-05-2018 https://dl.acm.org/citation.cfm?id=2343988

Kuo, C., Romanosky, S., Cranor L. F., (2006), Human selection of mnemonic phrase-based passwords, Proceedings of the second symposium on Usable privacy and security, July 12-14 2006,Pittsburgh, Pennsylvania. Pages 67-78, doi:10.1145/1143120.1143129

Kävrestad, J., Eriksson, F., and Nohlberg, M., (2018) The development of a Password Classification Model. In Dhillon, G. & Samonas, S., eds. 17th Annual Security Conference, Securing the interconnected world, March, 26-28 2018 Las Vegas, NV. In Proceedings of the Annual Information Institute Conference.

Mason, W., Watts, J. D., (2009), Financial incentives and the “performance of crowds”, In ACM SIGKDD Explorations Newsletter, Volume 11, Pages 100-108, doi:10.1145/1809400.1809422

Paolacci G., Chandler J., 2014, Inside the Turk Understanding Mechanical Turk as a Participant Pool. Sage journals. doi:10.1177/0963721414531598

Peer. E., Vosgerau. J., Acquisti. A., (2013), Reputation as a sufficient condition for data quality on Amazon Mechanical Turk, Springer, doi:10.3758/s13428-013-0434-y

Riley, S., (2006), Password Security: What Users Know and What They Actually Do, In Usability News, Volume 8, Retrieved 14-05-2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.597.5846&rep=rep1&type=pdf

Robson, C., McCartan, K., (2016), Real World Research. A Resource for Users of Social Research Methods in Applied Settings. 4th Ed. John Wiley & Sons Ltd. ISBN: 9781118745236

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., Tomlinson, B., (2010) Who are the crowdworkers?: Shifting demographics in mechanical turk, In proceeding CHI EA '10 CHI '10 Extended Abstracts on Human Factors in Computing Systems, Pages 2863-2872, doi:10.1145/1753846.1753873

Snow, R., O’Connor, B., Jurafsky, D., Ng, A. Y., (2008) Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks, in EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language

Processing, 254-263, Retrieved 13-05-2018 from https://dl.acm.org/citation.cfm?id=1613751

What are masters, (n.d.), FAQs: Who are Amazon Mechanical Turk Masters?, mturk.com Retrieved 19-04-2018 from: https://requester.mturk.com/help/faq#what_are_masters

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M, C., Regnell, B., Wesslen, A., (2012) Experimentation in Software Engineering. Springer Heidelberg New York Dordrecht London, ISBN 978-3-642-29043-5, doi:10.1007/978-3-642-29044-2

Zhu, B. B., Yan, J., Bao, G., Yang, M., Xu, N., (2014), Captcha as Graphical Passwords - A New Security Primitive Based on Hard AI Problems. IEEE Transactions on Information Forensics and Security, Vol. 9, No. 6, pp.891-904, IEEE, doi:10.1109/TIFS.2014.2312547