EFFECTS OF COGNITIVE ABILITIES ON RELIABILITY OF CROWDSOURCED RELEVANCE JUDGMENTS IN INFORMATION RETRIEVAL EVALUATION

PARNIA SAMIMI

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2016

ABSTRACT

Test collection is extensively used to evaluate information retrieval systems in laboratory-based evaluation experimentation. In a classic setting of a test collection, human assessors involve relevance judgments which is costly and time-consuming task while scales poorly. Researchers are still being challenged in performing reliable and low-cost evaluation of information retrieval systems. Crowdsourcing as a novel method of data acquisition provides a cost effective and relatively quick solution for creating relevance judgments. Crowdsourcing by its nature has a high level of heterogeneity in potential workers to perform relevance judgments, which in turn cause heterogeneity in accuracy. Therefore, the main concern for using crowdsourcing as a replacement for human expert assessors is whether crowdsourcing is reliable in creating relevance judgments. It is an important concern, which needs to identify factors that affect the reliability of crowdsourced relevance judgments. The main goal of this study is to measure various cognitive characteristics of crowdsourced workers, and to explore the effect(s) that these characteristics have upon judgment reliability, as measured against a human assessment (as the gold standard). As such, the reliability of the workers is compared to that of an expert assessor, both directly as the overlap between relevance assessments, and indirectly by comparing the system effectiveness evaluation arrived at from expert and from worker assessors. In this study, we assess the effects of the three different cognitive abilities namely verbal comprehension skill, general reasoning skill and logical reasoning skill on reliability of relevance judgment in three experiments. Furthermore, workers provided some information about their knowledge about the topics, their confidence in performing given tasks, the perceived tasks' difficulty, as well as their demographics. This information is to investigate the effect of various factors on the reliability of relevance judgments. In this work, we hypothesized that workers with higher cognitive abilities can outperform the workers with lower level of cognitive abilities in providing reliable relevance judgments in crowdsourcing.

All of the three experiments show that individual differences in verbal comprehension skill, as well as general reasoning skill and logical reasoning skill are associated with reliability of relevance judgments, which leaded us to propose two approaches. These approaches are to improve the reliability of relevance judgments. Filtering approach suggests recruiting workers with certain level(s) of cognitive abilities for relevance judgment task. Judgment aggregation approach incorporates scores of cognitive abilities into aggregation process. These approaches improves the reliability of relevance judgments while have a small effect on system rankings. Self-reported difficulty of a judgment and the level of confidence in performing a given task have significant correlations with reliability of judgments. Unexpectedly though, self-reported knowledge about a given topic and demographics data have no correlation with the reliability of judgments. This study contributes to the information retrieval evaluation experimental methodology by addressing the issues faced by those researchers who use test collections for information retrieval system evaluation. This research emphasizes the importance of the cognitive characteristics of crowdsourcing workers as important factors in performing relevance judgment tasks.

ABSTRAK

Koleksi ujian digunakan dengan meluas untuk menilai sistem capaian maklumat berasaskan penilaian uji kaji makmal. Dalam tetapan yang klasik, penilai manusia melibatkan penghakiman kerelevanan di mana ianya mahal dan tugas yang memakan masa manakala keupayaan penskalaan kurang baik. Penyelidik masih dicabar dalam melaksanakan penilaian yang boleh dipercayai dan penilaian system capaian yang berkos rendah. *Crowdsourcing* sebagai kaedah novel pengambilalihan data menyediakan kos penyelesaian yang berkesan dan agak cepat untuk mewujudkan penghakiman kerelevanan. *Crowdsourcing* dengan sifatnya mempunyai tahap kepelbagaian yang tinggi dengan pekerja yang berpotensi untuk melaksanakan penghakiman kerelevanan, yang seterusnya menyebabkan kepelbagaian dalam ketepatan. Oleh yang demikian, masalah utama untuk menggunakan *crowdsourcing* sebagai pengganti penilai pakar manusia adalah, adakah *crowdsourcing* boleh dipercayai dalam mewujudkan penghakiman kerelevanan. Ia adalah satu masalah yang penting, perlu mengenal pasti faktor yang mempengaruhi kebolehpercayaan penghakiman kerelevanan yang dihasilkan melalui *crowdsourcing*. Matlamat utama kajian ini adalah untuk mengukur pelbagai ciri-ciri kognitif pekerja *crowdsource*, dan untuk meneroka kesan ciri-ciri ini terhadap kebolehpercayaan, penghakiman, berbanding dengan penilaian manusia (sebagai standard piawaian). Oleh itu, kebolehpercayaan pekerja dibanding dengan penilai pakar, bagi kedua-dua pertindihan langsung antara penilaian kerelevanan, dan secara tidak langsung dengan membandingkan keberkesanan system diperolehi daripada pakar dan juga daripada penilai pekerja. Dalam kajian ini, kami menilai kesan daripada tiga kebolehan kognitif yang berbeza iaitu kemahiran lisan, kemahiran penaakulan umum dan logik pemikiran kemahiran ke atas kebolehpercayaan penghakiman kerelevanan dalam tiga eksperimen berbeza. Tambahan pula, pekerja menyediakan beberapa maklumat mengenai pengetahuan mereka tentang topik, keyakinan mereka dalam tugas-tugas yang diberikan, kesukaran tugas, serta demografi mereka.

Maklumat ini adalah untuk mengkaji kesan pelbagai faktor ke atas kebolehpercayaan penghakiman yang kerelevanan. Dalam penyelidikan ini, hipotesisnya ialah pekerja dengan kebolehan kognitif yang lebih tinggi boleh mengatasi pekerja dengan tahap kognitif yang lebih rendah dalam menyediakan penghakiman kerelevanan yang boleh dipercayai melalui *crowdsourcing*. Ketiga-tiga eksperimen menunjukkan bahawa perbezaan individu dalam kemahiran lisan kefahaman, serta keseluruhan kemahiran penaakulan dan kemahiran pemikiran logik boleh dikaitkan dengan kebolehpercayaan penghakiman kerelevanan, yang membawa kita untuk mencadangkan dua pendekatan. Pendekatan ini adalah untuk meningkatkan kebolehpercayaan penghakiman kerelevanan. Pendekatan penapisan mencadangkan menggunakan pekerja dengan tahap kebolehan kognitif tertentu untuk tugas penghakiman relevan. Penghakiman pendekatan pengagregatan menggabungkan skor kebolehan kognitif dalam proses pengagregatan. Pendekatan in meningkatkan kebolehpercayaan penghakiman kerelevanan mempunyai kesan kecil ke ataskedudukan sistem. Kesukaran yang dilaporkan sendiri dan tahap keyakinan dalam melaksanakan tugas yang diberikan mempunyai korelasi yang signifikan ke atas kebolehpercayaan penghakiman. Kerelevanan didapati tahap, pengetahuan yang dilaporkan sendiri mengenai topik yang diberikan dan data demografi tidak mempunyai korelasi dengan kebolehpercayaan penghakiman. Kajian ini menyumbang kepada metodologi penilaian system capaian maklumat secara eksperimen dengan menangani isu-isu yang dihadapi oleh penyelidik yang menggunakan koleksi ujian untuk penilaian. Kajian ini menekankan pentingnya ciri-ciri kognitif pekerja crowdsourcing sebagai faktor-faktor penting dalam melaksanakan tugas-tugas penghakiman kerelevanan.

ACKNOWLEDGMENT

Many people have supported and encouraged me to undertake this experience at University of Malaya. First, I am offering my insubstantial gratitude to my supervisor Dr. Sri Devi Ravana for providing guidance and supporting me throughout these years, reading this thesis, and being patient during various stages of this study. I would also like to acknowledge the faculty of computer science and information technology.

I would also like to thank my parents and my sister, Parisa, for the extraordinary support they provided. Finally, a huge thank you goes out to my husband, Shahram, without whose reassurance, love and editorial assistance, I would not have finished this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronym	Definition
AMT	Amazon Mechanical Turk
ANOVA	Analysis of Variance for Repeated Measures
AP	Average Precision
AUS	Australia
BHS	Bahamas
CAN	Canada
CEM	Crowdsourcing Event Monitoring
CLEF	Cross Language Evaluation Forum
DCG	Discounted Cumulative Gain
EM	Expectation Maximization
FRCT	Factor Referenced Cognitive Test
GBR	Great Britain
HIT	Human Intelligence Task
ICT	Information and Communication Technologies
INEX	Initiative for The Evaluation of Xml Retrieval
IR	Information Retrieval
IRL	Ireland
MAP	Mean Average Precision
MV	Majority Voting
NB	Naive Bayes
NDCG	Normalized Discounted Cumulative Gain
NfC	Need for Cognition
NIST	National Institute of Standards and Technology
NZL	New Zealand
PMF	Probabilistic Matrix Factorization
TREC	Text Retrieval Conference
USA	United States

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Introduction

Since late 1990s, search engines have gradually become the most crucial systems for seeking information in the Web. Search engines provide a link between information inquired by a user and the outcome by matching information in a query format to documents, with computer system like Web or personal computer. In Information Retrieval (IR) process, a user submits a search query, and a search engine returns information relevant to the query, retrieving a set of relevant documents or Webpages. There are a large amount of available information on the Web therefore, it is essential to return the most relevant documents to the users' queries. As such, it is important to have a proper evaluation method to qualify retrieval of search engines, which provokes the growth and advancement of retrieval evaluation methods.

Retrieval evaluation approaches are either user-based or system-based methods. The user-based methods monitor user's behavior for searching information to find out whether a user is satisfied with the returned search results. The user-based methods are not reproducible because they are dependent on a bunch of users and each user has a different information need. Users are not involved in a retrieval evaluation through system-based methods. Instead, retrieval systems are evaluated according to document rankings. Test collection model such as Text Retrieval Conference (TREC)[1] is used by the system-based experiments to measure system performance, which is reproducible method with lower cost. Test collections consist of document corpora and search queries (topics) with respective relevance judgments.

[1] http://trec.nist.gov/

In traditional method, test collections are created under controlled conditions: expert searchers create topics and the documents retrieved by numerous IR systems are pooled to be assessed by trusted human assessors. The compiled set of documents, topics and relevance labels are then used to compute performance metrics across IR systems, e.g., *precision* and *recall*. Formerly, relevance judgments set has been created by hiring human experts who are trained to interpret topics precisely and judge their relevancy to documents. As the size and diversity of test collections have massively increased, hiring expert assessors appeared expensive and burdensome for performing judgments. Indeed, major challenges of TREC-like test collection approach are time and cost for relevance judgments which makes that an unsuitable approach for scaling up (Alonso & Mizzaro, 2012).

The recent growth of the test collections has led to adapt crowdsourcing method for creating relevance judgments. The term crowdsourcing was coined by Howe based on Web 2.0 technology in a Wired Magazine article (Howe, 2006). Crowdsourcing is defined as outsourcing tasks, which were formerly accomplished inside a company or institution by employees, to a huge, heterogeneous mass of potential workers in the form of an open call through Internet. Crowdsourced workers (henceforth "workers") are hired through online web services such as Crowdflower, and work online to perform repetitive cognitive piece-work (known as HITs) at low cost, with many workers potentially working in parallel to quickly complete a task. Crowdsourcing is an efficient method particularly for tasks in which human participations are necessary, such as creating relevance judgments in IR evaluation (Alonso, Rose, & Stewart, 2008; Grady & Lease, 2010; Kazai, Kamps, Koolen, & Milic-Frayling, 2011). The main feature that makes this approach attractive is its flexibility, low cost and fast outcome (Alonso & Mizzaro, 2012).

Despite the popularity of crowdsourcing in creating relevance judgments, its reliability has been questioned for various reason. For instance, do the workers have adequate expertise for a given task (Quinn & Bederson, 2011)? Are demographics and personality traits of workers affect the quality of crowdsourced relevance judgments (Kazai, Kamps, & Milic-Frayling, 2012)? Moreover, the quality of the final relevance judgments is highly subjective to how a worker is interested and incentive in performing a given task (Kazai, Kamps, & Milic-Frayling, 2011).

A range of quality assurance and control techniques are developed to reduce noise that produced during or after completion of a given task. However, little is known about the workers themselves and the role of individual differences in reliability of crowdsourced relevance judgment. Cognitive performance is of individual differences, which is also referred to "cognitive abilities". Cognitive abilities are mainly brain-based skills, concerning learning, remembering, problem-solving, and attention and mindfulness (Ekstrom, French, Harman, & Dermen, 1976). This study focuses on three specific aspects of cognitive abilities, including (i) verbal comprehension skill, (ii) general reasoning skill, and (iii) logical reasoning skill, to determine their relationship with the reliability of crowdsourced relevance judgment.

1.2 Motivation

Crowdsourcing becomes popular in recent years because of the ubiquity of the Internet. Despite the popularity of crowdsourcing, it comes with the risk of a heterogeneous mass of potential workers who create the relevance judgments with varied levels of accuracy. This heterogeneity prevents enterprises from participating in such an open and cost effective innovation setting and using digital workers in replacement of the usual employees. Some workers are less reliable and less accurate that may mess up with those workers who are efficient and accurate in performing tasks in crowdsourcing.

Such variations among workers do not promote both the optimization of performances and the accuracy of the crowdsourced results. According to Li, Zhao, and Fuxman (2014), reliability of workers is a long-lasting issue in crowdsourcing and therefore it is important to find a way to screen workers based on their levels of quality. Therefore, it is important to understand how worker's cognitive abilities affect the reliability of crowdsourcing results to identify proper workers for performing tasks. Our study investigates the association between individual difference in cognitive abilities of crowdsourced workers and level of reliability of their relevance judgments. If some features and characteristics of workers are associated with their quality, these characteristics should be considered in estimation of worker quality (Li et al., 2014).

The association between cognitive abilities and reliability of relevance judgments performed by crowdsourcing will provide an important insight for IR practitioner to consider cognitive characteristics of assessors in recruiting them. Outcomes of this research may help to assign proper workers with certain qualities to perform the tasks optimally and produce accurate results. Besides, the outcome of this study would benefit businesses or individuals to be able to select the right crowd and achieve reliable business results.

1.3 Statement of the Problem

One of the main concerns about crowdsourcing is its low quality output due to heterogeneous workers including various behavior, characteristics, skills, levels of attention and accuracy (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010; Zhu & Carterette, 2010; Kazai, Kamps, Koolen, et al., 2011). Several studies investigated the effect of different factors on reliability of crowdsourced judgments and assessed a range of quality assurance and control techniques to reduce noises in crowdsourcing. However, little is known about the workers themselves and the effects of cognitive abilities on the

reliability of relevance judgment produced by crowdsourcing. A bunch of previously published works in information science area evaluated how cognitive differences might influence IR process (Allen, 1992; Allen & Allen, 1993; Ford, Wilson, Ellis, Foster, & Spink, 2000; K. S. Kim & Allen, 2002; Brennan, Kelly, & Arguello, 2014).

According to our literature investigation by the time of performing our experiments and writing the thesis, no specific research work was published to report the association between cognitive abilities and reliability of relevance judgments in crowdsourcing. It seems essentially important to understand whether human factors significantly influence the reliability of relevance judgment performed by crowdsourced in order to improve the quality of crowdsourcing outputs, for instance by choosing a right group of workers for creating relevance judgment. Understanding the relationship between workers' cognitive abilities and reliability of relevance judgments may convey some new ideas to propose new approaches in crowdsourcing to enhance the reliability of relevance judgments.

1.4 Objectives of this Study

The main objective of this study is to investigate the effect of cognitive abilities of crowdsourced workers on the reliability of relevance judgments performed through crowdsourcing. The objectives of this study are as followed:

i. To investigate the effects of verbal comprehension skill, general reasoning skill and logical reasoning skill on reliability of crowdsourced relevance judgments

ii. To investigate if verbal comprehension skill, general reasoning skill and logical reasoning skill affect IR systems performance rankings in IR evaluation experimentation.

iii. To enhance the reliability of relevance judgments performed by crowdsourced workers through the two proposed approaches: a filtering approach and a judgment aggregation approach.

1.5 Contributions

System-based IR evaluation is a main method for the assessing and comparing IR systems. This study presented throughout this thesis makes contribution to the experimental methodology addressing some issues of using test collections for IR system evaluation:

- ***Factors affecting the reliability of crowdsourced relevance judgments:*** One of the contributions of this work is to provide a comprehensive survey of various factors affecting the quality and reliability of crowdsourcing outcome as well as crowdsourced relevance judgments. This survey highlights missing factors which have an effect on reliability of relevance judgments.

- ***Association between cognitive abilities and reliability of relevance judgments***: The verbal comprehension, general reasoning and logical reasoning experiments conducted to address the first and the second objectives is to investigate the association between cognitive abilities and reliability of relevance judgments performed by crowdsourcing. The findings of these experiments provides a crucial insight for IR practitioner in predicting workers' accuracy based on their cognitive characteristics. In fact, these findings can be beneficial in determining high quality workers for performing relevance judgments.

- ***Workers filtering approach based on level of cognitive abilities***: As cognitive abilities of workers are associated with their reliability of relevance judgments, these characteristics can be considered to estimate the quality of their outcomes. In this

study, a filtering approach is proposed to select a certain group of workers according to their level of cognitive abilities for creating relevance judgments. This approach provides an insight over the crowdsourcing experiments to effectively achieve reliable relevance judgments and rank system performance in IR system evaluation by choosing certain group(s) of workers according to their cognitive abilities.

- ***Judgment aggregation approach***: The interesting results of relationship between workers' cognitive abilities and reliability of relevance judgments motivate to utilize this competence in judgment aggregation approach. A judgment aggregation approach introduced in this work is to aggregate the relevance judgments based on the workers' cognitive ability scores. This approach is to derive a reliable relevance judgment from multiple judgments.

1.6 Thesis Structure

This chapter provided an introduction about IR evaluation approaches, test collections and crowdsourcing method. Motivation, problem statement and objectives of this study explained in detail providing contributions of this study. In the next chapter (Chapter 2), IR evaluation, crowdsourcing and cognitive abilities are discussed in detail giving some explanations about IR evaluation methods, user-based and system-based methods. History of TREC and test collections, evaluation metrics of this study as well as details about relevance evaluation and its challenges are presented in Chapter 2. Crowdsourcing and its application in different areas as well as influential factors on reliability of crowdsourcing outputs and quality control methods are elaborated in this chapter. Different studies are reviewed to evaluate the crowdsourcing in relevance judgments in IR evaluation. Experimental methodologies used throughout this study are presented in Chapter 3. Firstly, the experimental design, experimental data and task design are explained in detail. Subsequently, different metrics used to compute reliability

of relevance judgments are explained. Finally, analysis methods and the pilot study are explained and discussed in this chapter.

Chapters 4 to 6 are presenting and discussing results for the three experiments in this study. In Chapter 4 results for verbal comprehension experiment are presented and discussed to determine the effect of verbal comprehension skill on reliability of relevance judgments. Chapter 5 provides results and discussion for the general reasoning experiment to find out the effect of general reasoning skill on reliability of relevance judgments. Results of the effect of logical reasoning skill on reliability of relevance judgments is discussed in Chapter 6. For Chapter 4 to 6, the filtering spam method is explained separately followed by some descriptive statistics of data. The effect of certain cognitive abilities on the reliability of relevance judgment and on system rankings are also investigated separately. Various self-reported competences including difficulty of task, confidence in judgment and knowledge on the given topic for each worker are assessed to find their associations with the level of accuracy attributing their relevance judgments. Furthermore, relationship between demographic data and reliability of relevance judgments performed by crowdsourcing are assessed separately for each experiment.

According to the findings of the three experiments, two proposed approaches are discussed in Chapter 7 for improving reliability of relevance judgments. A filtering approach for choosing workers with higher level of cognitive abilities is explained in this Chapter to enhance the reliability of relevance judgments. Subsequently, a judgment aggregation approach is introduced and compared with a commonly used method for aggregation. The two proposed approaches are tested for each of the three experiments (verbal comprehension, general reasoning and logical reasoning experiment) as well.

Finally, Chapter 8 summarizes the results and concludes achievements of this research on the basis of the objectives of this study, providing some suggestion for future studies.

CHAPTER 2: LITERATURE REVIEW

This chapter addresses three main areas of the scopes of the current research study named "Information Retrieval Evaluation", "Crowdsourcing" and "Cognitive Abilities". Information Retrieval Evaluation section reviews different methods of IR evaluation, the formation of TREC, with some explanation about test collections and their components. Evaluation measures used in this study and relevance evaluation are also reviewed. During the course of the second chapter, Crowdsourcing section provides some basic definitions, and describes different factors influence the reliability of crowdsourcing. Quality control methods, which are applicable in crowdsourcing, are reviewed provided by a review on some magnificent studies in the area of crowdsourcing in IR evaluation. The third section, which is Cognitive Ability section, describes cognitive abilities, and discusses cognitive abilities in IR process.

2.1 Information Retrieval Evaluation

2.1.1 Background

Information retrieval is associated with representation, storage, organization of, and the access to information items. User interprets his information need to a query, which consists of a set of keywords that can be processed by IR system. The main goal of IR system is to return information relevant to a user's query. In other words, the returned items are supposed to be related to the user query and provide meaningful outcomes (Baeza-Yates & Ribeiro-Neto, 2011). Figure 2.1 presents a schematic view of a typical IR process. A user sends information needs (in a query format) to an IR system. Then the IR system returns some information relevant to the user need. Effective IR system is supposed to return highly relevant information according to a user's query (user satisfaction). In this regard, and to assure about the level of effectiveness of an IR system,

evaluation of the performance of this system is critically important. IR evaluation is to assess how effectively IR system addresses the information needs of the users.

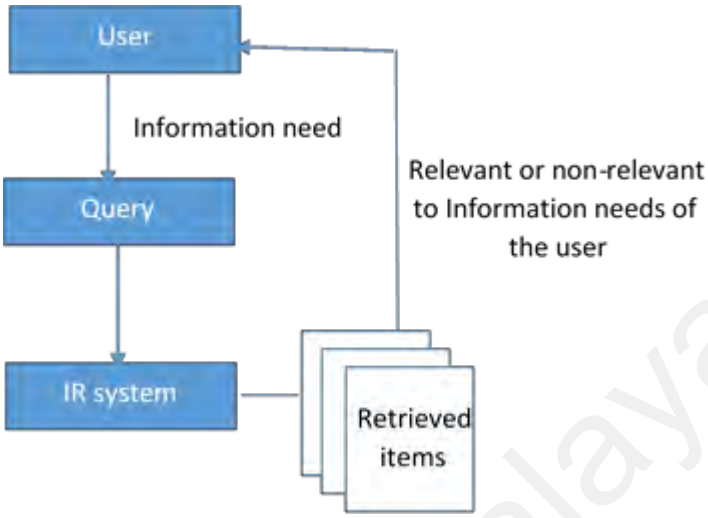


Figure 2.1: Schematic view of a typical IR process

With an appropriate evaluation technique, it can be defined how well an IR system is acting. Furthermore, IR evaluation makes it possible to compare retrieval quality of different IR systems. In other words, systematically linking a quantitative metric to the results, which returned by an IR system for a set of queries is IR evaluation. The quantitative metrics should represent how well the results are relevant to a user query and it is commonly calculated by comparing results returned by an IR system with that of suggested by human judges for a particular set of queries (Baeza-Yates & Ribeiro-Neto, 2011). There are basically two types of methods for evaluating the effectiveness of IR systems: (i) system-based evaluation, and (ii) user-based evaluation (Voorhees, 2002).

i. System-based evaluation

The Cranfield tests, which was performed in the 1950s and 1960s by Cyril Cleverdon, were the basis of system-based evaluations (Cleverdon, 1967). Cleverdon recognized two types of devices that have effect on effectiveness: (i) precision devices, which increased the proportion of relevant documents among those retrieved and (ii)

recall devices, which increased the proportion of all relevant documents found. Cleverdon asked a group of users (authors of research papers in aeronautics engineering) to define the research question that inspired the work. Then, he requested them to identify the rate of their cited references on a scale of 1 to 5 for relevance to the research question. Finally, he could simulate a user study by having both research questions and ratings relevancy. This methodology is called "Cranfield paradigm" or system-based evaluations. In system-based evaluations or batch evaluations, a batch of queries derived from a set of information needs is submitted to a system and then the relevance of the ranked documents are measured without human interference (Carterette & Voorhees, 2011).

The batch evaluation can also be done speedily. In this approach, the differences in effectiveness would be noticeable for developers with a careful sufficient measurement process while it may not be noticeable to individual users. System effectiveness is calculated based on a selected evaluation metric. System-based experiments has been started since the Cranfield paradigm and has continued through TREC (Voorhees & Harman, 2005). One of the TREC goal in the past few years has been to discover and evaluate inventive retrieval approaches over large-scale subsets of the Web (Collins-Thompson, Bennett, Diaz, Clarke, & Voorhees, 2014).

ii. User-based evaluation

The user-based evaluation method quantifies the satisfaction of users by monitoring the user's interactions with the system. This approach for evaluation is believed to capture users' real behavior and the actual system performance (Goker & Davies, 2009). Other aspects of user satisfaction can be monitored and measured including how much the user willing to work with the system; how the relevancy of information needs of users and the retrieved documents is; how is the speed of information seeking; in operational setting. Al-Maskari and Sanderson (2010) introduced

four factors, which affect user satisfaction: user effectiveness, system effectiveness, user effort and characteristics. User effectiveness refers to how well the user performs the task such as number of retrieved documents, which are relevant, and the completion time. System effectiveness is calculated by observing how well a system accomplished in retrieving relevant documents. Another factor is user effort measured by computing the time and energy applied by a user to accomplish a task such as number of clicks. The last factor is user characteristics such as users' information searching experience and skill. However, user satisfaction is difficult to measure and needs careful observation and control of many variables while designing an experiment.

Table 2.1 illustrates different user-based evaluation methods (Baeza-Yates & Ribeiro-Neto, 2011). One drawback of "human in the lab" method is the limitation of a small number of humans and a small set of information. Moreover, the cost of a setup and repeating an experiment is high. "Side-by-side panels" allow a comparison of two systems but it is not applicable to multiple systems. "A/B testing" is mainly important with those sites with many users since a poor alteration that is launched may lead to irritation to millions of users. Due to the low cost of collecting data, "using click-through data" seems attractive, but it needs a precise setup to avoid noise. User-based methods deal with obtaining and analyzing users' feedback on retrieval performance, therefore these methods require human participation, which makes this method costly, and time consuming. The concern is that whether using the user-based experiments is the best way to compare and measure the systems effectiveness. According to Al-Maskari and Sanderson (2010), user-based experiments are able to differentiate system effectiveness (measuring the user satisfaction factors), however, these kind of experiments are not repeatable due to the needs of different resources, which are costly (Alonso & Mizzaro, 2012).

Table 2.1: User-based evaluation methods

User-based methods	Description
Human in the lab	This method involves human experimentation in the lab to evaluate the user-system interaction.
Side-by-side panels	This method is defined as collecting the top ranked answers generated by two IR systems for the same search query and representing them side by side to the users. To evaluate this method, in the eyes of human assessor, a simple judgment is needed to see which side retrieve better results.
A/ B testing	A/ B testing involves numbers of preselected users of a Web site to analyse their reactions to the specific modification to see whether the change is positive or negative.
Using click-through data	Using click-through data is to observe how frequently users click on retrieved documents for a given query.

2.1.2 Text Retrieval Conference

Before 1990, different research groups were interested in evaluating the retrieval systems' performance independently but they could not compare their results together. Jones (1981b) found this lack of coordination. The main pitfalls of the experiments include (i) lack of a framework for system evaluation, (ii) huge cost for large retrieval tests, (iii) inaccessibility of data, and (iv) difficulty in comparing results across different projects because of the inconsistency in methodology used. Voorhees and Harman (2005) described Sparck Jones' idea, the unavailability of a platform for researchers to use the same data and measures and later being able to compare results; and secondly, the small size of the test collection not simulating the real world data led to the TREC project initiation.

In 1990, the National Institute of Standards and Technology (NIST) built a large collection of documents for the TIPSTER project (Voorhees & Harman, 2005). By the formation of TREC, this collection was later become openly accessible for the researchers. The TREC established in 1992 to support IR researches providing an infrastructure for large-scale evaluation of retrieval methodologies. TREC-1 was the first TREC conference held in 1992 and the positive findings supported the significance of

large test collections. Twenty-five participating systems submitted runs (*i.e.* a set of documents retrieved by a system for a set of topics) for evaluation in TREC-1. In 1993, TREC-2 for retrieval algorithms developed using large collections were adapted by the commercial world. A TREC workshop includes a collection of *tracks* with different aims. Each *track* focuses on a specific retrieval area and issue such as question answering. The concept of *tracks* at TREC-4 was coined in 1995 and continued with various track topics.

Generally, the TREC aims to improve in terms of the components of test collections including data collection and topics to enhance better simulate the Web for reliable result. Many efforts enhanced the validity of relevance judgments, built from a huge collection and retrieved tasks which were designed to emphasize on effectiveness of retrieval methods. TREC was very successful in attracting commercial researchers and test collections were used by some commercial systems to evaluate their systems. The usage of test collections for evaluations of the Web search is not comparable with user-based experiments, however, it provides some facilities which user-based approaches are not able to provide such as reusability and repeatability (Voorhees & Harman, 2005).

2.1.3 Test Collections

Cranfield experiments were popular before 1990s (Cleverdon, 1967). Then, some other large test collections were established. The TREC and Cross Language Evaluation Forum (CLEF)[2] are two common test collections in 1990s and later. In fact, Cranfield experiments were the beginning of today's laboratory retrieval evaluation experiments.

A goal of a test collection is to model users with information needs which are examples of the task (Tague-Sutcliffe, 1996). These information needs are the

[2] http://www.clef-campaign.org/

representative of the users' needs from a system in general. If a system can perform well on a test collection, the system will then be supposed to perform thriving in general. Basically, a test collection consists of three components. (i) document corpus, which is a set of large size documents (for instance one billion web pages crawled from the general web (ClueWeb09)), (ii) topics that are a collection of search queries or information needs of users, and (iii) relevance judgments, which shows which documents relevant to which topics and involves human expert assessors. The assessors are retired information specialists paid to carry out the relevance judgment task. In another word, an expert assessor have to decide which documents are relevant to a given topic. Indeed, relevance judgment means judging every single document in the document corpus to every single search query, which is the only way to guarantee that all relevant documents are identified. However, this is impossible due to the limitation of time and budget. For instance, if an assessor judge 10 documents per minute, judging a million documents would take about ten months of 40 hour/week to judge one topic. A complete judgement collection for TREC-2010 Web needs expert assessors to assess 1 billion documents. Assuming that an expert assessor can assess two documents per minute, judging one billion documents need about 347,000 days. Therefore, a large number of human experts should be appointed, and of course be paid (Moghadasi, Ravana, & Raman, 2013).

Judging a small portion of document corpus may provide enough relevant documents for the purposes of evaluation and experimentation. In practice, pooling method is a handy approach to recognize a subset of documents for judging. However, creating relevance judgments is time-consuming and even for pooling method. [for instance, it took approximately 7 hours per topic to be accessed for INEX 2006 (Trotman & Jenkinson, 2007)]. The process of a typical IR evaluation through a test collection is shown in Figure 2.2. Participating systems run their retrieval algorithms against the

document corpus and topics in the test collection. The retrieval algorithms generate a set of documents called runs. The systems 1, 2, …, *m* are the contributing systems for the pool creation. A collection of top ranked documents for each topic (retrieved by contributing systems) is then selected for judgment.

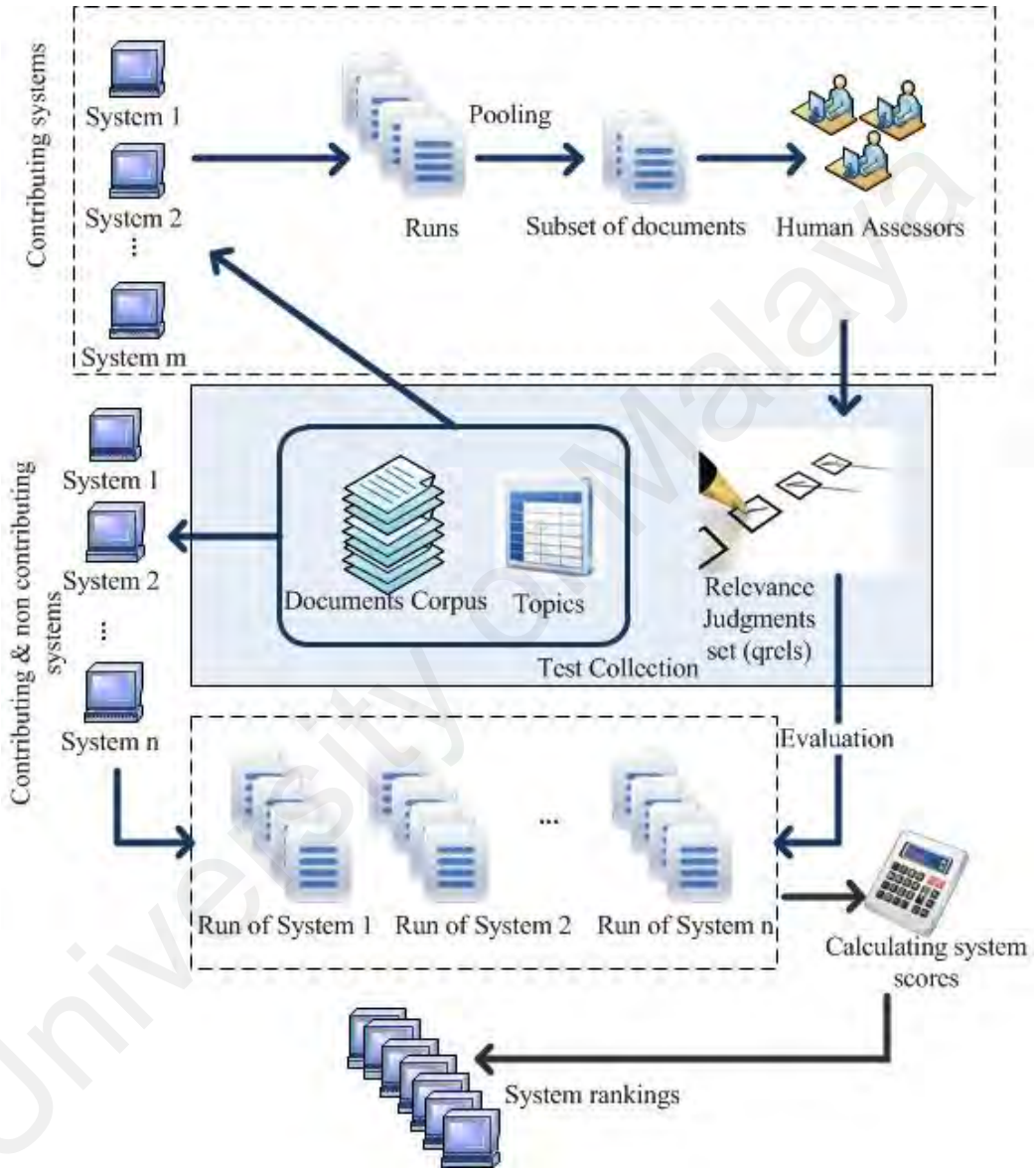


Figure 2.2: Schematic view of information retrieval evaluation process

Documents in the pool are judged by human assessors (to create relevance judgment set), and all of the other documents outside the pool are considered non-relevant documents. Once the relevance judgments are ready, the whole set of runs retrieved by both contributing and non-contributing systems (1, 2, …, *n*) is evaluated against relevance judgments to measure the accuracy and effectiveness of the retrieval systems through evaluation metrics. Each system receives a score for each topic and that is to be aggregated in order to achieve an overall performance score for the system. In each IR experiment, the system ranking is generated for all of the systems. However, the major drawback of test collections is the huge cost of creating relevance assessment (conducted by human expert assessors). Overall, this method needs additional resources in terms of time, infrastructure and budget whilst it does not scale up simply. Section 2.1.5 provides some detailed explanation about relevance assessment.

2.1.4 Evaluation Measures

System effectiveness can be measured by a comparison with an ideal answer set (Tague-Sutcliffe, 1996). In the TREC environment, an ideal answer set is a set of relevance judgments or *qrels* which created by human experts. An IR system returns a list of ranked documents for each topic. The ranked list is then compared to *qrels* to produce a numerical value called evaluation metric. An example of *qrels* is provided in Table 2.2. There are four columns in a *qrels* file named Topic, Iteration, Document ID and Relevance. Each document in the *qrels* with a value of 0 (Relevance) indicates that a given topic and a correspondence document are not relevant to each other. If they are relevant, the *qrels* is asserted 1 (Relevance). The second column (Iteration) is not usable in the evaluation. Table 2.3 presents a list of ranked documents returned by system1 according to similarity scores provided by the evaluation process. Consequently, this list is compared with the *qrels* to produce a numerical number using an evaluation measure.

Table 2.2: Example of *qrels* file

Topic	Iteration (unused)	Document ID	Relevance
451	0	WTX002-B01-101	0
451	0	WTX002-B30-306	0
451	0	WTX003-B26-249	1

Table 2.3: Document ranking

Topic	Unused	Document ID	Rank	Similarity score	System tag
451	Q0	WTX002-B30-306	1	0.547434	System1
451	Q0	WTX003-B26-249	2	0.543610	System1
451	Q0	WTX002-B01-101	3	0.464663	System1

Document rankings is based on similarity score created by a system

Evaluation of information retrieval systems is mostly done through calculation of *precision* and *recall*. *Precision* is the performance measurement of an information retrieval system that quantifies the ratio of the retrieved documents, which are highly or marginally relevant (see Equation (2.1)).

$$Precision = \frac{\#(relevant\ documents\ \ retrieved)}{\#(retrieved\ documents)} \tag{2.1}$$

Recall is the fraction of the documents relevant to the query that are successfully retrieved:

$$Recall = \frac{\#(relevant\ documents\ \ retrieved)}{\#(relevant\ documents)} \tag{2.2}$$

Recall measures the completeness of the results while *precision* measures the accuracy. Therefore, retrieving more items enhances *recall* but suppresses *precision* (and vice versa), but enhancement of retrieval method by itself, can improve both of the metrics (Jones, 1981a). A number of evaluation measures use a combination of *recall* and *precision* and convert them in a single metric, such as *average precision* (AP) which is

the average of precisions observed at every rank at which a relevant document *d* appears, and the *R* is the number of relevant documents in the set.

$$AP = \frac{1}{R} \sum_{d\ s.t.d\ relevant} precision@rank(d) \quad (2.3)$$

The performance of a retrieval system for an individual topic does not reflect the overall performance of the retrieval system, and that is why the above metrics can be measured over a set of topics and then average them to have a single measure of effectiveness such as *"mean average precision"* or *MAP*. The relative performance of IR systems can be compared with their ability to determine relevant documents over a set of topics using measures such as *MAP* (Järvelin & Kekäläinen, 2000). The contingency table (Table 2.4) defines some of evaluation measures concepts.

Table 2.4: Contingency

	Relevant	Non-relevant
Retrieved	True positives (tp)	False positives (fp)
Not-retrieved	False negatives (fn)	True negatives (tn)

Then, *precision* and *recall* can be shown as:

$$Precision=tp/(tp+fp) \quad (2.4)$$

$$Recall=tp/(tp+fn) \quad (2.5)$$

Another alternative to judge information retrieval system is *accuracy* (Manning, Raghavan, & Schütze, 2008). *Accuracy* is the ratio of the true results over all of the results:

$$Accuracy=(tp+tn)/(tp+fp+fn+tn) \quad (2.6)$$

Barhydt introduced two measures to quantify the similarity between two relevance judgments (Barhydt, 1964); one is *sensitivity* [which is another terms for *recall*] and

another one is *specificity*. *Specificity* shows the ability of a retrieval method to determine negative results (Jung & Lease, 2012).

$$Specificity = tn/(fp+tn) \tag{2.7}$$

Effectiveness is a combination of *sensitivity* and *specificity* (Barhydt, 1964):

$$Effectiveness = Sensitivity + Specificity - 1 \tag{2.8}$$

Another evaluation metric that is used in this study is Discounted Cumulative Gain (DCG). This metric considers ranks of documents. Järvelin and Kekäläinen (2002) define DCG as:

$$DCG@k = \sum_{i=1}^{k} \frac{r_i}{\log(i+1)} \tag{2.9}$$

where, k is evaluation depth, and r_i is relevance of the document at rank i. A *recall* adjusted normalized version of DCG was also suggested by (Järvelin & Kekäläinen, 2002), and the new value is within the range of 0 and 1. This value was achieved by normalizing DCG against an ideal ordering of the relevant documents, where R is the number of relevant documents for a query, then NDCG, normalized discounted cumulative gain is calculated as:

$$NDCG@k = \frac{\sum_{i=1}^{k} r_i . w_i}{\sum_{i=1}^{\min(k,R)} w_i} \tag{2.10}$$

Where

$$w_i = \frac{1}{\log(i+1)} \tag{2.11}$$

2.1.5 Relevance Evaluation

Previous studies in different theoretical contributions addressed relevance as subjective, situational and psychological issue (Wilson, 1973; Swanson, 1977, 1986; Schamber, Eisenberg, & Nilan, 1990; Harter, 1992). Therefore, the relevance evaluation depends on different factors such as topic, document's characteristics, and users' actual cognitive state. Situational and psychological theories indicate that the relevance relationship between a user and document can be fluctuating according to the actual of situational and psychological states. Many personal factors such as experience, knowledge, education, and training influence the relevance judgment of users (Schamber, 1994). In fact, relevance judgments is subjected to users' characteristics (Harter, 1996). Consistently, previous studies implied that relevance judgments are affected by a varied psychological and situational conditions and factors. In addition, differences in relevance judgments can be due to individual differences in information retrieval (Harter, 1996). In a study of factors affecting relevance judgments, individual differences were considered as the most general feature of the data (Rees & Schultz, 1967). There are a number of studies focused on individual differences in different IR subsystems, such as search term productivity (Harter, 1990), human-computer interaction (Borgman, 1989), search term selection (Saracevic & Kantor, 1988), and problems, information needs, and changes in relevance judgments over time (Smithson, 1994). Indeed, individual differences have a large impact on human decision- making (Saracevic, 1991) and the evidences from previous studies imply that human involved in IR process and the relevance judgments that they make vary from one another (Harter, 1996). There are around eighty factors that may influence relevance judgment (Schamber, 1994). Among those, users' characteristics such as cognitive style, education, intelligence, and knowledge/experience are of most impelling factors for relevance judgments.

Creating relevance judgment for IR is an expensive and tough task (Alonso et al., 2008). At the beginning years of the field, a number of graduate students thoroughly judged the relevancy of every document in a corpus to a collection of queries voluntarily, however, only a few set of small test collections were eventually built (such as Cranfield). In 1992, researchers had access to millions of full-text documents through TREC. However, the idea of TREC was only probable by rejecting the idea that every document in corpus would be judged. Instead, only the top ranked documents retrieved by participating systems should be judged (pooling approach). In TREC, a large number of expert assessors, who are retired intelligence analysts and were paid for their work, were responsible for relevance judgments (Voorhees & Harman, 2005). TREC collection (especially the relevance judgment and query set) has been invaluable during the time for IR researches as they are limited to tasks that TREC suggests. In addition, relevance judgments is subjective and can be varied among assessors (Kazai, Kamps, & Milic-Frayling, 2013). For instance, an agreement between two TREC assessors was reported 70 to 80 % in average (Voorhees & Harman, 2005). However, system rankings in some degrees are robust with this inconsistency. Despite high level of disagreement on relevance judgments, a high level of agreement on system rankings appeared between TREC assessors and non-TREC assessors (Carterette & Soboroff, 2010).

Creating a relevance judgment is a challenging issue and many researchers have tried to overcome this issue (Trotman & Jenkinson, 2007). Many researchers deal with creating their own relevance judgment set by using editorial resources to match with their needs in both academia and industry. Most of the web search engines use their editorial staffs to evaluate the relevancy of web pages and queries, for instance. Academic researchers mostly rely on students for doing relevance judgments as they usually do not expect to be paid (Saracevic, 2007), however, this approach is not that much applicable

due to the availability of students. Therefore, the test sets performed by students are not large enough to determine statistical differences in performance of systems. Although this approach provides a good understanding of student relevance behavior, it does not reveal a proper understanding of actual users in a real situation. Having a good understanding of relevance behavior of actual users need a diverse population (instead of students).

Utilizing user's behavior as an evaluation indicator is another approach to obtain relevance judgments (Joachims & Radlinski, 2007). In compared with editorial method, this approach is applicable for larger scale with a low cost for relevance evaluation. Behavioral approach seems advantageous but it requires a huge stream of real behavioral data, which is not always accessible for researchers evaluating an experimental system. Therefore, another approach is required to compensate editorial approach on a large scale. Crowdsourcing is a suggested method to create relevance judgments while it can scale up both the number of judgments and topics (Alonso et al., 2008).

2.2 Crowdsourcing

Crowdsourcing platforms enable the requesters to have a fast access to an on-demand, global, scalable workforce and the workers free to choose as many tasks as they want to accomplish. The use of crowdsourcing in information system is relatively new, and is widely applied in a various fields of computer science (Zhao & Zhu, 2012). Different applications of crowdsourcing is presented in Table 2.5.

Crowdsourcing is provided through various platforms such as Amazon Mechanical Turk (AMT)[3] and Crowdflower[4]. These platforms allow requesters to submit tasks and the workers to perform the tasks. Human Intelligence Task (HIT) or microtask is a unit of accomplished work. Crowdsourcing scheme is presented in Figure 2.3. It

[3] https://www.mturk.com/mturk/welcome
[4] http://www.crowdflower.com/

includes multiple requesters for publishing tasks and workers to accomplish tasks. Crowdsourcing process starts with publishing tasks by requesters to the crowdsourcing platform. Workers select their tasks and complete them. The requesters then assess the results of the performed tasks. If the results were acceptable to the requesters, they would proceed the payment to the workers. Otherwise, the workers are rejected because of performing the task carelessly (Figure 2.4).

Table 2.5: Different applications of crowdsourcing

Domain	Description
Natural language processing	Crowdsourcing technology was used to investigate linguistic theory and language processing (Munro *et al.*, 2010).
Machine learning	Automatic translation by using active learning and crowdsourcing was suggested to reduce the cost of language experts (Callison-Burch, 2009; Ambati, Vogel, & Carbonell, 2010).
Software engineering	The use of crowdsourcing was investigated to solve the problem of recruiting the right type and number of subjects to evaluate a software engineering technique (Stolee & Elbaum, 2010).
Network event monitoring	Using crowdsourcing to detect, isolate and report service-level network events was explored which called Crowdsourcing Event Monitoring (CEM) (Choffnes, Bustamante, & Ge, 2010).
Sentiment classification	The issues in training a sentiment analysis system using data collected through crowdsourcing was analysed (Brew, Greene, & Cunningham, 2010).
Cataloguing	The application of crowdsourcing for libraries and archives was assessed (Holley, 2009).
Transportation plan	Using crowdsourcing was argued to enable the citizen participation process in public planning projects (Brabham, 2009).
Information retrieval	To create relevance judgements, crowdsourcing was suggested as a feasible alternative (Alonso et al., 2008).

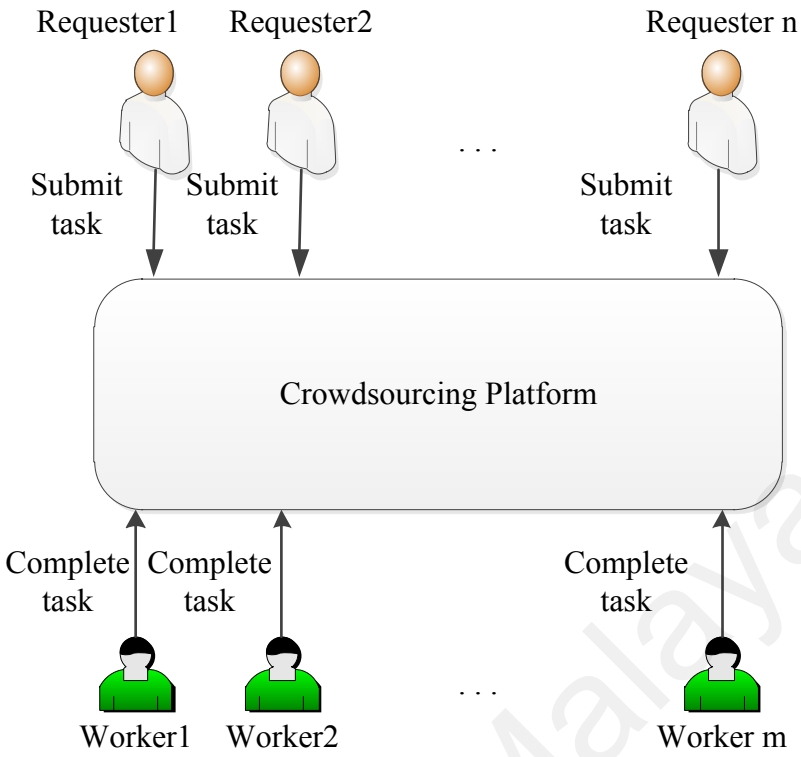


Figure 2.3: Crowdsourcing scheme

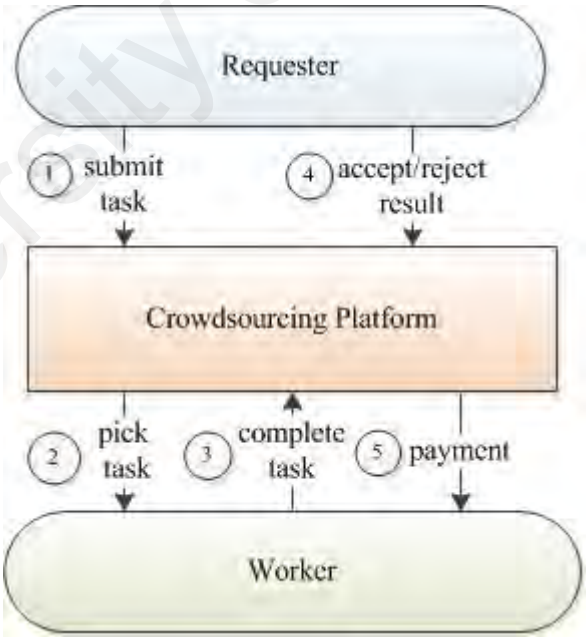


Figure 2.4: Procedure of crowdsourcing

The main feature of crowdsourcing is its simplicity. Crowdsourcing platforms was also suggested for data collection as a viable choice (Paolacci, Chandler, & Ipeirotis, 2010). Three advantages of crowdsourcing platforms are (i) allowing large number of workers to take part in experiments with low payment, (ii) workers are from diverse language, culture, background, age and country and (iii) low cost at which the researches can be carried on (Mason & Suri, 2012). But, crowdsourcing is engaged with low quality outputs due to variation in the workers' behaviors (Zhu & Carterette, 2010). Different researchers invetsigated different factors influencing reliability of crowdsourcing output which is discussed in the subsequent section (Section 2.2.1).

2.2.1 Factors that Affect the Reliability of Crowdsourcing Output

There are several studies investigating the effect of different factors on the reliability of crowdsourcing experiment comprises experimental design, human features and monetary factors.

i. Experimental design in crowdsourcing

Experimental design is the most critical part of the crowdsourcing (Alonso, 2012). Beyond the workers' levels of attention, diversity of cultures, and variations in preferences and skills, the presentation and properties of HITs is the key factor for the quality of crowdsourcing. The quality of the user interface, instructions and the design of crowdsourcing has a direct relationship with the quality of task performed by a worker. In the experimental design, the first information required to be presented to the workers is the definition of the given task. Task description, a part of task preparation, is an important topic in implementing a crowdsourcing experiment. A clear instruction is a part of task description, is crucial to have a quick result. Ideally, all of the workers should have a common understanding about a chosen task, and the task must be understandable in terms of language by different workers (Alonso, 2012). Task description should be

prepared according to the variation of general characteristics of workers such as their languages and/or the level of their expertise in the field (Allahbakhsh *et al.*, 2013).

Creating relevance judgment requires reading text. Plain English is the choice of interest for a diverse population, to avoid jargon (Alonso & Baeza-Yates, 2011). Therefore, using plain English words impacts on a successful experiment. The use of phrase "I do not know" is recommended as a possible choice because it allows workers to be able to indicate if they cannot answer the question logically (Alonso, 2012). Getting user feedback by asking an open-ended question is recommended to improve the quality of the experiment and this kind of question can be optional. If the answer to this kind of question is useful, the requester can pay bonus (Alonso & Baeza-Yates, 2011).

Interface provides users the accessibility and contributing to perform tasks. There are mainly some general recommendations for efficiently design interface. First, designing a user interface and instructions of experiments should be generally understandable. In some cases, perhaps instructions had better be provided in local languages as well (Khanna, Ratan, Davis, & Thies, 2010). Colors, highlights, style of typefaces and formatting improve cohesion comprehensibility (Alonso, 2012). Some verifiable questions about common-knowledge can be used to validate the workers' performance in doing tasks. This approach helps to find worthless results (Kittur, Chi, & Suh, 2008). Workers are mostly attracted to a user friendly interface and prefer a non-sophisticated interface rather than unreasonably complex user interface. As a result, when workers like the interface the speed and the quality of outcome increase (Allahbakhsh et al., 2013).

Tasks, in terms of their complications can be divided into three groups named routine, complex and creative tasks. Routine tasks are those that do not need specific expertise, such as creating relevance judgment. Complex tasks require some general skills for instance, rewriting a given text. Certain skills and expertise are essential in undertaking creative tasks such as performing research and development (Hirth, Hoßfeld, & Tran-Gia, 2012). Splitting a long and complex tasks into smaller tasks is a recommended approach because smaller tasks are more attractive to workers (Alonso, 2012). On the other, creative tasks are more prone to distract cheaters and attract more qualified and reliable workers (Eickhoff & de Vries, 2012; Kazai et al., 2013). Designing complicated tasks is a key to filter out the cheaters (Difallah, Demartini, & Cudré-Mauroux, 2012). There are different opinion about which types are workers are more reliable in performing tasks. For instance, Difallah et al. (2012) showed that, among various types of workers, those seeking fun or swaggering around are less truthful in tasks as compared with those who provoked by fulfilment and fortune. But, Eickhoff and de Vries (2012) found that the workers attracted for entertainment of doing tasks are more reliable compared to those who are doing tasks for money. To sum up, one way to achieve high-quality results (cost efficiently) is to enhance features of interface through better designing and representing HITs.

ii. Human features in crowdsourcing

A worker profile is of reputation (in accomplishments of tasks) and expertise (credentials and experience) of a worker, which can be influential in quality of results. Provided feedback by requesters about the quality of work accomplished by a worker scores in the systems and creates a worker's reputation (De Alfaro, Kulshreshtha, Pye, & Adler, 2011). Mutually, requesters need to enhance their reputations in order to increase the probability of being accepted by workers to accept their HITs (Paolacci et al., 2010).

Information such as language, location and academic degree builds credentials, but knowledge that a worker achieves through crowdsourcing system refers to experience (Allahbakhsh et al., 2013).

iii. Monetary factors in crowdsourcing

In crowdsourcing, payment affects the accuracy of results. Workers satisfied by the payment more accurately accomplished tasks than those who were left unsatisfied (Kazai et al., 2013). Monetary or non-monetary reasons can be the motivation for the workers of crowdsourcing platforms (Hammon & Hippner, 2012). A study conducted by Ross et al. (2010) to find out about motivations for workers in crowdsourcing. Money is of main incentive of 13% of Indian and 5% US workers. Another study carried out by Panagiotis Ipeirotis (2010) reported that AMT was the main income of 27% of Indians and 12% of US workers. Kazai (2011) reported that increasing the payment enhances the quality of work while Potthast, Stein, Barrón-Cedeño, and Rosso (2010) reported that higher payment only has effect on completion time rather than quality of results. In other studies reported that by increasing payment, it leads to increase in quantity rather than quality while some studies showed that considering greater payment may only influence getting the task done faster but not better as increasing payment incentive speeds up work (Heer & Bostock, 2010; Mason & Watts, 2010). However, a reasonable pay is a better cautious solution as high pay tasks attract spammers as well as legitimate workers (Grady & Lease, 2010). Indeed, the level of payment for accomplishment of a task should be reasonable and be assigned according to the level of complexity of the task.

The quality of the results can be enhanced with a set of compensation policies and inducement (Dow *et al.*, 2011; Scekic, Truong, & Dustdar, 2012). Incentives can be extrinsic like monetary bonus (e.g. extra payment), and/or intrinsic such as personal enthusiasm (Allahbakhsh et al., 2013). Rewards are generally categorized into

psychological, monetary and material (Scekic et al., 2012). The use of non-financial compensation (such as enjoyable tasks or social rewards) is more attractive to the workers and leads them to produce better quality results compared with the use of financial rewards (Mason & Watts, 2010). Workers can be motivated to provide their feedbacks and justifications by offering them some bonus. Moreover, workers who performed tasks with accuracy and high quality can be recommended for similar tasks (Alonso, 2012).

2.2.2 Quality Control in Crowdsourcing

Some workers are quite sloppy in doing their tasks, and perform them carelessly. Besides, some workers are in fact spammers using some tricks to complete their tasks. The outcome of these two type of workers are usually inaccurate. Therefore, quality control is a crucial part of crowdsourcing and it is defined as an extent to which provided outcomes fulfill requirements of the requester. There are two main approaches for quality control in crowdsourcing; (i) design-time approaches which can be applied before submitting a task, and (ii) runtime approaches used during or after submitting a task. Another classification for quality control methods are: (i) filtering workers, and (ii) aggregating labels (Tang & Lease, 2011) as discussed further in this section. These approaches and methods can be combined. In fact, filtering workers are design-time approaches and aggregating labels methods can be considered as run-time approaches (Allahbakhsh et al., 2013).

i. Design-time approaches

There are various methods of filtering workers to identify sloppy workers. Credential-based, reputation-based, and open-to-all are three methods for selecting workers in design-time approaches and can be combined as well. Credential-based method is suitable in systems where users are well profiled. Therefore, this approach is not applicable in crowdsourcing systems since users are usually unknown. Reputation-

based method selects workers according to their reputation like AMT by using approval rate parameters. Wikipedia is an example of open-to-all method through which any worker is allowed to contribute. This method is relatively easy to use and implement, and unsurprisingly unreliable workers may contribute (Allahbakhsh et al., 2013). Further to the defined methods for quality control, requesters can implement their own qualification methods (Mason & Suri, 2012) or combine different methods as they wish. Table 2.6 summarizes design-time methods in five discrete categories.

There is always a possibility for workers to perform tasks carelessly even if using qualification tests. Two main issues are complied with qualification test. First, tasks with qualification test may need longer time for being accomplished, therefore some of workers may choose those tasks without qualification test. The second issue is the cost for developing and maintaining tests continuously. Honey pots or gold standard data are pre-defined questions with known answers (Le, Edmonds, Hester, & Biewald, 2010). Those workers who answer these questions correctly will be appropriate worker for doing the task. The honey pots are faster than qualification test. They assist to identify workers who answer the questions randomly. The use of the combination of qualification test and honey pots is also applicable as a quality control method (Alonso, 2012).

Some qualification settings such as filtering workers against their origins may have some major impacts for decreasing cheater rates for certain tasks (Eickhoff & de Vries, 2012). AMT uses setting approval rate providing a metric called approval rate to pre-filter workers. The approval rate is presenting in a percentage of assignments that a worker had performed and confirmed by the corresponding requester. Therefore, this technique makes it possible to limit a task to a certain group of workers according to a range of approval rates. In AMT, those people who accomplished HITs with a high degree

of accuracy across a variety of requesters refer as master workers which of course expect higher wage demand while they usually return better quality results (Kazai et al., 2013).

Table 2.6: Design-time methods

Method	Description	Platform
Qualification test	A set of questions for the workers to qualify their performance for doing given tasks.	AMT
Honey pots or gold standard data	Pre-defined questions with known answers (Le et al., 2010). If the workers answer these questions correctly, they will be marked as appropriate workers for that task.	Crowdflower
Qualification settings	These settings are set when creating HITs.	AMT, Crowdflower
Trap questions	In designing HITs, this set of questions (with known answers) can be included to identify unreliable workers (Zhu & Carterette, 2010).	-
CAPTCHAs and reCAPTCHA	CAPTCHAs is an anti-spamming technique to discriminate humans from machines in order to filter out answers provided by computers (Von Ahn, Blum, Hopper, & Langford, 2003). reCAPTCHA is a development of CAPTCHAs (Von Ahn, Maurer, McMillen, Abraham, & Blum, 2008).	-

A higher approval rate for the quality control, the longer time it may needs to complete an experiment since a lesser number of workers (according to the approval rate) are considered alligible (Alonso & Baeza-Yates, 2011). This setting is generally task dependent. For instance, a more complicated task needs more stringent approval rate and setting amongst master workers. Whilst, a routine task needs ordinary workers through a simple setting. Trap questions are helpful to detect uncaring workers who do the tasks carelessly. Kazai, Kamps, Koolen, et al. (2011) designed two trap questions to avoid this situation; “Please tick here if you did NOT read the instructions” and “I did not pay attention”. Unsurprisingly, not all of the unreliable workers may be detected by trap questions but it showed strange behavior and it can be effective in both discouraging and identifying spammers.

Some programs or bots are designed to accomplish HITs automatically in crowdsourcing (McCreadie, Macdonald, & Ounis, 2010) which provide poor quality results (Mason & Suri, 2012). In this scenario, CAPTCHAs and reCAPTCHA (easily and cost effectively) help experiments to detect automated results generated by malicious software in crowdsourcing (Khanna et al., 2010; Kazai et al., 2013). It is also possible to combine above mentioned methods to filter out uncaring workers. For example, in recruiting workers and monitoring the quality of their works a real time strategy was to applying different methods to filter out workers. At first, a qualification test was used to filter sloppy workers. The completion time of HITs was then calculated to reflect on the truthfulness of the workers. Besides, a set of gold questions were used to evaluate the skills of workers (Tao Xia, Zhang, Li, & Xie, 2011).

ii. Run-time approaches

Although design-time techniques can intensify quality, there is the possibility of low quality because of misunderstanding while doing the tasks. At the same time, runtime techniques are essential for high quality results. Indeed, quality of the results would be increased by applying both approaches (Allahbakhsh et al., 2013). Table 2.7 outlines various methods of run-time quality control. In crowdsourcing, if we assume that one judgment per example or task called single labeling method, the time and cost of the experiment in a crowdsoursing may be saved. However, the quality of work is dependent to an individual‘s knowledge about the task. Integrating labels from multiple workers is to solve the issue of single labeling methods (Sheng, Provost, & Ipeirotis, 2008; Welinder & Perona, 2010). If labels are noisy, multiple labels can be desirable to single labeling. Integrating labels increase the level of accuracy for relevance judgments (Hosseini, Cox, Milić-Frayling, Kazai, & Vinay, 2012). The main issue for the multiple labeling is to

accurately and effieciently comply a single consensus label from aggregating various labels.

Table 2.7: Run-time methods

Method	Description
Majority Voting (MV)	is a straightforward and common method, which discriminates wrong results according to the decision of majority (Sheng et al., 2008; Snow, O'Connor, Jurafsky, & Ng, 2008; Hirth et al., 2012).
Expectation Maximization (EM) Algorithm	measures the quality of a worker according to the accuracy of answers to the tasks on the basis of labels completed by different workers using maximum likelihood. This algorithm has two phases; (i) the correct answer is estimated for each task through multiple labels submitted by different workers, accounting for the quality of each worker (ii) comparing the assigned responses to the concluded accurate answer in order to estimate quality of each worker (Dawid & Skene, 1979).
Naive Bayes (NB)	is a method to model the biases and reliability of single workers and to correct them in order to intensify the quality of the workers' results. According to gold standard data, a small amount of training data that labeled by an expert was used to correct the individual biases of workers. The idea is to recolibrate answers of workers to be more matched with experts (Snow et al., 2008).
Observation of the Pattern of Responses	some untrustworthy workers have certain patterns to answer tasks. For example, they may select the first choice of every question. Therefore, pattern of answers provides an effective way to filter out unreliable responses.
Probabilistic Matrix Factorization (PMF)	is an standard method in collaborative filtering through converting a crowdsourcing data to collaborative filtering data to predict unlabeled labels from workers (Jung & Lease, 2012). PMF defines a latent feature vector for each worker and example to infers unobserved worker assessments for all examples (Salakhutdinov & Mnih, 2008).
Contributor Evaluation	the workers are evaluated according to certain quality factors such as reputation of workers in the field, their experience, and/or their credentials. Requesters accept the tasks if the workers have enough quality factors. Tasks submitted by the workers of higher approval rates would be assumed correct. Wikipedia, for instance, accepts those article written by administrators without further evaluation (Allahbakhsh et al., 2013).
Real-Time Support	is to provide workers with the requesters' feedbacks about their quality of work in a real time manner. This method enhances the quality of tasks performed by workers because they receive the feedback about results, which provides a kind of self-assessment to improve their performances (Dow et al., 2011). Requesters can also follow the workflows of workers solving tasks (Kulkarni, Can, & Hartmann, 2012) . Turkomatic is a tool to identify workers of tasks through which requesters are able to monitor the process and review the status of a task in real time manner.

The Majority Voting (MV) is a reasonable choice for routine tasks as it is usually of lower payments and it is relatively easy to implement and to achieve acceptable results which is of course depend on the truthfulness of workers (Tang & Lease, 2011; Hirth et al., 2012). The weakness of this method is that the consensus label is measured for a specific task without considering the accuracy of the workers in other tasks. Moreover, MV considers all workers are equally good. For example, if there is a minority of experts and a majority of novices who provided the same but inaccurate responses to a task, the MV conclude that the novices' answer is the correct answer just because they are of the majority. A set of estimated accurate answers for each task and a set of matrixes that include the list of workers errors produces Expectation Maximization (EM) algorithm, a quality control method, by which the error rate of each worker can be accessed. In order to measure the quality of a worker by EM algorithm, the error rate is not an adequate measurement since the workers may have completed the task carefully but with bias. Given the example of labeling websites, parents with younger children were more conservative in classifying the websites. Therefore, to compensate this situation, a single scalar score would be assigned to each worker, corresponding to the completed labels. The scores separated the error rates from worker's bias and satisfactory treatment.

Variety of EM algorithms have been proposed such as a bayesian version of the EM algorithm using confusion matrix (Carpenter, 2008; Raykar *et al.*, 2010). A probabilistic framework was proposed by Raykar et al. (2010) which is usable when there is no gold standard with multiple labels. A specific gold standard is created repeatedly by proposed algorithm and based on this gold standard the performances of workers are assessed. Considering the time which takes to complete a task is an example of observing responses' patterns, that is to determine random answers produced by unreliable workers (Kittur et al., 2008). Those tasks completed fast are deemed poor quality; completion time

is a robust method for detecting sloppy workers. In another study, the time that each worker spent on judgment was assessed as a quality control (Zhu & Carterette, 2010), during which three types of pattern were found. The first pattern was considered normal pattern whereby the workers began slowly and get faster when they learn about the task. The second pattern called periodic pattern was a peculiar behavior since some of the judgments were performed fast and some slow. Interrupted pattern referred to disruption in performing tasks. The method of observation of the pattern of responses along with the other methods for quality control enhance the effectiveness of crowdsourcing experiments. Different quality control methods were applied in crowdsourcing experiment, which are listed and discussed in both design-time approaches and run-time approaches. Accordingly, proper quality control methods should be well-suited to crowdsourcing platform. The next section is mostly focused on the studies in which crowdsourcing was applied for IR evaluation.

2.2.3 Crowdsourcing in IR Evaluation

Alonso et al. (2008) are the pioneers of using crowdsourcing for obtaining relevance judgments in IR evaluation through AMT on TREC data. Crowdsourcing method in IR evaluation has been adapted massively in recent years (Alonso & Mizzaro, 2009; Grady & Lease, 2010; Lease & Kazai, 2011; Zuccon *et al.*, 2012; Kazai et al., 2013; Lease & Yilmaz, 2013; Kazai, 2014). Anonymous workers can online for crowdsourcing to create relevance judgments as a replacement for editorial staff who are expert in the fields. Crowdsourced workers usually are not trained for relevance judgments and potentially are from various backgrounds and have different levels of motivation for performing tasks. Therefore, crowdsourced relevance judgments may be varied in terms of quality and accuracy. The main criticism against crowdsourcing is for its diverse quality outputs (inconsistency), which is of course the core challenge to ensure about the

quality of crowdsourcing output (Kazai et al., 2013). This challenge is still vague whether crowdsourcing can overtake the traditional methods for relevance judgments.

Alonso and Mizzaro (2009) ran five preliminary experiments by different alternatives, such as qualification tests and changing interface, through AMT using TREC data and measured the agreement between relevance judgments made by crowdsourced workers and TREC assessors. The findings showed that the judgments of crowdsourced workers were comparable with that of the TREC assessors. In some cases, the workers detected TREC assessors' errors. In another study conducted by the same research group in 2012, a comprehensive experiment validated the use of crowdsourcing for creating relevance judgments (Alonso & Mizzaro, 2012). The experimental results show that crowdsourcing is relatively lower in cost, but reliable giving quick solutions as an alternative for creating relevance judgments by expert assessors. However, it is not a replacement for current methods because of several gaps and shadows, which are left for future research. For instance, scalability of crowdsourcing has not been fully investigated yet, although the reproducibility of crowdsourced evaluation was investigated in a study (Blanco *et al.*, 2011). In this study after a period of six months and with the use of different evaluation measures and system rankings, the crowdsourcing experiment was repeated and produced a similar output, showing that crowdsourcing experiments can be repeated over time in a reliable manner. Despite some differences in judgments between human expert and crowdsourced, the system ranking was the same.

Another study was also examined the reliability of using crowdsourcing in multi labelled images by conducting different experiments. For instance, in an experiment the agreement between 11 expert annotators showed high consistency and agreement, showing a high level of correlation in system rankings. In another experiment doing the same tasks but with non-expert annotators in crowdsourcing showed a high level of

agreement between crowdsourced annotators and an expert using MV to aggregate the non-expert annotators. System rankings of non-expert was highly correlated with system ranking generated by expert annotator. In the other word, MV method (for aggregating the annotations and to filter out noisy judgments) is reasonably beneficiary especially when there are disagreements between experts and non-experts judgments, through reducing the effects of variation in relevance judgments on system rankings (Nowak & Rüger, 2010).

In 2011, Kazai et al. investigated the relationship between workers' behavioral patterns, their personality profiles, and the accuracy of their judgments. The difference was based on behavioral observation including (i) label accuracy, (ii) HIT completion time and (iii) fraction of useful labels. The study investigated, whether the behavior and personality of workers are able to influence the label accuracy through designing two different HITs (namely Full Design (FD), a strict quality control, and Simple Design (SD), reduced the quality control compared with FD). The study correlated the worker types and personality trait information, with the accuracy of labels, considering the 'Big Five' personality dimensions (John, Naumann, & Soto, 2008) (namely openness, conscientiousness, extraversion, agreeableness and neuroticism). Using behavioral patterns method, various types of workers (spammer, sloppy, incompetent, competent, and diligent) were identified and as a result a strong correlation between the accuracy of judgments and the openness trait were reported (Kazai, Kamps, & Milic-Frayling, 2011).

The impact of task design on the quality of labels has been assessed in several researches. In a study for book search evaluation with two different HIT design, FD and SD. The FD leads to higher label quality compared with SD. Moreover, it was reported that crowdsourcing is a useful method for creating relevance judgments for IR evaluation, but tasks design needs to be done carefully, as different HIT designs lead to a significant

difference in agreement between crowdsourcing and the gold set (Kazai, Kamps, Koolen, et al., 2011). In 2012, Kazai et al. studied the relationship between demographics, personality of workers and label accuracy with two different HIT designs, the FD and SD. The results showed that the demographics and personality of the workers were strongly related to label accuracy. Among demographic factors, location had the strongest relationship with label accuracy, with the lowest accuracy from Asian workers, and the higher accuracy from American and European workers. Asian workers were more likely to undertake the SD, while American and European workers were more likely to undertake the FD—though the difference may have been an artifact of the pre-filtering in FD, in which workers without a sufficient AMT reputation score were filtered out (Kazai et al., 2012).

The effects of the level of pay, effort to complete tasks, and qualification needed to do the tasks, on the quality of the labels were investigated while correlating them with various human factors. Variety of information including perceived task difficulty, satisfaction with the offered pay, motivation, interest, and familiarity with the topic, were obtained from the workers to see how they influence label quality, along with aspects of the task design. A higher level of payment led to high quality of an output. However, this may also attract unethical workers to participate. On the other hand, higher efforts for HITs increased the probability of inaccurate labels, but enticed workers with higher performances. In addition, when the number of judgments that need to be made in a HIT increased, it led to increase productivity. Since achieving fewer judgments per HIT, decreased the possibility of detecting low quality judgments due to workers' limited exposure. Lower effort HITs had a faster overall task completion. Limiting HITs to workers that were more reliable increased the quality of the results. Therefore, a simple pre-filtering, such as filling captcha fields, helps to find unreliable workers. Obviously,

the pre-filtering application was not sufficient but aided to enhance the quality of the labels. Earning money was the main reason and motivation for workers to do the tasks. In fact, those workers who performed tasks for "Fortune and Fulfilment" were of the most precise workers, comparing to those who accomplished for "Fun and Fame". Self-reported information about familiarity with the topics seems unreliable. For instance, workers who reported higher familiarity with a given topic showed lower performances as compared with the workers who described the tasks boring. Satisfaction with the pay had a strong relationship with label accuracy, since the workers who were satisfied with pay were the most accurate workers (Kazai et al., 2013).

Clough, Sanderson, Tang, Gollins, and Warner (2012) compared the reliability of crowdsourced and expert judgments when used in IR evaluation. They evaluated two search engines on informational and navigational queries, using crowdsourced and expert judgments. The study found that the crowdsourced judgments are comparable to expert judgments, with a strong positive correlation between search effectiveness measured by each class of judgments. In terms of correlation between expert judgment and crowdsourced workers, the disagreements were more common on documents returned by the better performing system and on documents returned for informational queries.

Various studies in recent years, which was explained in this section, have focused on the reliability of using crowdsourcing in IR evaluation and the factors, which influence reliability of crowdsourced relevance judgments. None of these studies considers cognitive characteristics of crowdsourced workers and its possible effect on reliability of crowdsourced relevance judgments in IR evaluation. However, the cognitive characteristics of users have been recognized as a factor, which affect IR process. In the following section, the cognitive abilities and its application in previous studies especially IR process are reviewed.

2.3 Cognitive Abilities

2.3.1 Cognitive Ability Definition

Individual differences in cognitive performance is defined as cognitive abilities. The terms intelligence, aptitude and cognitive abilities, which are substitutable, are commonly defined as the learning ability, adapting new situations and solving problems. Internal cognitive abilities are also associated with problem solving performance (Mayer, 1992). In another word, a complex combination of cognitive abilities is intelligence. Cognitive abilities are the ability to understand, remember, reason language and locate material, which presented visually. Variety of research studies investigated these cognitive abilities. For instance, cognitive abilities have been introduced as an important indicator of people performance in certain jobs in management research. Library research has shown that users with high level of cognitive abilities utilize IR systems more efficiently. Various studies have shown that cognitive abilities are important in performance of technology-based tasks (Charness, Kelley, Bosman, & Mottram, 2001; Czaja, Sharit, Ownby, Roth, & Nair, 2001; Sharit, Czaja, Nair, & Lee, 2003). Cognitive abilities is also an important factor in everyday activities (Czaja *et al.*, 2006).

Cognitive style is about learning performance and preferences, and is independent from intelligence. Sometimes the differentiation between cognitive ability and cognitive style is confusing. The cognitive style indicates person's patterns of thinking and problem solving. Cognitive style is basically related to observable behaviours such as learning performance (Karahoca, Karahoca, & Güngör, 2008). This current study focuses on cognitive abilities rather than cognitive style. Hunter (1986) showed that cognitive abilities could anticipate the level of performance in work. Therefore, it seems essential to assess the impact of individual differences in cognitive characteristics on reliability of relevance judgments in crowdsourcing.

There are a number of measuring instruments to examine cognitive abilities such as Wonderlic Cognitive Ability Test (Wonderlic, 1961). The Wonderlic test is a group intelligence test applied to examine the aptitude of employees for problem solving. The Quick Word Test is another test used to measure verbal knowledge. This test consists of 100-item vocabulary test and considered as a substitute for measures of cognitive abilities (Borgatta & Corsini, 1964). The cognitive abilities can be also evaluated based on the IQ test (Karahoca et al., 2008). One of the popular instruments to assess cognitive abilities is Kit of Factor-Referenced Cognitive Test (FRCT) (Ekstrom et al., 1976). This test contains 72 tests to measure 23 different cognitive factors. A number of these 23 factors are listed in Table 2.8 (Ekstrom et al., 1976).

Table 2.8: Cognitive abilities in FRCT

Cognitive ability	Definition
Verbal Comprehension	"The ability to understand the English language"
General Reasoning	"The ability of select and organize relevant information for the solution of a problem"
Logical Reasoning	"The ability to reason from premise to conclusion, or to evaluate the correctness of a conclusion"
Perceptual Speed	"Speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual perception".
Spatial Scanning	"Speed in exploring visually a wide or complicated spatial field"
Visualization	"The ability to manipulate or transform the image of spatial patterns into other arrangements"
Associative Memory	"The ability to recall one part of a previously learned but otherwise unrelated pair of items when the other part of the pair is presented"

Verbal comprehension is depend on the contents of the long-term memory. Logical reasoning is the skill of evaluating the correctness of the answer. Mathematical reasoning is mostly used to evaluate general reasoning. General reasoning is similar to logical reasoning (Carroll, 1974), in which the type of content of long-term memory (retrieved and applied) are of differentiating factors. Some researchers combine logical

reasoning and general reasoning (Ekstrom et al., 1976). Some people with high perceptual speed are able to scan contents and judge about what they see. Spatial scanning is the ability to scan quickly for comprehension. Visualization is like the imagination of a piece of paper in its various stages from being folded to the end, being completely unfolded. Therefore, thinking sequentially is also required for this ability. Associative memory or intermediate memory involves when a person deliberately think of specific information.

2.3.2 Cognitive Abilities in Information Retrieval Process

Information search principally is of cognitive activity. Therefore, understanding the effect of cognitive abilities on search behaviour is an important topic. During a search process, the mental process is organized by cognitive abilities of the user (Brennan et al., 2014). Cognitive abilities have been considered as individual difference in information science research. Cognitive abilities prevent a user from confusion during a search process. Information retrieval is in fact the interaction between the systems (provide information) and users (need information). During a retrieving information process, users of IR systems use a variety of cognitive abilities such as memorizing, comprehending, and problem solving. Various studies in IR systems highlight the importance of cognitive abilities in information work (Allen & Allen, 1993; Ford et al., 2000), emphasizing the critical roles of cognitive abilities in IR processes.

Table 2.9 summarizes related researches in cognitive abilities in IR process. A study which investigated the roles of knowledge and cognitive abilities in older adult information seeking on the web, claimed that these cognitive abilities have a greater role for older adults since the problem solving process is complex for them (Sharit, Hernández, Czaja, & Pirolli, 2008).

Table 2.9: Summary of researches in cognitive abilities in IR process

Objective	Findings	Reference
To investigate the roles of knowledge and cognitive abilities in older adult information seeking.	Cognitive abilities have a greater role for older adults.	(Sharit, Hernández, Czaja, & Pirolli, 2008)
To investigate the effect of perceptual speed, logical reasoning, spatial scanning and verbal comprehension abilities on how their performance were in searching.	Students had higher level of perceptual speed and librarians had higher level of logical reasoning and verbal comprehension abilities.	(Allen & Allen, 1993)
To investigate the effects of three cognitive abilities (visualization ability, perceptual speed and memory) on search behaviors.	Perceptual speed and visualization ability were highly correlated with search behavior.	(Teitelbaum-Kronish, 1984)
To explore effects of perceptual speed, verbal comprehension, logical reasoning on search effectiveness of the users in a CD-ROM bibliographic search task.	verbal comprehension and logical reasoning influenced information seeking and those users with higher perceptual speed performed better in searching	(Allen, 1992)

Studies in the field of library setting showed a correlation between cognitive abilities and performance. One study, investigated (using FRCT) the effect of perceptual speed, logical reasoning, spatial scanning and verbal comprehension abilities on how well academic librarians suited their jobs and how their performance were in searching (Allen & Allen, 1993). The results of this study showed that students had higher level of perceptual speed and librarians had higher level of logical reasoning and verbal comprehension abilities. As the cognitive abilities influence IR performance, different methods to IR may be suitable for students and librarians. In a separate study, there was a positive relationship between logical reasoning ability and the performance in online searching (Teitelbaum-Kronish, 1984). Measured by the FRCT, three cognitive abilities (visualization ability, perceptual speed and memory) were tested to assess the effects of cognitive abilities on search behaviours during search tasks (Brennan et al., 2014). The

result showed that perceptual speed and visualization ability were highly correlated with search behaviour.

Study on the effect of certain cognitive style on IR effectiveness showed that users with verbalizer styles had deprived retrieval performance as compared with those with imager styles (Ford, Miller, & Moss, 2001). Although there are a number of studies assessing the effect of cognitive style on IR process, limited researches have been conducted for the evaluation of the effect of cognitive abilities on IR process. Cognitive abilities such as, perceptual speed, verbal comprehension, logical reasoning and spatial scanning are influential in search performance (Allen, 1992). For instance, in a CD-ROM bibliographic search task, verbal comprehension and logical reasoning influenced information seeking and those users with higher perceptual speed performed better in searching (Allen, 1992). Verbal comprehension ability was also found to be strongly predict performance in a simulated telecommuting task in which older adults were required to respond to queries from fictitious customer emails by navigating through a database configured in the form of a hyperlinked information environment similar to the Internet (Sharit *et al.*, 2004). Other specifications of users such as prior search experience and cognitive abilities on search effectiveness of the users were studied among 56 users given 56 topics and assessed by the TREC test collection (Al-Maskari & Sanderson, 2011). The users with higher perceptual speed and prior search experience performed better than those users with less experience and slower perceptual speed abilities. Need for Cognition (NfC) defines as an individual difference measure of "the extent to which a person enjoys engaging in effortful cognitive activity". Study of the impacts of NfC on relevance assessments showed that the participants with high NfC had a significantly higher level of agreement with expert assessors in terms of relevance assessment than low NfC participants (Scholer, Kelly, Wu, Lee, & Webber, 2013).

The results from the previous studies in cognitive abilities led the author to wonder if cognitive abilities influence the crowdsourced relevance judgments. The author thought it is possible that people with higher level of cognitive abilities would be more likely to create more accurate relevance judgments. This idea is derived from previous studies, which demonstrated that cognitive abilities influence IR processes. In the same way, the author predicts that IR practitioner, in choosing individuals to create relevance judgments would be likely to select workers with higher level of cognitive abilities. This idea is also suggested by management practitioner, which emphasized that administrators search for individuals with specific skills for their institutions. Accordingly, the hypothesis for this study is that crowdsourced workers with higher level of cognitive abilities would display higher reliability in creating relevance judgments. To the authors' knowledge, however, no studies have investigated the cognitive abilities of crowdsourced workers and their effect on the workers' reliability in judging the relevance of documents.

2.4 Summary

In this chapter, first, the process of information retrieval evaluation has been discussed. Followed by explanation on TREC, which providing a large test collection for experiments. Details on test collection were provided, along with explanation of common evaluation measures in IR evaluation. Relevance evaluation were explained in detail in this chapter as well. Then, an introduction into crowdsourcing was provided along with explanation on factors affect the reliability of crowdsourcing output. Followed by description on different quality control methods in crowdsourcing. Different studies which used crowdsourcing in IR evaluation was discussed in this chapter. Finally, cognitive abilities were defined along with its application in IR process. The next chapter will discuss the methodology used in this study in detail.

CHAPTER 3: RESEARCH METHODOLOGY

Prior to the full-length methodology, a brief introduction on the study design provides a comprehensive overview on the study. The initial idea about the project was reinforced by reviewing several articles and research work in the field of study, by which the research problem could be defined. The research problem was then evaluated by conducting a pilot study. Accordingly, we finalized the methods and developed an appropriate study design, which was based on the hypothesis, and a pilot study. The experimental design consisted of three experiments. Subsequently, acquired data was analyzed statistically to evaluate our findings, which were eventually compared with other research works in the field of study. Finally, based on the results of the experiments, two approaches were proposed. Figure 3.1 provides the research flow chart of the steps through this study. This chapter explains methods, experimental design, experiment data, task design and filtering method that we used in this study. Moreover, this chapter provides detailed explanation about methods of measuring judgment reliability and the statistical methods used in this study. Finally, the pilot study is explained in detail.

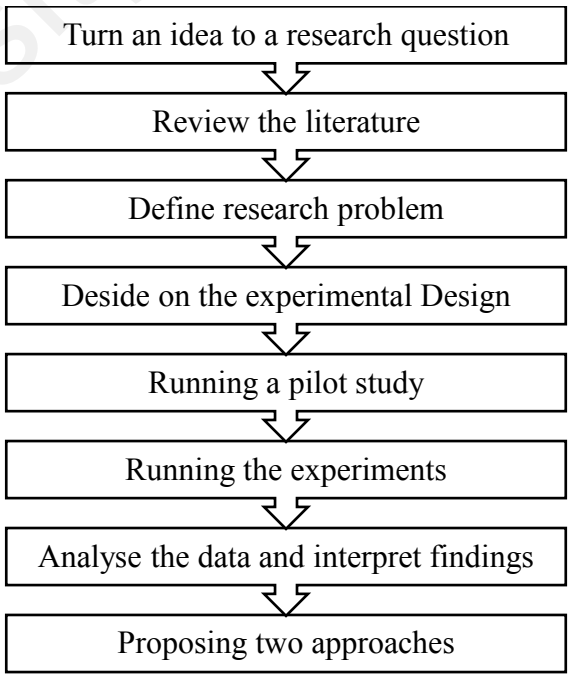


Figure 3.1: Flowchart of this study

3.1 Experimental Design

In a searching process, users accomplish a variety of tasks such as defining the search query, choosing proper search vocabulary, issuing commands, observing retrieved information and judging about relevancy and usefulness (Allen, 1994b). In 1998, Sutcliffe and Ennis introduced a theoretical framework for modelling information-seeking behaviour which were consisted of four cyclical cognitive activities: "problem identification, need articulation, query formulation, and results evaluation" (Sutcliffe & Ennis, 1998). *Problem identification* consists of defining the information need. *Need articulation* involves expressing the information need by selecting terms from long-term memory. The process of query generation performs during *Query formulation.* The last is the decision making process, *results evaluation*, during which the users make their decisions about retrieved results. Exploring the cognitive abilities of workers in the current study was chosen based on a theoretical understanding of the IR process, which suggests these abilities may be most likely to influence IR effectiveness, as indeed previous studies have found. In this study, we hypothesize that the same relationship will be pertained to the relevance assessment, as understanding the content of documents and topics in the relevance judgment task requires reading and understanding text, which refer to "verbal comprehension skill".

Evaluating the relevancy of given topics and documents requires arguments about evidences to conclude and solve the problem which refers to "logical reasoning skill "and "general reasoning skill". Those three cognitive abilities (verbal comprehension, general reasoning and logical reasoning) were selected since they were already proved to predict successful performance in IR tasks. Moreover, "verbal comprehension", "general reasoning" and "logical reasoning" are of the main cognitive abilities influencing informational retrieval behavior (Allen, 1992; Allen & Allen, 1993). Therefore, these three skills were chosen in this study, as they are potentially important in judging the

relevance of a document. However, these three cognitive abilities should not be considered to be representative of all aspects of cognitive abilities. This research work were consisted of three experiments. The first experiment was verbal comprehension experiment to examine the effect of verbal comprehension skill on reliability of crowdsourced relevance judgments. The second experiment was to evaluate the effect of general reasoning skill on reliability of relevance judgments in general reasoning experiment.

Logical reasoning experiment was the third experiment during which the effects of logical reasoning skill on reliability of crowdsourced relevance judgments were assessed. The main hypothesis of this study is that crowdsourced workers with different levels of cognitive abilities have a positive correlation (association) with their reliability of relevance judgments. Furthermore, we investigated whether when the assessments of workers with higher cognitive abilities are used to evaluate and rank retrieval systems by effectiveness, they provide a similar ranking to that of the TREC expert assessments. Besides, the relationship between self-reported difficulty of the task, confidence of the worker, and worker's knowledge about the topic, and the reliability of the relevance judgments were tested. Therefore, through the course of this study we will see if more reliable judgments are produced by those workers who report the task easy, and have higher levels of confidence in their judgment, and in their knowledge about the topics. The assessment about the relationship between demographics (age, gender, education, country, computer and Internet experience) of workers and reliability of their relevance judgments was the last part of our study.

3.2 Experimental Data

Ten topics were taken from the TREC 2011 Crowdsourcing Track[5] . All of the topics were open-ended information need, which was referred to answering an open-ended question (See Table 3.1). Ten documents from the ClueWeb09[6] dataset were chosen randomly for each topic. All documents and topics were in English. The chosen documents contained highly-relevant, relevant and non-relevant documents, as judged by the original TREC assessors. Topics and documents were chosen based on the number of available relevance judgments.

Table 3.1: Topics in this study

Number	Topic	Description
20644	vice president richard nixon	Information about Richard Nixon as Vice President.
20696	stars supernova	What are stars supernova?
20714	paramount pictures	What do we know today about Paramount pictures?
20764	green darner dragonfly	I am looking for information on the green darner dragonfly.
20766	prescription diet pills	Find information about prescription diet pills.
20814	elvish language	I want information about the Elvish language.
20916	doughton park	What is Doughton Park?
20922	virtual earth	What is virtual earth?
20958	lake murray fishing	Why is Lake Murray fishing popular?
20976	Sudoku	I am looking for information on sudoku puzzles.

For each of the 100 <topic, document> pairs, 50 graded relevance judgments (highly-relevant, relevant, non-relevant) were obtained through crowdsourcing using Crowdflower, each one from a different worker, in total 5000 judgments made by workers for each experiment. In Section 3.3, the explanation of how these tasks designed is provided in more details. The relevance judgment set created by the official TREC

[5] https://sites.google.com/site/treccrowd/2011
[6] http://www.lemurproject.org/clueweb09.php

assessors (*qrels*) were deemed as our gold standard dataset, to which relevance judgments of the crowdsourced workers were compared.

3.3 Designing Tasks

In this study, there were 20 HITs designed in Crowdflower. Each HIT was to be completed by 50 workers, for a total number of 1000 HITs. Figure 3.2 presents a schematic of the task design procedure for the three experiments.

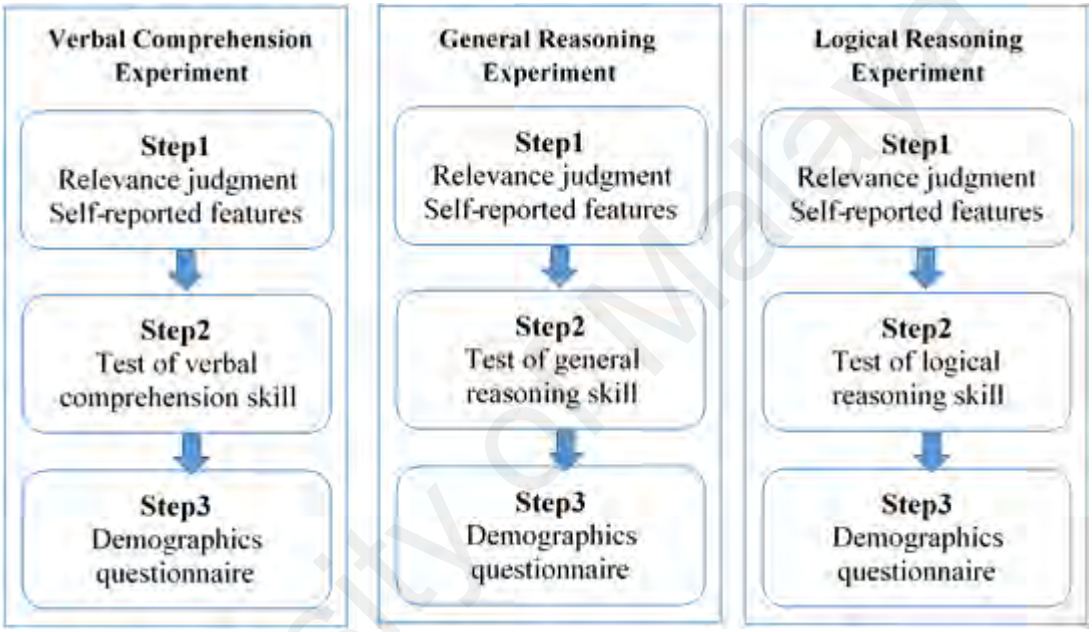


Figure 3.2: Task design of each experiment

From the three steps, the first and the third steps were the same in the three experiments. However, the second step was different for each experiment. The first step was designed to collect relevance judgments from workers in order to evaluate their reliability of relevance judgments. Each HIT consisted of one topic, and five documents to be assessed for a relevance judgment against that topic. After completing each judgment, the workers were asked to fill out a self-reporting questionnaire (prepared in the 4-point scale format), composed of three items, to declare the level of difficulty of the judgment, their knowledge about the given topic and how confidence they were in their

judgment. The procedure provided useful information about the association (if any) between the level of self-reported competence and reliability of relevance judgments.

Q1) Rate your knowledge on the topic: (Minimal 1 2 3 4 Extensive)

Q2) How difficult was this evaluation: (Easy 1 2 3 4 Difficult)

Q3) How confident were you in your evaluation: (Not confident 1 2 3 4 Very confident)

The relevance judgment task setting was different from the classic judgment method by TREC assessors. In the classic method, an assessor judges all documents from the same topic. In this study, the use of crowdsourcing platform lead to have a topic and five documents in each task. One of the main attraction of the crowdsourcing to the workers is that crowdsourcing tasks are commonly short (Alonso, 2012). Therefore, a long and complex task can be split into some short and simple tasks. Kazai et al. (2013) also emphasized to break-down tasks into simply digestible units for the workers and it was mentioned that smaller tasks are a better fit within crowdsourcing model.

In the second step, the three cognitive abilities (verbal comprehension, general reasoning and logical reasoning) were assessed to evaluate the level of cognitive abilities of the workers. The tests for this assessment were based on the suite of evaluation exercises known as the Factor-Referenced Cognitive Tests (FRCT), produced by the US-based Educational Testing Service[7] (Ekstrom et al., 1976). FRCT is a widely accepted and standardized research tool for studying the cognitive processes (Geary, Hoard, Nugent, & Bailey, 2013; Beaty, Silvia, Nusbaum, Jauk, & Benedek, 2014; Salthouse, 2014). The aim of the test kit is to provide researchers with a mean for identifying specific aptitude factors. In this study, verbal comprehension was measured by a test called *Extended Range Vocabulary Test,* which consisted of 24 vocabulary questions. The

[7] http://www.ets.org

workers required to choose one of five words that has the same meaning as the given word. The verbal comprehension score is the number of correct answers minus a fraction of number of wrong answers from the given 24 vocabulary questions. An example of this test is presented in Figure 3.3. In this study, *Necessary Arithmetic Operations Test* was the measurement tool to assess general reasoning skill. The workers completed a test of 10 general reasoning questions out of 15 questions (obtained from *Necessary Arithmetic Operations Test*) to examine their general reasoning skill. Helping them for the calculations, the workers were asked to choose proper numerical operations to solve arithmetic problems. General reasoning score was then calculated through; number of correct answers minus a fraction of number of wrong answers from the given 10 questions. An example of the *Necessary Arithmetic Operations Test* is shown in Figure 3.4.

Cottontail: (1) Squirrel (2) Poplar (3) boa (4) marshy plant (5) rabbit

The correct answer of this question is "rabbit".

Figure 3.3: Example of Extended Range Vocabulary Test

There are 4 quarts in a gallon and 4 cups in a quart. How many cups are there in a gallon?

(1) add (2) subtract (3) multiply (4) divide

The correct answer is "multiply".

Figure 3.4: Example of Necessary Arithmetic Operations Test

The intention behind using 10 questions from the *Necessary Arithmetic Operations Test* out of 15 questions was that as explained before, due to the using crowdsourcing it is better to limit the number of questions in a task. Moreover, each question needs thinking and calculation and we thought having more questions in a task may lead to be boring for workers and as a result, the workers may do the task carelessly.

But there was a question if the use of 10 questions instead of 15 questions for *Necessary Arithmetic Operations Test* could be statistically acceptable. Therefore, we assessed relationship between the 10-questions and the 15-questions for general reasoning skill of 47 participants, by comparing the outcomes in a separate experiment in crowdsourcing.

Two general reasoning scores were calculated for each participant; one for the 15-questions set and another one for the 10-questions set. According to the percentile of general reasoning scores split, the workers were categorized into three groups, namely low, moderate and high general reasoning scores two times, based on two scores. The kappa for goodness of fit was then calculated to find out whether there was an agreement for the grouping (Pallant, 2001) between the 10-questions set and the 15-questions set. Kappa measure of agreement was to evaluate the consistency of the two sets of questions, showing a substantial agreement between the two sets (Kappa=0.61). Therefore, in this study, the use of the 10-questions set (instead of the 15-questions set) could compensate the limitation of crowdsourcing, and could provide a statistically meaningful cohort to assess workers' general reasoning skill.

The logical reasoning was measured in this study using *Nonsense Syllogisms Test*. The workers required to answer a test consisted of 10 questions (out of 15 questions) to measure their ability to tell whether the conclusions drawn from certain statements were either correct or incorrect. The logical reasoning score was calculated by subtracting the number of wrong answers from the number of correct answers for the 10 given questions. Figure 3.5 shows an example of *Nonsense Syllogisms Test*. The number of questions used for the *Nonsense Syllogisms Test* was 10 (out of 15 questions). To find out whether 10 questions conveyed statistically acceptable output for logical reasoning skill, in another experiment using Crowdflower, we compared the outcomes of the 10-questions with the 15-questions set for 40 participants. Scores for each of logical reasoning question sets

were used to calculate the percentile of logical reasoning scores split. Accordingly, the participants were categorized into three groups, namely low, moderate and high logical reasoning scores. Agreement between the 10-questions set and the 15-questions set was calculated using kappa for goodness of fit. The results showed that there is a substantial agreement between the two sets of questions (kappa=0.72). Therefore, in order to reduce the number of questions for logical reasoning skill to be used in crowdsourcing, the 10-questions set was implied in this study to assess workers' logical reasoning skill.

All birds have purple tails. All cats are birds. Therefore, all cats have purple tails.

(1) True (2) False

The correct answer is "True".

Figure 3.5: Example of Nonsense Syllogisms Test

In each experiment, according to their scores for certain cognitive abilities, workers are divided into three groups. This grouping is based on percentile split, Group 1 consists of low scores workers, Group 2 of moderate scores, and Group 3 of high scores workers. The rationale behind categorizing workers into three groups is to see whether relevance judgments generated by workers with higher level of cognitive abilities are more agree with relevance judgments provided by TREC assessors than relevance judgments created by workers who have lower level of cognitive abilities.

During the last step, the workers completed a set of five questions (including a trap question- discussed in the following section) about their demographic information. The demographic questions acquired some information about age, gender, educational, computer experience and the Net experience for each worker. This information was then used to find statistical association (if any) with reliability of their relevance judgments. A summary of the demographic items is presented in Table 3.2.

Table 3.2: Survey questions

Demographics	Level
Age	not yet 20
	in my 20's
	in my 30's
	in my 40's
	in my 50's
	60+ years old
Gender	Male
	Female
Level of Education	No education
	Basic schooling
	High school
	Bachelor degree
	Master degree
	PhD or higher
Level of Experience with Computer	1
	2
	3
	4
Level of Experience with Internet	1
	2
	3
	4

3.4 Filtering Spam

Crowdflower provides the requesters with several qualification settings to filter the workers based on certain specifications. For instance, by choosing various parameters within "job setting", a requester can apply a pre-filter to the workers for performance, ranged from highest speed to highest quality. In this study, we chose high quality option, which of course took longer time to complete a task. Therefore, the HITs was open only to certain group of worker who had the qualification. Another available filtering parameter is to specify geographical regions to contribute in the tasks. As a result, it is possible to restrict the workers for example to English-speaking countries such as USA, UK and Australia. This setting was also applied to this study because all of the tasks were prepared in English.

Crowdsourcing is quite attractive to those workers who are not trustworthy in their performance in the tasks. This group of workers are those who may complete tasks fast but carelessly, and their main motivation is just to earn money (of least effort). One of the common quality control is to filter out this group of workers (Kazai et al., 2013). In this study, we applied a filtering method of HITs consisted of two assurance criteria and a number of qualification setting for each experiment. As a result, workers whose tasks could not fulfill the quality criteria were removed from the experiment. The filtering included two criteria; a trap question, to test if workers have read question carefully, and spent time for completing a HIT. The trap question used in the three experiments is shown in Figure 3.6.

Task completion time as a filtering method has been extensively used in various studies especially in crowdsourcing (T. Xia, Zhang, Xie, & Li, 2012; Kazai et al., 2013). The completion time is an indicator to find out malicious random answers (Difallah et al., 2012). This method was used as the second filtering step used in this study, by which those tasks that were completed in less than 2 minutes for each task identified as spammers. The threshold was determined on the basis of a clear observed separation of poor and reasonable levels of performance.

To be sure that you are paying attention, please select 'Neither Agree nor Disagree' for this item.

1) Strongly Disagree
2) Disagree
3) Neither Agree nor Disagree
4) Agree
5) Strongly agree

Figure 3.6: Trap question used in this study

3.5 Reliability of Relevance Judgments

The use of crowdsourcing in creating relevance judgment for IR evaluation can be validated through measuring the agreement between crowdsourcing workers and human assessors. This is to evaluate the reliability of crowdsourcing as a replacement for human assessors. In crowdsourcing experiments, the inter-rater or inter-annotator agreement is used to measure the performance and to analyze the agreement between crowdsourcing workers and human assessors. The score of homogeneity in the rating list given by judges is the inter-rater agreement. Two common methods are defined to calculate the inter-rater agreement. The first is the percentage agreement, which is the simplest and easiest scale-base [i.e. dividing number of times for each rating (e.g. 1, 2, ... 5), assigned by each assessor, by the total number of the ratings]. Cohen's kappa, the second method, is an adjusted accuracy based on the probability of chance of agreement (Alonso & Mizzaro, 2012). Cohen's kappa in comparison with percentage agreement is more robust because the effects of random agreement between two assessors is considered in Cohen's kappa (Cohen, 1960). A five level scale for qualitatively interpreting Cohen's kappa was proposed by Landis and Koch (1977);

- 0.01–0.20: Slight
- 0.21–0.40: Fair
- 0.41–0.60: Moderate
- 0.61–0.80: Substantial
- 0.81–0.99: Perfect

3.5.1 Individual Agreement

Ternary agreement and binary agreement are two measuring methods to find out individual agreement between two assessors (Kazai, Kamps, Koolen, et al., 2011). Ternary agreement is an exact degree of relevance on which crowdsourced judgments

and judgments by TREC assessors have agreement. In other words, if a worker and the TREC assessor had judged the same <topic, document> similarly, they are considered as "have agreement". For instance, five workers made relevance judgments on a given topic and two documents as shown in Table 3.3.

Table 3.3: Example for calculating individual agreement

			Judgments	
Topic	Document	Worker ID	Worker	TREC Assessors
100	1	Worker 1	R	R
100	1	Worker 2	HR	
100	1	Worker 3	NR	
100	1	Worker 4	HR	
100	1	Worker 5	R	
100	2	Worker 1	NR	R
100	2	Worker 2	NR	
100	2	Worker 3	R	
100	2	Worker 4	R	
100	2	Worker 5	NR	

Note: HR= Highly-relevant, R= Relevant and NR= Non-relevant.

The topic and document, <100, 1> were judged by five workers. Two workers (worker 2 and worker 4) judged this topic and document as highly-relevant while worker1 and worker 5 judged as relevant. Only worker 3 judged the topic and document as non-relevant. Five workers also judged the topic and document <100, 2>. Two workers judged as relevant while other three workers judged as non-relevant. The last column in Table 3.3 shows the relevance judgments made by TREC assessors. Table 3.4 shows the ternary individual agreement between relevance judgments made by workers and relevance judgments provided by TREC assessor.

Table 3.4: Ternary agreement (workers and TREC assessors)

TREC assessors				
Workers		HR	R	NR
	HR	0%	20%	0%
	R	0%	40%	0%
	NR	0%	40%	0%

Note: HR= Highly-relevant, R= Relevant and NR= Non-relevant.

In four cases (40%), both the worker and the respective TREC assessor judged the pairs of <topic, document> as relevant. In two cases (20%), the workers judged highly-relevant while the corresponding TREC assessors judged the pairs of <topic, document> as relevant. In four cases (40%), workers judged the pairs of <topic, document> as non-relevant while the corresponding TREC assessors judged as relevant. Therefore, the total ternary individual agreement was 40%. Table 3.5 shows the binary individual agreement between relevance judgments made by workers and relevance judgments provided by TREC assessor. In comparison with ternary agreement to which the agreement between two assessors should be exact, binary agreement is not exact and it considers highly-relevant and relevant as an agreement.

In the example for calculating individual agreement (Table 3.3), in six cases both worker and the corresponding TREC assessors judged either relevant or highly-relevant. In four cases, the workers judged the pairs of <topic, document> as non-relevant while the corresponding TREC assessor judged as relevant. Therefore, the binary agreement between relevance judgments made by workers and relevance judgments provided by TREC assessors would be 60% (Table 3.5).

Table 3.5: Binary agreement (workers and TREC assessors)

		TREC assessors	
		R	NR
Workers	R	60%	0%
	NR	40%	0%

Note: R= Relevant and NR= Non-relevant.

3.5.2 Group Agreement

In this study, there are 50 different judgments created by 50 different workers, for each topic and each document, for total number of 1000 HITs for each experiment. In order to reduce a group of assessments to a single assessment, an aggregating method is required. As discussed in Section 2.2.2, there are a variety of aggregating methods in

crowdsourcing. MV method is one of the common and straightforward methods for aggregating labels. MV seems easy to implement and powerful to achieve meaningful results which makes the method popular for routine tasks (Tang & Lease, 2011). In this study, MV method was applied to aggregate the judgments. The rationale behind finding the group agreement between relevance judgments made by workers and relevance judgments provided by TREC assessors is twofold. The first is to find out whether the group agreement shows a higher value than individual agreement, the second, to assess whether the aggregate of multiple high cognitive abilities workers is better than that of low cognitive abilities ones. As a result, five relevance judgment sets were used in this study for each experiment.

Each set consisted of 100 relevance judgments for 10 topics and 10 documents. The relevance judgment sets are binary data; either relevant or non-relevant:

i. Relevance judgment set which is a subset of *qrels* created by TREC assessors consisting only relevance judgments for 10 documents for each of the 10 topics used in this study (100 relevance judgments).

ii. Relevance judgment set is created based on the judgments made by all of the workers.

iii. Relevance judgment set is created based on the judgments made by workers who have low cognitive abilities (Group 1).

iv. Relevance judgment set is created based on the judgments made by workers who have moderate cognitive abilities (Group 2).

v. Relevance judgment set is created based on the judgments made by workers who have high cognitive abilities (Group 3).

Table 3.6 provides an example for calculating group agreement. MV method of five workers for the pair of topic and document <100, 1> is relevant since two workers judged as highly-relevant which considered as relevant and another two workers judged as relevant and in total four workers judged as relevant while one workers judged as non-relevant. Therefore, MV method for the pair of topic and document <100, 1> is relevant. Group agreement between workers and the TREC assessors using MV method is shown in Table 3.7. In the example provided in Table 3.6, since there are only two pairs of topic and document, MV for the pair of topic and document <100, 1> is relevant which is as the same as the relevance judgment made by TREC assessors. MV for <100, 2> is non-relevant which is not as the same as the relevance judgment provided by TREC assessors. In one case, both workers and the TREC assessors agree while in another case, they disagree. Therefore, group agreement between relevance judgments (by workers) and relevance judgments (by TREC assessors) is 50%.

Table 3.6: Example for calculating group agreement

Topic	Document	Worker ID	Worker Judgment	Binary Worker Judgment	MV method	TREC assessors Judgment
100	1	Worker 1	R	R	R	R
100	1	Worker 2	HR	R		
100	1	Worker 3	NR	NR		
100	1	Worker 4	HR	R		
100	1	Worker 5	R	R		
100	2	Worker 1	NR	NR	NR	R
100	2	Worker 2	NR	NR		
100	2	Worker 3	R	R		
100	2	Worker 4	R	R		
100	2	Worker 5	NR	NR		

Note: HR= Highly-relevant, R= Relevant and NR= Non-relevant.

Table 3.7: Group agreement (workers and TREC assessors)

	TREC assessors		
Workers		R	NR
	R	50%	0%
	NR	50%	0%

Note: R= Relevant and NR= Non-relevant.

3.5.3 Reliability of Relevance Judgment for Each Task

In addition to individual and group agreement which explained previously, five evaluation metrics were used to analyse the pattern and reliability of relevance assessments for each task named ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness. These metrics are explained as followed.

i. Accuracy

Accuracy of the relevance judgment is the proportion of judgments on which the worker and the TREC assessors agreed (Equation 2.6) (Kazai et al., 2013). Accuracy can be either 0 for "no agreement" or 1 "complete agreement", and is measured over the number of documents in a single HIT. Ternary accuracy (agreement on the exact degree of relevance) and binary accuracy are two types of measurements for accuracy.

In this study, the number of documents was five in a single HIT. Table 3.8 provides an example for accuracy, assuming there are a topic and five documents in a single HIT. A worker judges one document (Document 4) "accurate" as the same as TREC assessors. Therefore, the ternary accuracy for this example is 1/5 or 0.2. Accordingly, the binary accuracy for three documents accurately (Document 1, Document 4 and Document 5) is 3/5 or 0.6.

Table 3.8: Example for calculating accuracy

	Document 1	Document 2	Document 3	Document 4	Document 5
Worker	R	R	R	NR	HR
TREC assessor	HR	NR	NR	NR	R

Note: HR= Highly-relevant, R= Relevant and NR= Non-relevant.

ii. Precision and recall

The precision is also calculated for each HIT. Precision is the number of relevant retrieved documents over the number of retrieved documents (Equation 2.4). Considering the example given in Table 3.8, the number of relevant documents retrieved is two and the number of retrieved documents is four. Therefore, the precision for this example is 2/4 or 0.5. The recall is the number of relevant documents retrieved over total number of relevant documents (Equation 2.5). In this example, the recall is 2/2 or 1. Accordingly, in this study we used the same procedures to calculate precision and recall.

iii. Normalized Discounted Cumulative Gain

In the example provided in Table 3.8, NDCG is calculated in three steps. Firstly, the NDCG for TREC assessors was calculated (Figure 3.7 (a)). Secondly, the NDCG was calculated for crowdsourced worker (Figure 3.7 (b)). Finally, the NDCG was calculated as shown in Figure 3.7 (c). NDCG calculation for each HIT is based on based on the Equation 2.10.

iv. Sensitivity, specificity and effectiveness

In order to measure performance of workers, we used sensitivity and specificity as previously was suggested by Raykar et al. (2010). Sensitivity is the same as recall and would be 1 for the example shown in Table 3.8. Specificity would be 1/3 or 0.33 (See Equation 2.7). The number of non-relevant documents judge by human experts is 3 and the number of non-relevant documents judge by workers is 1. According to the Equation 2.8, the effectiveness would be 1+0.33-1=0.33.

Rank(i)	R	Log(i)	R/log(i)
1	2	0	N/A
2	1	1	1
3	0	1.58	0
4	0	2	0
5	0	2.32	0
NDCG=2+(1+0+0+0)=3			

(a) Calculating TREC assessors NDCG

Rank(i)	Documents	R	Log(i)	R/log(i)
1	Document5	1	0	N/A
2	Document1	2	1	2
3	Document2	0	1.58	0
4	Document3	0	2	0
5	Document4	0	2.32	0
NDCG=1+(2+0+0+0)=3				

(b) Calculating worker NDCG

NDCG= NDCG(worker) / NDCG(TREC assessors)=3/3=1

(c) Calculating NDCG

Figure 3.7: Example of calculating NDCG

3.6 System Rankings

The aim of this section is to observe how is the effect of different relevance judgment sets created by crowdsourced workers on the ranks of systems in a benchmark scenario. We investigated whether when the assessments of workers with higher cognitive ability are used to evaluate and rank retrieval systems by effectiveness, they give a ranking more similar to that of the official TREC assessments than when the assessments of workers with lower cognitive abilities are so used. As explained earlier (see Section 3.5.2), there are five sets of relevance judgments in each experiment (a subset of *qrels*, relevance judgment set created by all of the workers, relevance judgment set created by Group 1, relevance judgment set created by Group 2 and relevance judgment set created by Group 3). Each relevance judgment set consists of 100 relevance judgments. The twenty-five IR systems that participated in the TREC 2009 Million Query Track were then scored five times (a subset of *qrels*, all workers, Group 1, Group 2 and Group 3)

using *MAP*, which is, calculated for each system by averaging the precision for both depth (*k*) 10 and 1000. Finally, the results of the ranked lists for each relevance judgment set were compared to the rankings achieved by the relevance judgments provided by the original TREC assessments using the Kendall's tau correlation coefficient.

3.7 Statistical Analysis

Correlation coefficient and significance test are of two main statistical methods applied throughout this study.

3.7.1 Correlation Coefficient

Correlation coefficient is to measure the strength of relationship between two variables. Pearson (for parametric variables) and Kendall are two commonly used correlation coefficient measures in IR evaluation experiments. Cohen defined an acceptable cutoff to interpret data (Cohen, 1977), where small correlation for r =0.10 to 0.29, medium for r =0.30 to 0.49 and large for r =0.50 to 1.0. The correlation of 1.0 showed a perfect positive linear correlation (i.e. the factors form an upward-sloping straight line if plotted on a graph); a correlation of 0 means no correlation (as would occur if the factors were independent); and a correlation of -1 means perfect negative linear correlation (a downward-sloping straight line).

It is an standard procedure in information retrieval to measure the similarity between effectiveness scores of two lists of ranked systems (Scholer, Turpin, & Sanderson, 2011) with Kendall's tau (τ) (Kendall, 1938). For example, Kendall's tau identified how system rankings are similar for different *qrels* systems (Sakai & Kando, 2008). Moreover, Kendall's tau is commonly applied to measure the correlation between system rankings for different evaluation metrics (Sakai & Kando, 2008; Nowak & Rüger, 2010). In fact, Kendall's tau measures agreement in the ranking (instead of exact scores) between two sets of paired values. Using Kendall's tau in our study, we assess whether

one system is better than another system, and find out the level of agreement between the system rankings produced by crowdsourced judgments and TREC expert assessors' judgments.

3.7.2 Significance Test

As explained earlier in this chapter, workers were categorized into groups of low, moderate and high cognitive abilities in each experiment. For example, in the verbal comprehension experiment, workers were divided into three groups of low verbal comprehension skill, moderate verbal comprehension skill and high verbal comprehension skill. Significance tests are to explore differences between groups. One-way ANOVA test is a parametric significance test to examine differences among groups for reliability of judgments. Prior to exploring differences, Levene's test for homogeneity of variances was used to test homogeneity of variances. Not violating the assumption of homogeneity of variance, ANOVA test was used, otherwise, Welch's test was applied to report significant differences.

Significance tests tell us whether there is a significant difference among groups. Effect size is to measure the strength of association (Tabachnick & Fidell, 2001). One of the most commonly used effect size is *eta squared* which indicates the proportion of variance of the dependant variable that is explained by the independent variable. The value of effect size ranges from 0 to 1, including three cut-off points for small effect size (0.01), medium effect size (0.06) and large effect size (0.13) (Cohen, 2013).

Chi-square test for independence was applied to examine the effect of self-reported competence on crowdsourced judgment reliability. The Chi-square test for independence is used to explore relationship between categorical variables. This test compares observed proportion of cases for each category, and test the null hypothesis that the population proportions are identical (Soboroff, Nicholas, & Cahan, 2001). Moreover,

the effect of demographics on reliability of relevance judgments is examined by Chi-square test for independence.

3.8 Pilot Study

Prior to developing this study, a pilot study was conducted to assess whether cognitive abilities of workers is associated with reliability of crowdsourced relevance judgments. In the pilot study, the cognitive ability investigated is "verbal comprehension skill". Further, the relationship between the reliability of the crowdsourced relevance judgments on one hand, and self-reported difficulty of the task, confidence of the worker, and worker's knowledge of the topic on the other was investigated. The reliability of crowdsourced relevance judgment was evaluated for the agreement (if any) between workers' judgments and that of TREC expert assessors.

3.8.1 Experimental Data

Eight topics were taken from the TREC-9 Web Track[8] , and 20 documents were chosen randomly for each topic from the WT10g document collection[9]. All documents and topics were in English. Of the 20 chosen documents, 10 were relevant and 10 non-relevant, as judged by the original TREC assessors. For each of the 160 <topic, document> pairs, 10 binary judgments were obtained through crowdsourcing, each from a different worker, totally 1600 judgments made by workers. The number of workers who performed the tasks was 154.

3.8.2 Experimental Design

This experiment was conducted using Crowdflower and each HIT involved two steps (Appendix A); Step 1, which consisted of 40 tasks, to be completed by 10 workers. In this step, each task had four topics, and each topic had a document to be assess for the

[8] http://trec.nist.gov/data/t9.web.html
[9] http://ir.dcs.gla.ac.uk/test_collections/wt10g.html

relevance judgment against a given topic. Upon completing each judgment, the workers were required to complete a questionnaire. The questionnaire consisted of the following three items, to be answered on a 4-point scale;

Q1) Rate your knowledge on the topic: (Minimal 1 2 3 4 Extensive).

Q2) How difficult was this evaluation: (Easy 1 2 3 4 Difficult).

Q3) How confident were you in your evaluation: (Not confident 1 2 3 4 Very confident).

Step 2 was to examine the workers verbal comprehension. In this step, the workers were asked to complete a vocabulary test of 10 questions. These questions were randomly sampled from the FRCT. The workers were required to choose one of four given words that had the similar meaning as a given word. The verbal comprehension score was then calculated based on the overall vocabulary task.

3.8.3 Results and Discussion

Filtering untrustworthy workers was based on their verbal comprehension scores. As the vocabulary test is a multiple-choice test with 4 choices per question and 10 questions, a worker selecting at random has an expected score of 2.5. Put another way, a worker selecting at random has less than a one in four chance of achieving a score of 4 or higher. From 400 HITs, 81 HITs were recognized as unreliable. Therefore, of the 1600 judgments submitted, 1276 judgments was deemed as reliable, constituting 147 of the 154 workers. Further, workers were divided into two groups based on their verbal comprehension scores, the high values above the median and the low values below the median. A median split is one method for turning a continuous variable into a categorical one (Reis & Judd, 2000):

- Group 1- low scores: verbal comprehension score between 4 and 8, consisting of 156 HITs.

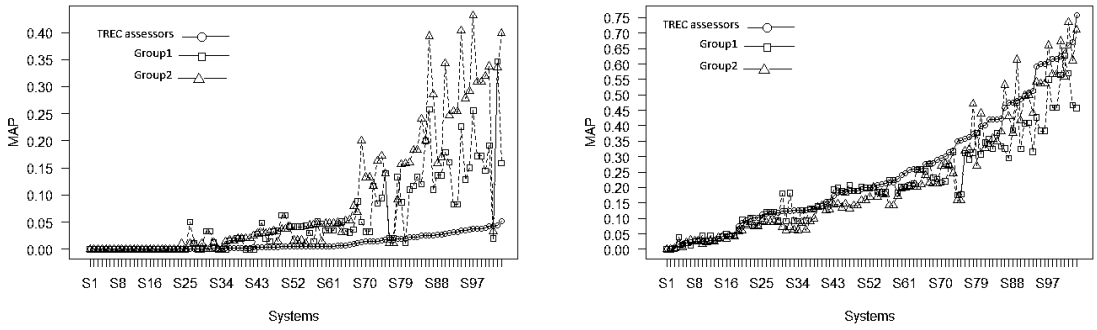
- Group 2- high scores: verbal comprehension score between 9 and 10, consisting of 163 HITs.

Agreement between Group 1 and TREC assessors (35.93% on relevant and 26.01% on non-relevant) is 61.94%. The level of disagreement between them is 34.2% while 3.7% of workers chose "Don't know". The level of agreement between Group 2 and TREC assessors is 75.9% (32.7% on relevant and 43.1% on non-relevant), which is higher than that of observed for Group 1. The disagreement between Group 2 and TREC assessors and the percentage of those workers who chose "Don't know" is 21.4% and 2.6%, respectively. Cohen's kappa agreement between the relevance judgments of crowdsourced workers and TREC assessors is 0.3 (fair agreement) for Group 1 and is 0.57 (moderate agreement) for Group 2. Pearson's correlation between verbal comprehension score and accuracy is 0.32 ($p<0.001$). The verbal comprehension score shows a moderate but significant correlation with accuracy. There are significant differences in accuracy between Group 1 ($M=0.62$, $SD=0.25$) and Group 2 ($M=0.76$, $SD=0.21$; $t(317) = -5.20$, $p<.001$) using the independent-samples t-test.

In addition, the influence of crowdsourced judgments on system ranking was examined, to see if crowdsourced judgments are reliable for evaluation purposes. One set of relevance judgments was generated from Group 1, and another from Group 2; where multiple assessors assessed the one document, MV method was used to determine its judgment. Each relevance judgment set consisted of 160 relevance judgments. The IR systems that participated in the TREC-9 Web Track were then scored using MAP and ranked using the Group 1 judgments, and again using the Group 2 judgments. MAP was calculated for both 1000 and to a lower depth 10. Each of these rankings was compared to the ranking achieved by the systems on the original TREC assessments, and Kendall's tau was computed for this rank comparison. The Kendall's tau correlation coefficients

between system rankings based on relevance judgments made by workers and system rankings based on relevance judgments provided by TREC assessors is shown in Table 3.9. The system rankings are shown in Figure 3.8. There is a slightly higher correlation between TREC rankings and those using the Group 2 judgments (for depth 10 and 1000) than those using the Group 1 judgments. This trend reveals that when the assessments of workers with high verbal comprehension were used to evaluate and rank retrieval systems by effectiveness, they gave a ranking more similar to that of the official TREC assessments than when the assessments of workers with low verbal comprehension were so used. However, the difference was not great.

After judging the relevance of each topic and document, workers rated their confidence in their evaluation using a 4-point Likert scale, from not confident to very confident. Table 3.10 shows the accuracy for each level of confidence, across the 1276 relevance judgments. It shows that workers who were less confident with their judgments achieved a lower accuracy in making relevance judgments, while the more confident workers achieved higher accuracy. A Chi-square test found the relationship between confidence and accuracy to be significant ($\chi 2 = 20.05$, $p < 0.01$). Once judging the relevancy of each topic and document, workers rated the difficulty of the evaluation using a 4-point Likert scale, from easy to difficult. From the 1276 relevance judgments, the accuracy is calculated for each level of difficulty to see whether judgment difficulty influences accuracy of judgments. Table 3.10 shows the accuracy for each level of difficulty. The workers who claim that the task is difficult achieve lower accuracy, while the workers who report the task to be easy achieve higher accuracy.



(a) (b)

Figure 3.8: System rankings. (a) System rankings based on TREC assessors, Group 1 and Group 2 for MAP (*k*=10) (b) System rankings based on TREC assessors, Group 1 and Group 2 for MAP (*k*=1000). The systems are sorted in ascending order of MAP scores generated using TREC assessors judgments.

Table 3.9: Kendall's tau (workers and TREC assessors)

Workers	Kendall's tau	
	MAP (*k*=10)	MAP (*k*=10)
Group 1	0.73	0.86
Group 2	0.85	0.90

Table 3.10: Relationship between self-reported competence and accuracy

	level	Judgments	Correct judgments	Accuracy
Confidence in judgment	1	44	21	0.47
	2	170	103	0.60
	3	530	368	0.69
	4	532	391	0.73
Difficulty of the judgment	1	342	276	0.80
	2	303	210	0.69
	3	541	345	0.63
	4	90	52	0.57
Knowledge on the topic	1	207	138	0.66
	2	319	241	0.75
	3	469	335	0.71
	4	281	169	0.60

Ratings are based on a 4-point Likert-type scale, confidence in judgment: Not confident 1 2 3 4 Very confident, difficulty of the judgment: Easy 1 2 3 4 Difficult, knowledge on the topic: Minimal 1 2 3 4 Extensive.

A Chi-square test found the relationship between difficulty and accuracy to be significant ($\chi 2 = 34.22$, $p < 0.01$). Workers also rated their knowledge of the given topic

using a 4-point Likert scale, from minimal to extensive. The accuracy statistics reported in Table 3.10 show a surprising result: workers with low and high self-reported knowledge are both less accurate than those with a moderate level of self-reported knowledge, with high-level knowledge workers being less reliable than low-level knowledge workers. The relationship was significant under a Chi-square test for independence ($\chi 2 = 18.56$, $p<0.01$). The findings of the pilot study showed that verbal comprehension skills influence the accuracy of crowdsourced workers who create the relevance judgments set. In light of the findings above, it is reasonable to argue that certain worker characteristics can be used to predict accuracy or to explain differences in accuracy between worker groups. However, as this pilot study was conducted on a small dataset, in the later experiments, the findings were confirmed on a large-scale experiment to investigate whether the findings remain stable. Additional research was conducted to investigate how other cognitive abilities namely general reasoning and logical reasoning can influence judgment reliability.

3.9 Summary

This chapter has discussed methodology used throughout this study. First experimental design was explained, together with explanation of experimental data. Followed by explanation of the task design and filtering method used in crowdsourcing. We also explained the metrics used to measure reliability of relevance judgments with giving examples. Usage of statistical analysis in this study, its underlying concepts, applicability and suitability was discussed. We end the chapter with details of the pilot study. In the next chapter, the results of verbal comprehension experiment is presented and discussed.

CHAPTER 4: VERBAL COMPREHENSION EXPERIMENT

This chapter presents the results of the verbal comprehension experiment, looking into the filtering method used in this experiment and details of the collected data. The main goal of this experiment is to evaluate the effect of verbal comprehension skill of the workers on reliability of crowdsourced relevance judgments. In this chapter, we will find out whether there is any relationship between workers' verbal comprehension skill and reliability of their relevance judgments. Moreover, the results for the effect of verbal comprehension skill of workers on system performance rankings in IR evaluation is also presented. The relationship between self-reported competence and reliability of relevance judgments, as well as, the effect of workers' demographics on reliability of relevance judgments is assessed. The screenshot of this experiment is provided in Appendix B.

4.1 Filtering Spam

Overall, 378 workers participated in the verbal comprehension experiment. After applying the trap filtering step, 969 HITs (out of 1000) proceeded to the second filtering step, the completion time, by which 106 HITs were recognized as unreliable HITs. In total 863 HITs with overall 4315 judgments were assigned "reliable HITs". The number of accepted HIT assignments and the total number of performed HITs, including the rejected ones, are shown in Table 4.1. "All HITs" is consisted of all of the collected judgments. "Cleaned HITs" (Reliable) comprises only approved HITs that passed the quality assurance criteria and were, thus, considered reliable. "Rejected HITs" comprises the HITs that were rejected based on trap question and time criteria.

Table 4.1: Summary of HITs

HITs type	Number of HITs	Judgments	Workers
All HITs	1000	5000	378
Cleaned HITs (Reliable)	863	4315	345
Rejected HITs	137	685	57

4.2 Descriptive Statistics

Demographic data of workers includes age, gender, education, country, computer and Internet experience is presented in Table 4.2. Female workers were the majority of workers in "all HITs" (57.9%) and "cleaned HITs" (61.2%). The relative excess of women is in line with a study (Paolacci et al., 2010) which reported more females than males across U.S.-based workers in their experiment using AMT and this trend may reveal women having greater access to computers or gender differences in motivation. The majority of workers in our study was from the USA which is in line with the previous studies which found that the majority of workers who participated in crowdsourcing platforms are from USA and India (Paolacci et al., 2010; Schulze, Seedorf, Geiger, Kaufmann, & Schader, 2011) . The range of workers' age for 38.9% of "all HITs", 39.3% of "cleaned HITs" and 36.5% of "rejected HITs" was between 30 to 40 years which is again consistent with a survey done in AMT and workers who participated to the survey were 36 years old on average (Paolacci et al., 2010). In agreement with Ipeirotis's findings (Panagiotis Ipeirotis, 2010), educational level was mostly bachelor degrees for workers in "all HITs" (44.5%) and in "cleaned HITs" (44.4%) groups. The majority of workers reported a high level of knowledge about computer and the Internet. Descriptive statistics for analyzed measures of each HIT is summarized in Table 4.3. The mean for verbal comprehension scores for "all HITs" is greater (11.88) than that of "rejected HITs" (5.92) and of "cleaned HITs" shows greatest value (12.83) compared with "All HITs" (11.88) and "rejected HITs" (5.92). The average of ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness for "rejected HITs" was less than that of parameters for "all HITs" and "cleaned HITs". The differences among "all HITs", "cleaned HITs" and "rejected HITs" is due to the filtering method applied in this study, which attributes for improving the quality of crowdsourced outputs. Table 4.4 presents descriptive statistics for self-reported competence of participants in doing their tasks. The

majority of workers in “all HITs” and “cleaned HITs” claimed a high level of confidence (level 4) in performing their judgments (47.0% and 48.7%, respectively). In “rejected HITs”, most of workers (39.1%) chose 3 for their level of confidence in their judgments. The level of difficulty reported for the relevance judgment was 1 representing that the task was not difficult for “all HITs” (51.8%), “cleaned HITs” (53%) and “rejected HITs” (44.5%). The majority of workers reported to be minimally (level 1) familiar with the topic for “all HITs” (45.7%), “cleaned HITs” (47.5%) and “rejected HITs” (34.3%).

Table 4.2: Demographics of participant

	level	All HITs (%)	Cleaned HITs (%)	Rejected HITs (%)
Age	Less than 20	2.8	0.8	15.3
	in my 20's	20.6	19.0	30.7
	in my 30's	38.9	39.3	36.5
	in my 40's	15.7	16.6	10.2
	in my 50's	16.3	17.7	7.3
	60+ years old	5.7	6.6	0
Gender	Male	42.1	38.8	62.8
	Female	57.9	61.2	37.2
Education	No education	0	0	0
	Basic schooling	2.4	1.5	8
	High school	41.2	43.8	24.8
	Bachelor degree	44.5	44.4	45.3
	Master degree	8	8	8
	PhD or higher	3.9	2.3	13.9
Computer Experience	1	0.3	.2	0.7
	2	1.3	1.4	0.7
	3	36.7	35.2	46
	4	61.7	63.2	52.6
Internet Experience	1	0.2	.2	0
	2	2.2	2.3	1.5
	3	33.3	30.7	49.6
	4	64.3	66.7	48.9
Country	AUS	2.2	2.1	2.9
	BHS	0.1	.1	0
	CAN	23.3	20.9	38.7
	GBR	23.7	24.7	17.5
	IRL	3.6	3.8	2.2
	NZL	1.3	1.5	0
	USA	45.8	46.9	38.7

Table 4.3: Descriptive statistics for analysed measures

	Measures	HITs	Minimum	Maximum	Mean	SD
All HITs	Verbal comprehension score	1000	-3.50	24	11.88	6.22
	Ternary accuracy	1000	0	1	0.42	0.23
	Binary accuracy	1000	0	1	0.69	0.21
	Precision	1000	0	1	0.74	0.22
	Recall	1000	0	1	0.86	0.23
	NDCG	1000	0.43	1	0.87	0.11
	Specificity	1000	0	1	0.40	0.43
	Effectiveness	1000	0	1	0.32	0.37
Cleaned HITs	Verbal comprehension score	863	-3.50	24	12.83	5.62
	Ternary accuracy	863	0	1	0.44	0.22
	Binary accuracy	863	0	1	0.72	0.22
	Precision	863	0	1	0.78	0.23
	Recall	863	0	1	0.86	0.24
	NDCG	863	0.43	1	0.88	0.11
	Specificity	863	0	1	0.43	0.43
	Effectiveness	863	0	1	0.34	0.38
Rejected HITs	Verbal comprehension score	137	-3.50	22.75	5.92	6.54
	Ternary accuracy	137	0	1	0.35	0.22
	Binary accuracy	137	0.2	1	0.64	0.20
	Precision	137	0.2	1	0.68	0.20
	Recall	137	0.25	1	0.88	0.20
	NDCG	137	0.43	1	0.86	0.12
	Specificity	137	0	1	0.24	0.37
	Effectiveness	137	0	1	0.19	0.34

The number of "cleaned HITs" accomplished by each worker for verbal comprehension experiment is shown in Figure 4.1. In the case of traditional expert-based evaluation for generating relevance judgment task, this distribution would be at as each expert would assess the same tasks. In our experiment, each worker may assess a different number of HITs. Some workers assessed a large number of HITs, with the most hard-working worker went through 14 HITs, while a long tail of workers worked on a single task only.

Table 4.4: Descriptive statistics for self-reported competence

	level	All HITs Percent	Cleaned HITs Percent	Rejected HITs Percent
Confidence in judgment	1	5.5	3.9	15.2
	2	13.7	14.3	9.6
	3	33.9	33.0	39.1
	4	47.0	48.7	36.1
Difficulty of the judgment	1	51.8	53	44.5
	2	25.8	26.8	20
	3	18.9	16.8	31.8
	4	3.5	3.4	3.6
Knowledge on the topic	1	45.7	47.5	34.3
	2	23.4	25.4	10.9
	3	22.9	22.2	27
	4	8	4.9	27.7

Ratings are based on a 4-point Likert-type scale, confidence in judgment: Not confident 1 2 3 4 Very confident, difficulty of the judgment: Easy 1 2 3 4 Difficult, knowledge on the topic: Minimal 1 2 3 4 Extensive.

Figure 4.1: Number of HITs judged by each worker

4.3 Effect of Verbal Comprehension Skill on Reliability of Judgments

As previously discussed in section 3.5, workers were divided into three groups according to their verbal comprehension scores:

- Group 1- low scores: with the verbal comprehension score less than or equal to 10.25, consisting of 295 HITs.
- Group 2- moderate scores: with the verbal comprehension score between 10.25 and 15.5, consisting of 291 HITs.
- Group 3- high scores: with the verbal comprehension score more than 15.5, consisting of 277 HITs.

Accordingly, this Section provides the results for correlation between workers' judgment reliability and verbal comprehension score. In addition to the results for individual agreement and group agreement between workers and TREC judgments for relevance judgments, results for statistical differences among groups for reliability of relevance judgments is described in this Section.

4.3.1 Correlation Coefficient

Association between verbal comprehension score and reliability of relevance judgments (ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness) is assessed by Pearson correlation matrix as summarized in Table 4.5. Binary accuracy (r=0.38), ternary accuracy (r=0.29) and effectiveness (r=0.34) are moderately correlated with verbal comprehension score (Cohen, 1977). Correlation between verbal comprehension score and either precision (r=0.24), recall (r=0.25), NDCG (r=0.23) or specificity (r=0.24) is low. Strong correlation between ternary accuracy and binary accuracy (r=0.53), and effectiveness (r=0.50) was seen. Ternary

accuracy showed a moderate correlation with other measures of judgment reliability including precision (r=0.32), recall (r=0.30), NDCG (r=0.43) and specificity (r=0.37).

Correlation of binary accuracy with precision (r=0.69), recall (r=0.59) and effectiveness (r=0.54) is strong whilst that of with NDCG (r=0.28) and specificity (r=0.28) is small. There is no correlation between precision and recall, which has a small correlation with NDCG (r=0.13) and effectiveness (r=0.17). Precision has a moderate correlation with NDCG (r=0.30), specificity (r=0.48) and effectiveness (r=0.45). Having a small negative correlation with recall (r=-0.25), specificity has a moderate correlation with NDCG (r=0.30) and strong correlation with effectiveness (r=0.87).

The positive correlation between each measure of judgment reliability and verbal comprehension score of workers shows that there is an association between judgment reliability and verbal comprehension skill of workers. Although, there is no previous research, which investigated the effect of verbal comprehension skill on reliability of crowdsourced relevance judgment, there are a few studies, which explored the effect of verbal comprehension skill on the search process.

Table 4.5: Pearson correlation matrix for eight measures

Measures	1	2	3	4	5	6	7	8
1.Verbal comprehension score	-							
2.Ternary accuracy	0.29**	-						
3. Binary accuracy	0.38**	0.53**	-					
4. Precision	0.24**	0.32**	0.69**	-				
5. Recall	0.25**	0.30**	0.59**	-0.03	-			
6. NDCG	0.23**	0.43**	0.28**	0.30**	0.13**	-		
7.Specificity	0.24**	0.37**	0.28**	0.48**	-0.25**	0.30**	-	
8.Effectiveness	0.34**	0.50**	0.54**	0.45**	0.17**	0.34**	0.87**	-

Note: p<0.01 (**)

In a previous study in librarianship, it was showed that a higher level of verbal comprehension ability is associated with a higher performance in searching (Allen & Allen, 1993). In a separate study (Allen, 1992), it was showed that high verbal abilities enable users to utilize a comprehensive search strategy as compared with those with low verbal abilities. The study found a significant Pearson correlation of 0.33 between verbal comprehension and "Number of Search Expressions" and 0.38 between verbal comprehension and "Number of High-Frequency Keywords". Verbal comprehension skill influences search process, and is the main predictor for choosing proper vocabularies for searching. Consistently, our results showed a moderate positive correlation between verbal comprehension and judgment reliability as users with high verbal abilities are better in understanding the context of articles than users with low verbal abilities

4.3.2 Individual Agreement (Workers vs. TREC Assessors)

Table 4.6 presents ternary agreement for relevance judgments between all of the workers and the TREC assessors. Overall, in 44.6% of cases the workers' judgments and the corresponding TREC assessors' judgments have the ternary degree of relevance, which include "highly-relevant documents" (15.21%), "relevant documents" (14.33%) and "non-relevant documents" (15.06%). Most disagreement are between those workers whose judgements were "highly-relevant documents" whilst the TREC considered them "relevant documents". Binary agreement for relevance judgments between all of the workers and the TREC assessors is also shown in Table 4.6. The binary agreement is 57.06% for "relevant documents" and 15.06% for "non-relevant documents", for overall binary agreement of 72.12%.

Table 4.6: Agreement (Ternary and Binary) for all workers

		TREC assessors				
		Ternary agreement			Binary agreement	
		HR	R	NR	R	NR
All workers	HR	15.21%	16.85%	5.33%	-	-
	R	10.68%	14.33%	12.4%	57.06%	17.73%
	NR	4.63%	5.51%	15.06%	10.15%	15.06%

HR, highly-relevant; R, relevant; NR, non-relevant

Ternary agreement and binary agreement for relevance judgments between each group of workers (consisted of Group 1 for low scores, Group 2 for moderate scores and Group 3 for high scores) and the TREC assessors is presented in Table 4.7. Ternary agreement with the corresponding TREC assessor on relevant documents is 13.83% for Group 1, 14.02% for Group 2 and 15.17% for group 3. Accordingly, the level of agreement on non-relevant documents is 10.11%, 14.70% and 20.72% for group 1 to 3, respectively. Overall ternary agreement between Group 1 and the TREC assessors is 36.62% (12.68% for "highly-relevant documents", 13.83% for "relevant documents" and 10.11% for "non-relevant documents"). The overall ternary agreement for Group 2 is 44.53% (15.81% for "highly-relevant documents", 14.02% for "relevant documents" and 14.70% for "non-relevant documents"), which is 53.15% for group 3 (17.26% on for "highly-relevant documents", 15.17% "relevant documents" and 20.72% for "non-relevant documents"). Binary agreement for relevance judgments between different groups of workers and the TREC assessors is also presented in Table 4.7. Binary agreement for Group 1 is 51.12% for "relevant documents" and 10.11% for "non-relevant documents". The overall binary agreement for Group 1, Group 2 and Group 3 is 61.23%, 73.81% and 81.95%, respectively.

Table 4.7: Agreement (Ternary and binary) for groups of workers

Workers		TREC assessors				
		Ternary agreement			Binary agreement	
		HR	R	NR	R	NR
Group 1 (low scores)	HR	12.68%	14.24%	6.10%	-	-
	R	10.37%	13.83%	16.13%	51.12%	22.23%
	NR	7.66%	8.88%	10.11%	16.54%	10.11%
Group 2 (moderate scores)	HR	15.81%	18.42%	6.05%	-	-
	R	10.86%	14.02%	12.51%	59.11%	18.56%
	NR	3.78%	3.85%	14.70%	7.63%	14.70%
Group 3 (high scores)	HR	17.26%	17.98%	3.75%	-	-
	R	10.83%	15.17%	8.30%	61.23%	12.06%
	NR	2.31%	3.68%	20.72%	5.99%	20.72%

HR, highly-relevant; R, relevant; NR, non-relevant

Similar to (Alonso & Mizzaro, 2012), our assessments for individual agreement showed that, there is a higher binary agreement on "relevant documents" for relevance judgments between workers and the TREC assessors as compared with "non-relevant documents". The disagreement is mostly appeared for workers who claimed "relevant documents" whilst the TREC assessors assigned them not relevant, which was consistent with previous study (Alonso & Mizzaro, 2012). Moreover, in a distinct study (Blanco *et al.*, 2013), it was highlighted that, in contrast to experts who are pessimistic in their judgment, non-expert assessors consider more items as relevant. A summary of the individual agreement for relevance judgments between workers and the TREC assessors is presented in Table 4.8 providing Cohen's Kappa agreement for ternary agreement and binary agreement.

Table 4.8: Summary of individual agreements

Workers	Ternary agreement		Binary agreement	
	Percentage	Kappa	Percentage	Kappa
All workers	44.6%	0.17	72.12%	0.33
Group 1(low scores)	36.62%	0.04	61.23%	0.07
Group 2 (moderate scores)	44.53%	0.17	73.81%	0.36
Group 3 (high scores)	53.15%	0.30	81.95%	0.57

Comparison between overall ternary agreement and overall binary agreement, showed that the binary agreement is higher. Binary agreement for relevance judgments between Group 2 and the TREC assessors is 73.81%, which is higher than that of ternary agreement for 29.28%. This is due to the easier and simpler assumptions for binary agreement. For instance, in ternary agreement workers are assessed against certain and exact levels of relevancy. Kappa agreement (ternary) for relevance judgments between all of the workers and the TREC assessors is 0.17, a slight agreement according to (Landis & Koch, 1977), whilst Kappa agreement for binary agreement is 0.33 (a fair agreement). Group 1 has a Kappa ternary agreement and Kappa binary agreement of 0.04 and 0.07 respectively, showing a slight agreement with the TREC assessors. A slight agreement is seen between Group 2 and the TREC assessors for a Kappa value of 0.17 for ternary agreement, but a fair agreement for Kappa binary of 0.36. Group 3 has a moderate binary agreement (0.57) and a fair ternary agreement (0.30) with the TREC assessors. The comparison among workers groups for binary agreement and ternary agreement with the TREC assessors showed that Group 3 has the highest percentage agreement and Kappa statistics. As a result, Group 3 appears more reliable in creating relevance judgments as compared with Group 1 (low scores) and Group 2 (moderate scores). In the other words, high scores workers are more reliable in comparison with moderate scores and low scores workers, respectively.

Our findings are consistent with a previous study in which relevance judgments between workers and TREC assessors was 68% for individual binary agreement (a fair agreement) and 59% for ternary agreement (Alonso & Mizzaro, 2012). In another study, the relevance judgments comparison between TREC and non-TREC assessors, showed 75% overall agreement (Al-Maskari, Sanderson, & Clough, 2008), similar to the binary agreement that we found between workers and the TREC assessors (72.12%) in our study.

In this study, none of the groups, however, has a strong individual agreement with TREC assessors. This can be due to subjective matters in performing relevance judgments and could be varied among assessors, as shown previously for an overall agreement of 70-80% between two TREC assessors in a relevance judgment (Voorhees & Harman, 2005). This is important to highlight that individual agreement is to scale the level of agreement between every worker and the corresponding TREC assessor. As explained in Chapter 2 (section 2.2.2), in crowdsourcing experiment, in order to make sure that the output is reliable, a group of workers assesses the relevancy of each topic and document and then to have a single relevance judgment, aggregating method has been applied. Therefore, in order to validate individual agreement, we assessed a group agreement between workers and the TREC assessors to find out whether group agreement is higher than individual agreement.

4.3.3 Group Agreement (Workers vs. TREC Assessors)

Agreements for relevance judgments between the TREC assessors and the rest of relevance judgment sets (including all workers, Group 1, Group 2 and Group 3) is presented in Table 4.9. The level of agreement between workers and the TREC assessors on "relevant documents" is 62% for all of the workers. Group comparison showed that 56% of workers in Group 1, 61% in Group 2 and 64% in Group 3 have an agreement with TREC assessors on "relevant documents". The agreement between workers and the TREC assessors on "non-relevant documents" is 15%, 10%, 14% and 22% for all workers and Group 1-3, respectively. There is most disagreement between workers and TREC assessors on documents, which workers marked relevant and TREC assessors marked not relevant (18% for all of the workers, 23% for Group 1, 19% for Group 2, 11% for Group 3). Table 4.10 summarizes percentage agreement and Cohen's Kappa for relevance judgments between workers and the TREC assessors.

Table 4.9: Group agreement (workers and TREC assessors)

		TREC assessors	
		R	NR
All workers	R	62%	18%
	NR	5%	15%
Group 1 (low scores)	R	56%	23%
	NR	11%	10%
Group 2 (moderate scores)	R	61%	19%
	NR	6%	14%
Group 3 (high scores)	R	64%	11%
	NR	3%	22%

R, relevant documents; NR, non-relevant documents

Table 4.10: Summary of group agreement

Workers	Group agreement	Kappa
All workers	77%	0.42
Group 1 (low scores)	66%	0.15
Group 2 (moderate scores)	75%	0.37
Group 3 (high scores)	86%	0.66

The highest level of group agreement for relevance judgments is between Group 3 and the TREC assessors for 86%, with Kappa statistic of 0.66. Group agreement for Group 2 is 75%, higher than that of Group 1 for 66%. The Kappa value shows a fair agreement (Kappa=0.37) for Group 2 and a slight agreement (Kappa=0.15) for Group 1.

In comparison with the individual agreement, group agreement between workers and TREC assessors for relevance judgments shows higher values. The higher agreement with TREC assessors reveals that group agreement is more reliable. Furthermore, the use of MV generates one judgment set out of several judgments seems to filter out noisy judgments. Kappa agreement for relevance judgments between workers and the TREC assessors (0.42) is in accord with a study which reported a moderate group agreement (Cohen's Kappa= 0.478) for relevance judgments between workers and TREC assessors (Alonso & Mizzaro, 2012).

In our study, group agreement for relevance judgments between Group 3 and the TREC assessors is substantial (Kappa 0.66) and higher than Group 2 and Group 1. As a result, high scores workers are more reliable in their relevance judgments than those workers with moderate or low scores. Group agreement between Group 2 and the TREC assessors for Kappa value of 0.37 is greater than Group 1 (Kappa=0.15), indicating a positive association between verbal comprehension scores and judgment reliability.

4.3.4 Difference of Reliability of Judgments among Groups

Table 4.11 and Figure 4.2 present means for ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness for each group. Group 3 has the highest values for all of the measures of judgment reliability. For all measures of judgment reliability, the mean values for Group 3 (high scores) is higher than Group 2 (moderate scores) and Group 1 (low scores) as presented in figure 4.2. Mean of specificity for Group 3 is 0.57 while that of for Group 2 and Group 1 is 0.41 and 0.31, respectively. One-way statistical significance test was conducted to find statistically significant differences among these three groups (if any) for different measures including ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness. A test of homogeneity of variances was also conducted for each measurement, showing the assumption of homogeneity of variances had been violated. Therefore, Welch's test was applied instead of ANOVA test for all measures of judgment reliability as shown in Table 4.12.

Table 4.11: Mean of the judgment reliability measures

Group	Ternary accuracy	Binary accuracy	Precision	Recall	NDCG	Specificity	Effectiveness
1	0.35	0.61	0.71	0.77	0.85	0.31	0.18
2	0.44	0.73	0.78	0.86	0.88	0.41	0.33
3	0.53	0.81	0.85	0.92	0.91	0.57	0.51

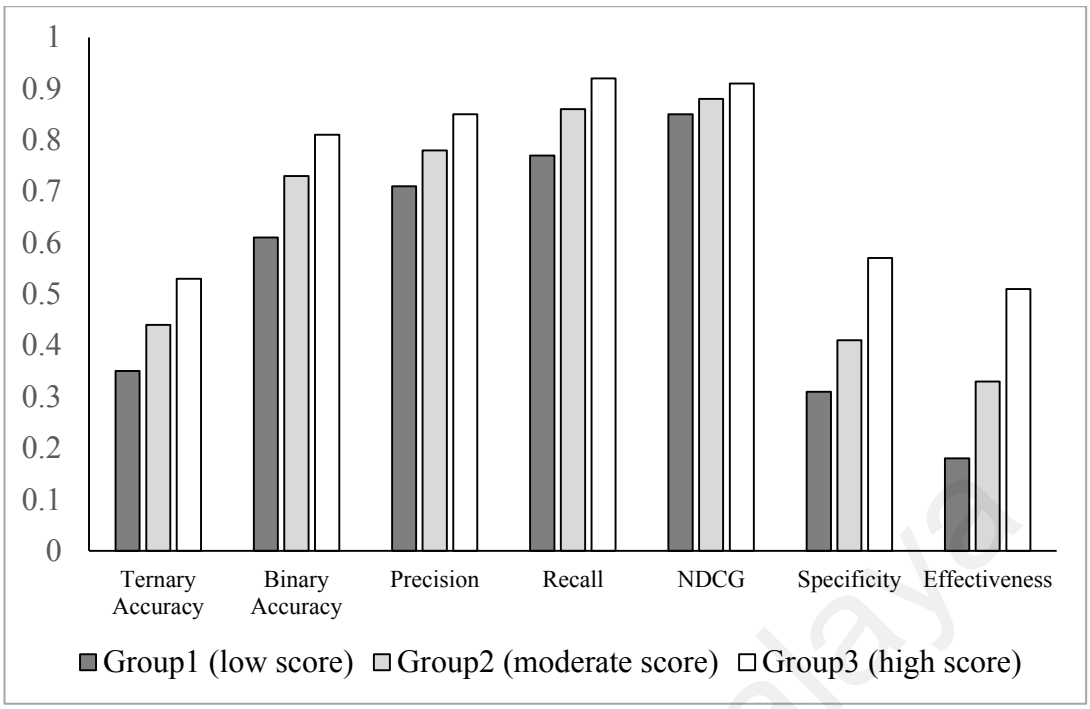


Figure 4.2: Mean of judgment reliability measures

There is a statistically significant difference between three groups for all measures (ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness) for *p* value less than 0.001. Post hoc comparisons using the Games-Howell test (Appendix E) reveal that workers with high verbal comprehension skill (Group 3) have a significantly greater binary accuracy, ternary accuracy, precision, NDCG, specificity and effectiveness in their relevance judgments as compared with Group 1 and Group 2. Workers with a moderate verbal comprehension skill showed a significantly higher means for binary accuracy, ternary accuracy, precision, recall, NDCG, specificity and effectiveness in their relevance judgments than those workers with a low verbal comprehension skill.

Our study is in line with the previous study (Scholer et al., 2013) which found a significant difference between participants with high NfC and low NfC in their relevance judgments using ANOVA test. The participants with high NfC had a higher agreement with the expert assessors (gold-set) than participants with low NfC. Effect size eta^2 values

were calculated for all measures of judgment reliability. Effect size value for binary accuracy (eta^2= 0.15) suggested a large significance effect size. Effect size value for effectiveness, ternary accuracy, precision, recall and specificity is 0.12, 0.09, 0.06, 0.08 and 0.06, respectively, which shows a moderate effect size, but a small for NDCG (0.05) among groups.

Table 4.12: Welch's test

Measures	Group	df	F	p
Ternary accuracy	Between Groups	2	40.130	0.000
	Within Groups	569.131		
Binary accuracy	Between Groups	2	72.451	0.000
	Within Groups	572.000		
Precision	Between Groups	2	27.712	0.000
	Within Groups	570.892		
Recall	Between Groups	2	29.825	0.000
	Within Groups	561.721		
NDCG	Between Groups	2	23.200	0.000
	Within Groups	568.455		
Specificity	Between Groups	2	28.340	0.000
	Within Groups	570.338		
Effectiveness	Between Groups	2	61.734	0.000
	Within Groups	553.642		

4.4 Effect of Verbal Comprehension Skill on Rank Correlation

As explained earlier (see Section 3.5.2), there are five sets of relevance judgments: (i) a relevance judgment set provided by TREC assessors (a subset of *qrels*), (ii) a relevance judgment set created by all of the workers, (iii) a relevance judgment set created by Group 1 (low verbal comprehension skill), (iv) a relevance judgment set created by Group 2 (moderate verbal comprehension skill), and (v) a relevance judgment set created by Group 3 (high verbal comprehension skill). System rankings based on relevance judgment set generated by all of the workers and relevance judgment set provided by TREC assessors using MAP (k=1000) is shown in Figure 4.3 and using MAP (k=10) in Figure 4.4.

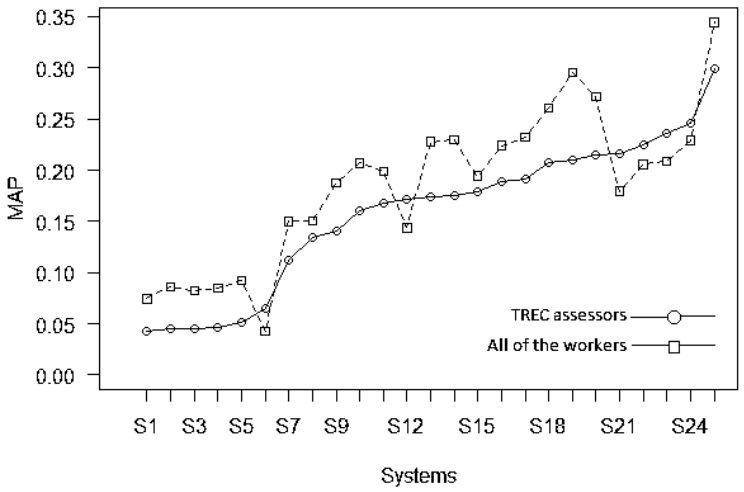


Figure 4.3: System rankings for all workers; MAP (*k*=1000). The systems are sorted in ascending order of MAP (*k*=1000) scores generated by TREC assessors judgments.

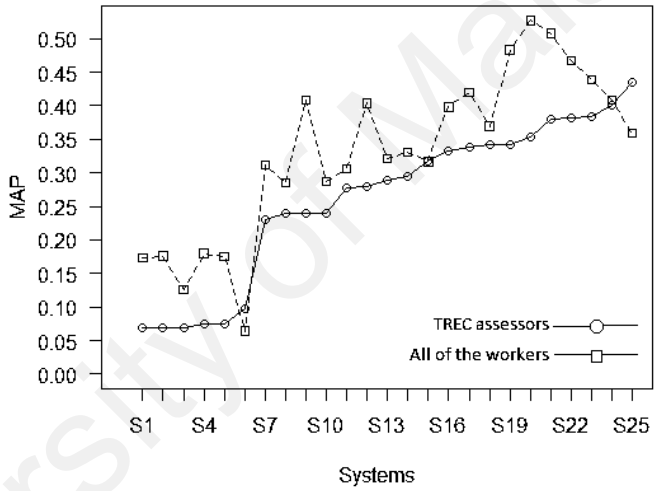


Figure 4.4: System rankings for all workers; MAP (*k*=10). The systems are sorted in ascending order of MAP (*k*=10) scores generated by TREC assessors judgments.

Using relevance judgment sets, 25 systems scored and ranked by Group 1 as the gold data. Then, the systems scored again based on relevance judgment set generated by Group 2 (as the gold data). Finally, the systems scored by Group 3. The System rankings for relevance judgment sets using MAP (*k*=1000 and *k*=10) is shown in Figure 4.5 and Figure 4.6. System rankings for relevance judgments made by Group 3 is relatively closer to that of by the TREC assessors as compared with Group 2 and Group 1. System rankings for Group 2 seems quite closer the system rankings for the TREC assessors as it is for

Group 1. Table 4.13 shows Kendall's tau correlation that was computed for the rank comparison between different sets of relevance judgments of workers and TREC assessors.

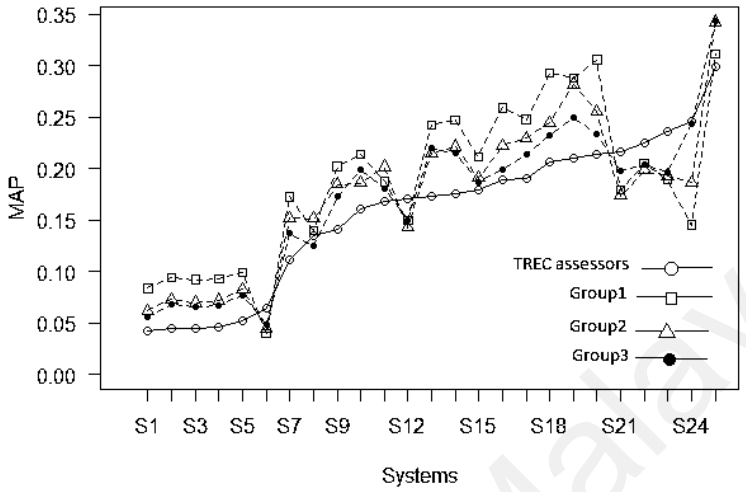


Figure 4.5: System rankings for groups; MAP (*k*=1000). The systems are sorted in ascending order of MAP (*k*=1000) scores generated by TREC assessors judgments.

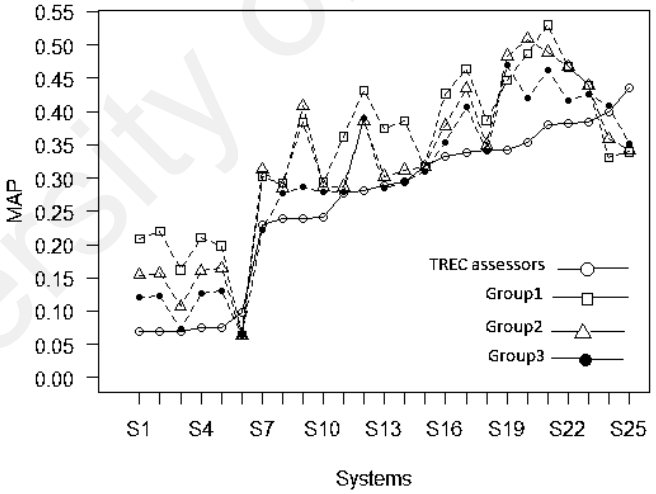


Figure 4.6: System rankings for groups; MAP (*k*=10). The systems are sorted in ascending order of MAP (*k*=10) scores generated by TREC assessors judgments.

Table 4.13: Kendall's tau correlation

Workers	Kendall's tau	
	MAP (*k*=1000)	MAP (*k*=10)
All workers	0.66	0.65
Group 1 (low scores)	0.56	0.60
Group 2 (moderate scores)	0.63	0.64
Group 3 (high scores)	0.69	0.75

Based on (Landis & Koch, 1977), there are high correlations between system rankings created by TREC assessors on one hand and the system rankings made by all of the workers, Group 1, Group 2 and Group 3 on the other hand for both MAP (10) and MAP (1000). The Kendall's tau correlation coefficients between Group 3 and the TREC assessors is the highest where MAP (1000) and MAP (10) are 0.69 and 0.75, respectively (Figure 4.5 and Figure 4.6). The Kendall's tau correlation coefficients between Group 2 and the TREC assessors is 0.63 for MAP (1000) and 0.64 for MAP(k=10), which is slightly higher than that of for Group 1 (0.56 for MAP (1000) and 0.60 for MAP(k=10)). The Kendall's tau correlation coefficients between Group 1 and the TREC assessors shows the lowest correlation which is in line with a previously published study by Soboroff et al. (2001), showing a tau correlation of 0.459 with the TREC assessors similar to that of Group 1 (0.56). However, none of the correlations is very high as compared with a previously published study by Kazai, Kamps, Koolen, et al. (2011), showing a tau correlation of 0.96 with the INEX assessors using MAP with 21 topics. The reason that the correlation between system rankings created by workers and system rankings provided by TREC assessors is not very high in our experiment may be due to the number of topics and total number of relevance judgments; if more were produced, the correlation with the official ranking would likely be higher.

In our study, the results of system rankings show that verbal comprehension skill of workers have a little effect on system rankings and system rankings made by Group 3 which is the high scores group is relatively more reliable as it has a highest correlation with system rankings provided by the TREC assessors. Recent studies also shows that system rankings can be affected by different HIT design (Kazai, Kamps, Koolen, et al., 2011), by assessors' task and domain expertise (Bailey *et al.*, 2008) and by assessor errors (Carterette & Soboroff, 2010). In a study conducted to assess effects of HIT design on

the system rankings, a full set of quality control methods lead to the better system rankings with a high level correlation with the system rankings based on the gold-set (Kazai, Kamps, Koolen, et al., 2011). Accordingly, by removing low accuracy workers, there is a slight effect on system ranking. In a separate study, there were three different groups of judges including gold standard (query originators and expert in information seeking), silver standard (task expert) and bronze standard (ordinary assessors) (Bailey et al., 2008). The authors found that task and domain expertise of the assessors influence the system rankings. Silver standard was highly correlated with gold standard and bronze standard judges were not reliable substitute for the gold standard judges. This trend reveals that disagreements were not just from non-originators and unfamiliarity with topic and task plays a major role. Carterette and Soboroff (2010) studied the effects of different kinds of assessors ("conservative" vs. "liberal") on the process of making relevance judgments. The finding showed stable system rankings with "conservative" assessors whilst "liberal assessor" presents noise to the system rankings.

4.5 Effect of Self-Reported Competence on Accuracy of Judgments

In this section, we investigated how various self-reported competence about the workers including confidence in relevance judgments, difficulty of the relevance judgments and knowledge on the topic relate to the reliability of their relevance judgments. After judging the relevance of each topic and document, workers rated their confidence in their evaluation using a 4-point Likert scale (ranged from not confident to very confident). The reason for questioning about confidence was to explore whether a worker's reported confidence in judging relevance for a topic was justified: does self-reported confidence lead to more accurate judgment? Table 4.14 shows the accuracy (Equation 2.6) of each level of confidence, across 4315 relevance judgments. As expected, those workers with less confident on their judgments achieved lower ternary and binary accuracy.

A Chi-square test shows a relationship between confidence and ternary accuracy to be significant ($\chi 2 = 14.179$, $p < 0.01$). The relationship between confidence and binary accuracy is also significant for chi-squared test ($\chi 2 = 94.32$, $p < .001$) as well. Further, judging the relevancy of each topic and document, the workers were asked to rate the difficulty of the evaluation with a 4-point Likert scale (ranged from easy to difficult). For 4315 relevance judgments, the accuracy was calculated for each level of difficulty to see whether judgment difficulty influences the accuracy of judgments. Table 4.14 shows the ternary and binary accuracy for each level of difficulty. The workers who had claimed that the tasks were difficult achieved lower accuracy. Chi-squared test shows a relationship between the self-report of the difficulty of a task and binary accuracy ($\chi 2 = 61.30$, $p < .001$). The association between the difficulty of a task and ternary accuracy is not significant ($\chi 2 = 7.14$, $p = .06$).

For knowledge on the topic, the results shows, the lowest ternary and binary accuracy levels among those workers claimed to have more knowledge (level 4 on the scale of 0-4) about a given topic, where a chi-square tests measure the relationship between knowledge and binary accuracy ($\chi 2 = 4.30$, $p = 0.23$) and ternary accuracy ($\chi 2 = 7.55$, $p = .05$) not significant.

4.6 Effect of Demographics on Accuracy of Judgments

In this study, some demographics were acquired about the workers, which consists of age, gender, level of education, level of computer experience and level of Internet experience and their country as provided by Crowdflower. In this section, the demographics is assessed to find out how various demographics information about the workers is related to the reliability of their relevance judgments. Table 4.15 shows the ternary accuracy and binary accuracy for demographic information.

Accuracy is presented in Table 4.15, where chi-square test shows no association between age and accuracy. Besides, there is no meaningful differences in gender (male and female) for the level of accuracy as tested by chi-square. Chi-square tests for the level of education are not significant. Furthermore, our statistical analysis shows no meaningful association between accuracy and either the level of computer experience or the level of Internet experience. Similarly, geographical distribution of the workers is not correlated with the level of accuracy in relevance judgments.

Table 4.14: Self-reported competence and accuracy of judgments

	level	Judgments	Ternary correct judgments	Binary correct judgments	Ternary accuracy	Binary accuracy
Confidence in judgment	1	169	61	96	0.36	0.57
	2	619	244	383	0.39	0.62
	3	1425	650	991	0.46	0.69
	4	2102	969	1642	0.46	0.78
Difficulty of the judgment	1	2286	1051	1751	0.46	0.77
	2	1155	511	804	0.44	0.67
	3	726	308	473	0.42	0.65
	4	148	54	84	0.36	0.57
Knowledge on the topic	1	2050	942	1472	0.46	0.72
	2	1095	500	805	0.46	0.73
	3	960	397	695	0.41	0.72
	4	210	85	140	0.40	0.67

Ratings are based on a 4-point Likert-type scale, confidence in judgment: Not confident 1 2 3 4 Very confident, difficulty of the judgment: Easy 1 2 3 4 Difficult, knowledge on the topic: Minimal 1 2 3 4 Extensive.

Table 4.15: Demographics and accuracy of judgments

	level	Judgment	ternary correct judgment	Binary correct judgment	Ternary accuracy	Binary accuracy
Age	not yet 20	35	17	25	0.49	0.71
	in my 20's	820	338	545	0.41	0.66
	in my 30's	1695	779	1228	0.46	0.72
	in my 40's	715	309	529	0.43	0.74
	in my 50's	765	356	578	0.46	0.76
	60+ years old	285	125	207	0.44	0.73
Gender	Male	1675	743	1210	0.44	0.72
	Female	2640	1181	1902	0.45	0.72
Education	no education	0	0	0	0	0
	primary school	65	20	34	0.31	0.52
	high school	1890	812	1306	0.43	0.69
	Bachelor degree	1915	898	1440	0.47	0.75
	master degree	345	147	248	0.43	0.72
	PhD or higher	100	47	84	0.47	0.84
Computer experience	1	10	2	4	0.20	0.40
	2	60	28	42	0.47	0.70
	3	1520	679	1099	0.45	0.72
	4	2725	1215	1967	0.45	0.72
Internet experience	1	10	2	4	0.20	0.40
	2	100	58	83	0.58	0.83
	3	1325	564	939	0.43	0.71
	4	2880	1300	2086	0.45	0.72
Country	AUS	90	43	74	0.48	0.82
	BHS	5	2	5	0.40	1
	CAN	900	382	631	0.42	0.70
	GBR	1065	502	789	0.47	0.74
	IRL	165	83	129	0.50	0.78
	NZL	65	22	51	0.34	0.78
	USA	2025	890	1433	0.44	0.71

4.7 Summary

This chapter presented findings from the verbal comprehension experiment. The findings support that verbal comprehension skill of workers do influence the reliability of relevance judgments in crowdsourcing and those workers with higher levels of verbal comprehension skill appeared relatively more reliable in performing relevance judgments. Our results showed that relevance judgments provided by workers with high verbal comprehension skills are relatively more reliable for system rankings than that of made by workers with low level of verbal comprehension skills. In the next chapter, the results of general reasoning experiment are presented.

CHAPTER 5: GENERAL REASONING EXPERIMENT

This chapter presents the results of the general reasoning experiment, looking into the filtering method used in this experiment and details of the collected data. This experiment aims to investigate the effect of general reasoning skill on reliability of crowdsourced relevance judgments. The results of this experiment are shown in five main parts: (i) the effect of general reasoning skill of workers on reliability of relevance judgments, (ii) the effect of general reasoning skill of workers on system rankings, (iii) the effect of self-reported competence on reliability of relevance judgments, and (iv) the effect of workers' demographics on reliability of relevance judgments. The screenshot of this experiment is shown in Appendix C.

5.1 Filtering Spam

Totally, 502 workers participated in the published tasks of general reasoning experiment. From 1000 HITs, 46 HITs were removed because of failing to answer the trap question. From 954 HITs, 49 HITs were recognized as unreliable HITs in the second step of filtering (time spent of less than 2 minute on a task). Finally, 905 HITs or 4525 judgments were recognized as "Cleaned HITs" (Reliable) including 471 workers (Table 5.1). "All HITs" is consisted of all of the collected judgments. "Cleaned HITs" comprises only approved HITs that passed the quality assurance criteria and were, thus, considered reliable. "Rejected HITs" comprises the HITs that were rejected based on trap question and time criteria.

Table 5.1: Summary of HITs

HITs type	Number of HITs	Judgments	Workers
All HITs	1000	5000	502
Cleaned HITs (Reliable)	905	4525	471
Rejected HITs	95	475	52

5.2 Descriptive Statistics

Table 5.2 shows the demographic data for "all HITs", "cleaned HITs" and "rejected HITs". Looking at the distribution of workers, the majority of workers were male (57.8% in "all HITs", 54.9% in "cleaned HITs" and 85.3% in "rejected HITs") from GBR (36.4% in "all HITs" and 36.5% in "cleaned HITs"), aged 30-40 (28.6% in all and cleaned HITs). Educational level was mostly bachelor degrees (44.4% in "all HITs", 44.6% in "cleaned HITs" and 42.1% in "rejected HITs") with a high level of computer knowledge (59.5% in "all HITs", 59.6% in "cleaned HITs" and 58.9% in "rejected HITs") and Internet experience (59.3% in "all HITs", 59.6% in "cleaned HITs" in 56.8% for "rejected HITs"). Moreover, the majority of workers in "rejected HITs" were from USA (38.9%) and aged 20-30 (42.1%). Through the use of crowdsourcing, the workers represent a variety of characteristics providing an interesting and representatively diverse population.

Descriptive statistics for analyzed measures of each HIT is shown in Table 5.3. The measures of judgment reliability comprises ternary accuracy, binary accuracy, precision, recall, NDCG, specificity, effectiveness are between 0 and 1. The general reasoning score is calculated based on a test of 10 questions. Therefore, the maximum score would be 10 for workers who answered all 10 questions correctly. A trend which is obvious in Table 5.3 is that the average of ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness for "rejected HITs" was less than that of parameters for "all HITs" and "cleaned HITs". For example, the mean for effectiveness for "all HITs" is greater (0.32) than that of "rejected HITs" (0.17) and of "cleaned HITs" shows greatest value (0.34) compared with "All HITs" and "rejected HITs". The differences among all, cleaned and rejected HITs is due to the filtering method applied in this study, which attributes for improving the quality of crowdsourced outputs.

Table 5.2: Demographics of participant

	level	All HITs Percent	Cleaned HITs Percent	Rejected HITs Percent
Age	Less than 20	2.3	2.2	3.2
	in my 20's	26.7	25.1	42.1
	in my 30's	28.6	28.6	28.4
	in my 40's	24.2	26	7.4
	in my 50's	9.2	9.9	2.1
	60+ years old	9	8.2	16.8
Gender	Male	57.8	54.9	85.3
	Female	42.2	45.1	14.7
Education	no education	0.4	0.3	1.1
	basic schooling	2.6	2.8	1.1
	high school	39.2	40.1	30.5
	Bachelor degree	44.4	44.6	42.1
	master degree	9.7	9.7	9.5
	PhD or higher	3.7	2.4	15.8
Computer experience	1	0.3	0.2	1.1
	2	6.6	5.3	18.9
	3	33.6	34.9	21.1
	4	59.5	59.6	58.9
Internet experience	1	0.4	0.2	2.1
	2	6.2	4.1	26.3
	3	34.1	36.1	14.7
	4	59.3	59.6	56.8
Country	AUS	1.5	1.7	0
	BHS	0.3	0.3	0
	CAN	19.1	19.2	17.9
	GBR	36.4	36.5	35.8
	IRL	7.9	8	7.4
	NZL	0.7	0.8	0
	USA	34.1	33.6	38.9

Table 5.3: Descriptive statistics for analysed measures

	Measures	HITs	Minimum	Maximum	Mean	SD
All HITs	General reasoning score	1000	-2.5	10	6.23	3.03
	Ternary accuracy	1000	0	1	0.47	0.24
	Binary accuracy	1000	0	1	0.72	0.21
	Precision	1000	0	1	0.77	0.23
	Recall	1000	0	1	0.87	0.23
	NDCG	1000	0.43	1	0.89	0.10
	Specificity	1000	0	1	0.40	0.44
	Effectiveness	1000	0	1	0.32	0.39
Cleaned HITs	General reasoning score	905	-2.50	10	6.64	2.72
	Ternary accuracy	905	0	1	0.47	0.24
	Binary accuracy	905	0	1	0.72	0.21
	Precision	905	0	1	0.77	0.23
	Recall	905	0	1	0.86	0.23
	NDCG	905	0.43	1	0.89	0.10
	Specificity	905	0	1	0.42	0.44
	Effectiveness	905	0	1	0.34	0.39
Rejected HITs	General reasoning score	95	-2.50	10	2.36	3.17
	Ternary accuracy	95	0	1	0.42	0.23
	Binary accuracy	95	0.2	1	0.68	0.21
	Precision	95	0.2	1	0.73	0.20
	Recall	95	0.25	1	0.87	0.10
	NDCG	95	0.54	1	0.87	0.10
	Specificity	95	0	1	0.24	0.39
	Effectiveness	95	0	1	0.17	0.32

Table 5.4 shows the descriptive statistics for self-reported competence of participants in performing tasks. The majority of workers claimed a high level of confidence (level 4) in performing their judgments (45.8% in "all HITs", 45.6% in "cleaned HITs" and 47.2% in "rejected HITs"). The level of difficulty reported for the relevance judgment was 1 representing that the task was not difficult (47.9% in "all HITs", 49.6% in "cleaned HITs" and 31.4% in "rejected HITs"). The majority of workers were less familiar (level 1) with the topics (40.5% in "all HITs" and 43.2% in "cleaned HITs").

Figure 5.1 shows the number of HITs judged by each worker in general reasoning experiment (cleaned HITs). Each worker may assess a different number of the total set of HITs. Some workers assessed all of the HITs, with the most hard-working worker went through all 20 HITs, while a long tail of workers worked on a single task only.

Table 5.4: Descriptive statistics for self-reported competence

	level	All HITs Percent	Cleaned HITs Percent	Rejected HITs Percent
Confidence in judgment	1	3.6	3.6	3.6
	2	15.7	15.5	16.8
	3	34.9	35.2	32.4
	4	45.8	45.6	47.2
Difficulty of the judgment	1	47.9	49.6	31.4
	2	30.1	30.8	22.9
	3	16.7	16.3	21.1
	4	5.4	3.3	24.6
Knowledge on the topic	1	40.5	43.2	14.7
	2	26.2	26.3	25.3
	3	22.6	22.7	22.1
	4	10.7	7.8	37.9

Ratings are based on a 4-point Likert-type scale, confidence in judgment: Not confident 1 2 3 4 Very confident, difficulty of the judgment: Easy 1 2 3 4 Difficult, knowledge on the topic: Minimal 1 2 3 4 Extensive.

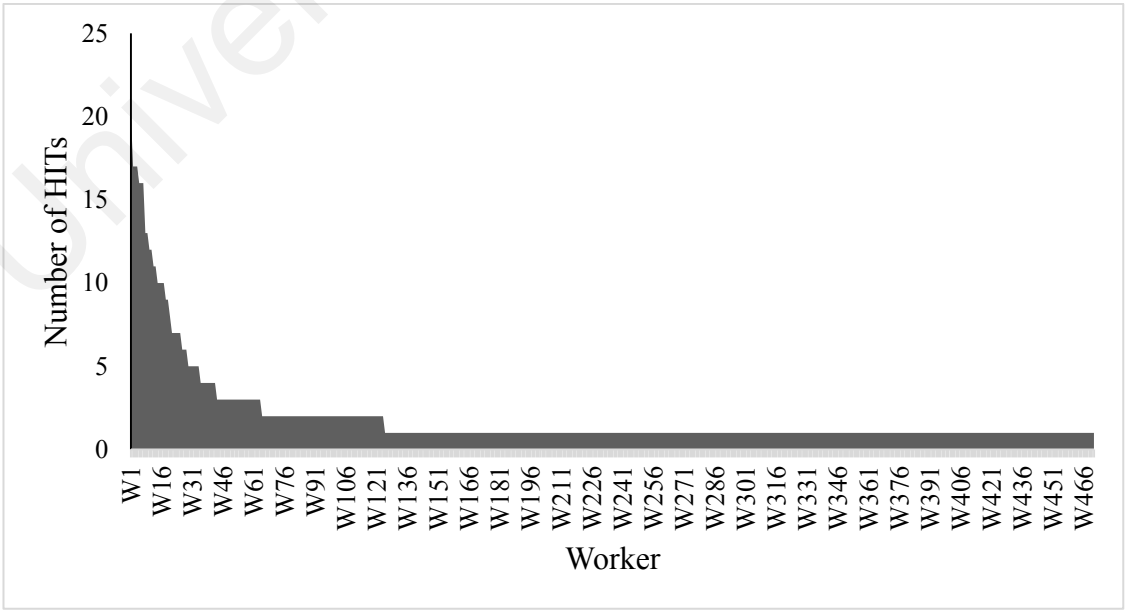


Figure 5.1: Number of HITs judged by each woker

5.3 Effect of General Reasoning Skill on Reliability of Judgments

Workers were divided into three groups according to their general reasoning scores:

- Group 1- low scores: with the general reasoning score less than or equal to five consisting of 265 HIT.
- Group 2- moderate scores: with the general reasoning score between five and eight consisting of 305 HITs.
- Group 3- high scores: with the general reasoning, score more than eight consisting of 335 HITs.

Accordingly, this Section provides the results for correlation between workers' judgment reliability and general reasoning score. In addition to the results for individual agreement and group agreement between workers and TREC judgments for relevance judgments, results for statistical differences among groups for reliability of relevance judgments is described in this Section. The rationale behind these investigations is to find out whether there is any relationship between general reasoning skill of workers and reliability of relevance judgments and whether the workers with higher level of general reasoning skill are more reliable than, workers with lower general reasoning skill in terms of making relevance judgments.

5.3.1 Correlation Coefficient

The correlation matrix using Pearson' correlation between different measures (ternary accuracy, binary accuracy, precision, NDCG, specificity, effectiveness and general reasoning score) is shown in Table 5.5. The general reasoning score is positively correlated with all measures of judgment reliability (ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness). Pearson's correlation between binary accuracy and general reasoning score shows a moderate correlation (r=0.31).

Ternary accuracy (r=0.21), precision (r=0.21), recall (r=0.19), NDCG (r=0.16), specificity (r=0.20) and effectiveness (r=0.24) have a small correlation with general reasoning score.

Ternary accuracy has a moderate correlation with precision (r=0.38), recall (r=0.37), NDCG (r=0.44) and specificity (r=0.41) whilst that of with binary accuracy (r=0.63) and effectiveness (r=0.54) is large. Binary accuracy is highly correlated with precision (r=0.73), recall (r=0.58) and effectiveness (r=0.57) while moderately correlated with NDCG (r=0.41) and specificity (r=0.35). Precision has a moderate correlation with NDCG (r=0.34), specificity (r=0.44) and effectiveness (r=0.47). Specificity has a small negative correlation with recall (r=-0.21) and small positive correlation with NDCG (r=0.29). The effectiveness have a small correlation with recall (r=0.16), moderate correlation with NDCG (r=0.38) and a large correlation with specificity (r=0.9). The main conclusion from this section is that there is an association between general reasoning skill and the reliability of relevance judgment.

Table 5.5: Pearson correlation matrix for eight measures

Measures	1	2	3	4	5	6	7	8
1. General reasoning	-							
2. Ternary accuracy	0.21**	-						
3. Binary accuracy	0.31**	0.63**	-					
4. Precision	0.21**	0.38**	0.73**	-				
5. Recall	0.19**	0.37**	0.58**	0.08*	-			
6. NDCG	0.16**	0.44**	0.41**	0.34**	0.26**	-		
7.Specificity	0.20**	0.41**	0.35**	0.44**	-0.21**	0.29**	-	
8.Effectiveness	0.24**	0.54**	0.57**	0.47**	0.16**	0.38**	0.9**	-

Note: p<0.01 (**)

5.3.2 Individual Agreement (Workers vs. TREC Assessors)

Table 5.6 shows ternary agreement for relevance judgments between all of the workers and TREC assessors. Overall, in 47.96% of the cases (14.74% for "highly-relevant documents", 17.75% for "relevant documents", 15.47% for "non-relevant

documents"), the judgments made by all of the crowdsourced workers were agreed with the corresponding judgments provided by TREC assessor on the exact degree of relevance. The binary agreement for relevance judgments between all of the workers and TREC assessors is also presented in Table 5.6. Binary agreement between relevance judgments made by all of the workers and the corresponding judgments provided by TREC assessor is 72.62% (57.15% for "relevant documents", 15.47% for "non-relevant documents"). Most disagreement are between those workers whose judgements were "relevant" whilst the TREC considered them "non-relevant" (17.83%).

Table 5.6: Agreement (Ternary and binary) for all workers

		TREC assessors				
		Ternary agreement			Binary agreement	
		HR	R	NR	R	NR
All workers	HR	14.74%	13.48%	3.84%	-	-
	R	11.18%	17.75%	13.99%	57.15%	17.83%
	NR	4.07%	5.48%	15.47%	9.55%	15.47%

HR, highly-relevant; R, relevant; NR, non-relevant

The ternary and binary agreement for relevance judgments between each group of workers (consisted of Group 1 for low scores, Group 2 for moderate scores and Group 3 for high scores) and TREC assessors is shown in Table 5.7. There is 39.77% (11.62% on "highly-relevant documents", 17.28% on "relevant documents" and 10.87% on "non-relevant documents") ternary agreement for relevance judgments between Group 1 and TREC assessors. The ternary agreement for relevance judgments between Group 2 and TREC assessors is 50.69% (16.72% on "highly-relevant documents", 19.02% on "relevant documents" and 14.95% on "non-relevant documents"). The ternary agreement for relevance judgments between Group 3 and TREC assessors is 51.95% (15.41% on "highly-relevant documents", 16.96% on "relevant documents" and 19.58% on "non - relevant documents"). The binary agreement for relevance judgments between each group of workers and TREC assessors is also presented in Table 5.7.

There is a 62.34% (51.47% on "relevant documents", 10.87% on "non-relevant documents") binary agreement between relevance judgments created by Group 1 and the judgments provided by TREC assessors. The binary agreement for relevance judgments between Group 2 and TREC assessors shows 75.08% agreement (60.13% on "relevant documents", 14.95% on "non-relevant documents"). The binary agreement between relevance judgments created by Group 3 and relevance judgments provided by TREC assessors is 78.51% (58.93% on "relevant documents" and 19.58% on "non-relevant documents"). The disagreement is mostly appeared for workers who claimed "relevant documents" whilst the TREC assessors assigned them not relevant.

Table 5.7: Agreement (Ternary and binary) for groups of workers

Workers		TREC assessors				
		Ternary agreement			Binary agreement	
		HR	R	NR	R	NR
Group 1 (low scores)	HR	11.62%	11.02%	4.15%	-	-
	R	11.55%	17.28%	18.64%	51.47%	22.79%
	NR	5.51%	9.36%	10.87%	14.87%	10.87%
Group 2 (moderate scores)	HR	16.72%	13.77%	4%	-	-
	R	10.63%	19.02%	13.05%	60.13%	17.05%
	NR	3.93%	3.93%	14.95%	7.87%	14.95%
Group 3 (high scores)	HR	15.41%	15.16%	3.46%	-	-
	R	11.41%	16.96%	11.16%	58.93%	14.63%
	NR	3.04%	3.82%	19.58%	6.86%	19.58%

HR, highly-relevant; R, relevant; NR, non-relevant

A summary of the individual agreement for relevance judgments between workers and the TREC assessors is presented in Table 5.8 providing Cohen's Kappa agreement for ternary agreement and binary agreement. Unsurprisingly, the amounts of ternary agreement (percentage and kappa) were lower than the amounts of binary agreement. As an instance, Group 2 has a binary agreement of 75.08%, which is higher than its ternary agreement by 24.39%. The reason is that a binary relevance judgment needs less effort to achieve than the graded relevance judgments as for the ternary agreement the workers have to agree on the exact level of relevancy.

Table 5.8: Summary of individual agreement

Workers	Ternary agreement		Binary agreement	
	Percentage	Kappa	Percentage	Kappa
All workers	47.96%	0.21	72.62%	0.34
Group 1(low scores)	39.77%	0.08	62.34%	0.10
Group 2 (moderate scores)	50.69%	0.25	75.08%	0.38
Group 3 (high scores)	51.95%	0.28	78.51%	0.49

Based on (Landis & Koch, 1977), the ternary agreement between relevance judgments made by all of the workers and relevance judgments provided by TREC assessors is relatively slight agreement with the Kappa value of 0.21, whilst the binary kappa agreement is a fair agreement with the Kappa value of 0.34. A slight agreement was also found between relevance judgments made by Group 1 and relevance judgments provided by TREC assessors with a Kappa value of 0.08 for the ternary agreement and 0.10 for binary agreement, respectively. A fair ternary and binary agreement for relevance judgments between Group 2 and TREC assessors was found with a Kappa value of 0.25 for ternary agreement and 0.38 for binary agreement. Group 3 has a moderate binary agreement (0.49) and a fair ternary agreement (0.28) with the TREC assessors. Based on individual agreement between relevance judgments made by different groups of workers and relevance judgments provided by TREC assessors, it can be concluded that Group 3 (high scores) is more reliable than Group 1 (low scores) and Group 2 (moderate scores) in terms of creating relevance judgments whilst Group 2 is also more reliable than Group 1. Considering (Carterette, Bennett, Chickering, & Dumais, 2008) as a baseline, they got a 43% individual agreement by using 5-points relevance scale and by grouping five categorize into two, that increase to 69% and 78%. The individual ternary agreement between all workers and the TREC assessors in our experiment is 47.96%, similar to 43%. The individual binary agreement between all workers and the TREC assessors in our experiment is 72.62% and that is in line with (Carterette et al., 2008).

5.3.3 Group Agreement (Workers vs. TREC Assessors)

Agreements for relevance judgments between the TREC assessors and the rest of relevance judgment sets including all workers, Group 1 (low general reasoning skill), Group 2 (moderate general reasoning skill) and Group 3 (high general reasoning skill) is presented in Table 5.9. Group comparison showed that 60% of workers in Group 1, 62% in Group 2 and 63% in Group 3 have an agreement with TREC assessors on "relevant documents". The agreement between workers and the TREC assessors on "non-relevant documents" is 12%, 9%, 14% and 18% for all workers and Group 1-3, respectively. Table 5.10 summarizes percentage agreement and Cohen's Kappa for relevance judgments between workers and the TREC assessors. Both the kappa and the percentage agreement show a greater agreement for relevance judgments between Group 3 and TREC assessors compared with two other groups (Group 1 and Group 2). The kappa value of, 0.53 shows a moderate agreement for relevance judgments between Group 3 and TREC assessors, whilst there is a slight agreement between Group 1 and TREC assessors with kappa value of 0.19 and a fair agreement between Group 2 and TREC assessors (Kappa=0.39). The kappa value of 0.33 shows a fair agreement between relevance judgments made by all of the workers and relevance judgments provided by TREC assessors.

Table 5.9: Group agreement (workers and TREC assessors)

		TREC assessors	
		R	NR
All workers	R	62%	21%
	NR	5%	12%
Group 1 (low scores)	R	60%	24%
	NR	7%	9%
Group 2 (moderate scores)	R	62%	19%
	NR	5%	14%
Group 3 (high scores)	R	63%	15%
	NR	4%	18%

R, relevant documents; NR, non-relevant documents

Table 5.10: Summary of group agreement

Workers	Group agreement	Kappa
All workers	74%	0.33
Group 1(low scores)	69%	0.19
Group 2 (moderate scores)	76%	0.39
Group 3 (high scores)	81%	0.53

Some conclusions can be drawn from the group agreement described in this section. The agreement between each worker and the TREC assessor is not high when calculated individually, but it increases when workers are grouped. Group 3 who had high general reasoning skill were again more accurate in their relevance judgments for group agreement than Group 2 and Group 1. Moreover, the workers with moderate level of general reasoning skill (Group 2) were more accurate than the Group 1 who had low level of general reasoning skill.

5.3.4 Difference of Reliability of Judgments among Groups

Table 5.11 and Figure 5.2 present means for ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness for each group. For all measures of judgment reliability, the mean values for Group 3 (high scores) is higher than Group 2 (moderate scores) and Group 1 (low scores). This trend is also noticeable in Figure 5.2 that the workers who have higher general reasoning skill have greater ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness.

One-way statistical significance test was conducted to find statistically significant differences among these three groups (if any) for different measures including ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness. First, a test of homogeneity of variances was conducted and the results were negative for binary accuracy, recall, NDCG, specificity and effectiveness. Therefore, the Welch's test was applied for these measures and the one-way ANOVA test was used for ternary accuracy and precision (Table 5.12).

Table 5.11: Mean of judgment reliability measures

Group	Ternary accuracy	Binary accuracy	Precision	Recall	NDCG	Specificity	Effectiveness
1	0.39	0.62	0.69	0.79	0.86	0.30	0.21
2	0.50	0.75	0.79	0.89	0.89	0.41	0.32
3	0.51	0.78	0.81	0.90	0.90	0.52	0.44

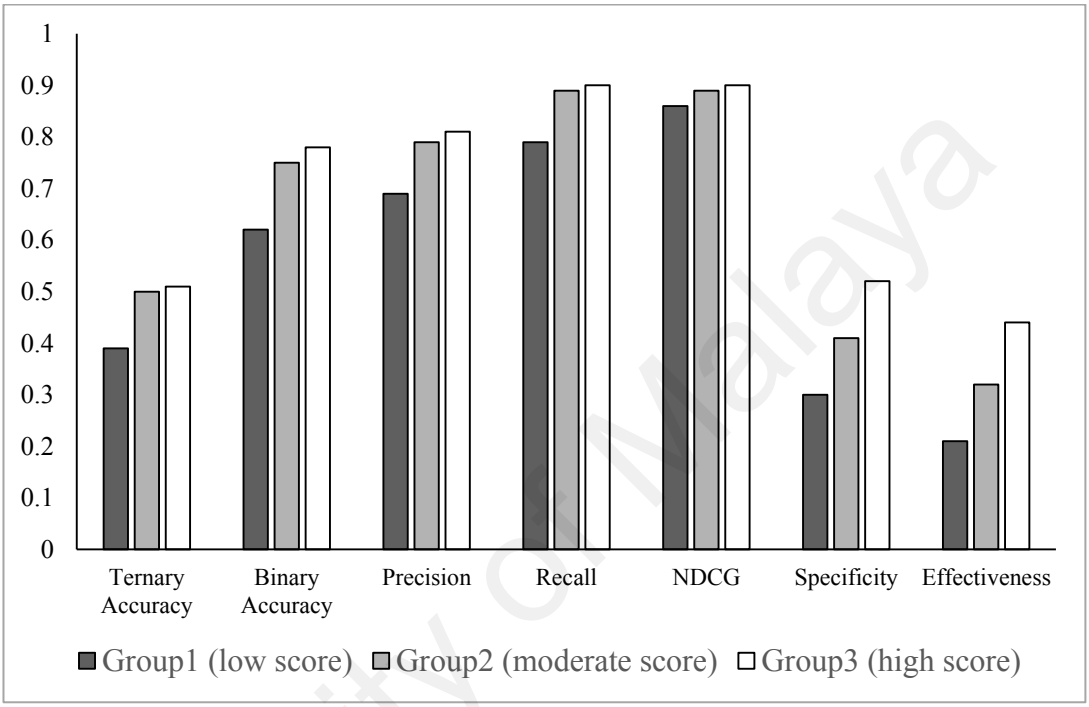


Figure 5.2: Mean of judgment reliability measures

Table 5.12: ANOVA and Welch's test

Measures	Group	df	F	Test	p
Ternary accuracy	Between Groups	2	21.39	ANOVA test	0.000
	Within Groups	902			
Binary accuracy	Between Groups	2	43.78	Welch's test	0.000
	Within Groups	576.123			
Precision	Between Groups	2	21.59	ANOVA test	0.000
	Within Groups	902			
Recall	Between Groups	2	15.89	Welch's test	0.000
	Within Groups	555.819			
NDCG	Between Groups	2	9.27	Welch's test	0.000
	Within Groups	565.801			
Specificity	Between Groups	2	18.69	Welch's test	0.000
	Within Groups	594.387			
Effectiveness	Between Groups	2	28.834	Welch's test	0.000
	Within Groups	599.387			

The tests showed statistically significant differences between three groups for all measures (ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness) for *p* value less than 0.001. Post hoc comparisons using the Games-Howell test (Appendix F) reveal that workers with high general reasoning skill have a significantly greater binary accuracy, recall, NDCG and specificity in their relevance judgments than those workers with low general reasoning score. Moreover, workers with moderate general reasoning skill show a significantly greater binary accuracy, recall, NDCG, specificity and effectiveness in their relevance judgments than those workers with low general reasoning score. The workers with high general reasoning skill show a significantly greater specificity and effectiveness than the workers with moderate general reasoning skill.

Using Bonferroni, workers with high general reasoning skill have a significantly greater ternary accuracy and precision in their relevance judgments than those workers with low general reasoning scores whilst the workers with moderate skill have a significantly greater ternary accuracy and precision in their relevance judgments than those workers with low general reasoning scores. However, no significant difference is found between the workers who have moderate and high general reasoning score in all measures except specificity.

Further, eta^2 effect size values were calculated for all measures of judgment reliability. Effect size value for binary accuracy (eta^2= 0.10) and effectiveness (eta^2= 0.06) suggested a moderate significance between groups. In addition, effect size value for ternary accuracy (eta^2= .04) and precision (eta^2= 0.05), recall (eta^2= 0.04), NDCG (eta^2= 0.02) and specificity (eta^2= 0.04) suggested a small practical significance between groups.

5.4 Effect of General Reasoning Skill on Rank Correlation

The effect of general reasoning skill of the workers on system rankings is examined in this section, to find out whether relevance judgments created by workers with higher level of general reasoning skill are more reliable for system rankings than the relevance judgments generated by workers with lower level of general reasoning skill. As explained earlier (see Section 3.5.2), there are five sets of relevance judgments: (i) a relevance judgment set provided by TREC assessors (a subset of *qrels*), (ii) a relevance judgment set created by all of the workers, (iii) a relevance judgment set created by Group 1 (low general reasoning skill), (iv) a relevance judgment set created by Group 2 (moderate general reasoning skill), and (v) a relevance judgment set created by Group 3 (high general reasoning skill). System rankings based on relevance judgments set created by all of the workers and relevance judgments set provided by TREC assessors using MAP (k=1000) is shown in Figure 5.3 and using MAP (k=10) in Figure 5.4.

Using relevance judgment sets, 25 systems scored and ranked by Group 1 as the gold data. Then, the systems scored again based on relevance judgment set generated by Group 2. Finally, the systems scored by Group 3. The System rankings for relevance judgment sets using MAP (k=1000 and k=10) is shown in Figure 5.5 and Figure 5.6. System rankings which are based on the relevance judgment sets which created by Group 1 and Group 2 are quite comparable. However, as it is noticeable in Figure 5.5 and 5.6, system rankings based on relevance judgments made by Group 3 (high scores) is relatively closer to the system rankings based on judgments provided by TREC assessors than Group 1 and Group 2. Table 5.13 shows Kendall's tau correlation that was computed for the rank comparison between different sets of relevance judgments of workers and TREC assessors.

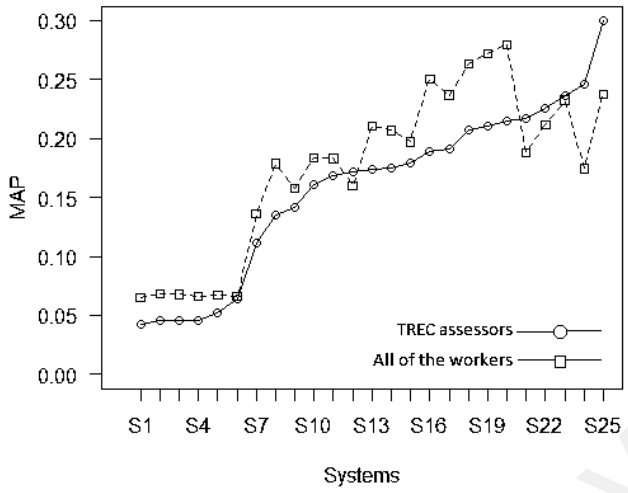


Figure 5.3: System rankings for all workers; MAP (*k*=1000). The systems are sorted in ascending order of MAP (*k*=1000) scores using TREC assessors judgments.

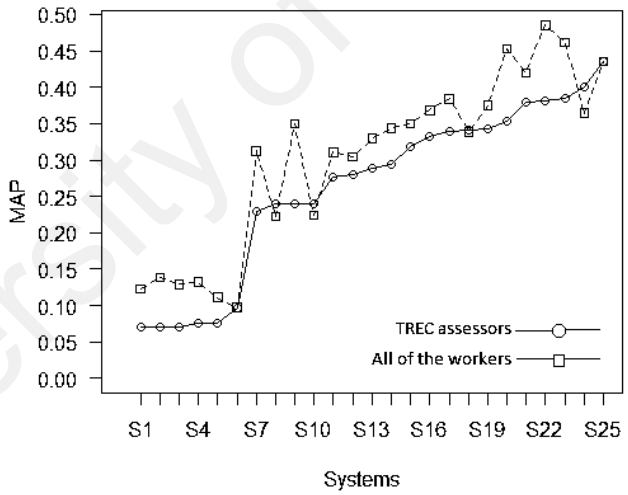


Figure 5.4: System rankings for all workers; MAP (*k*=10). The systems are sorted in ascending order of MAP (*k*=10) scores using TREC assessors judgments.

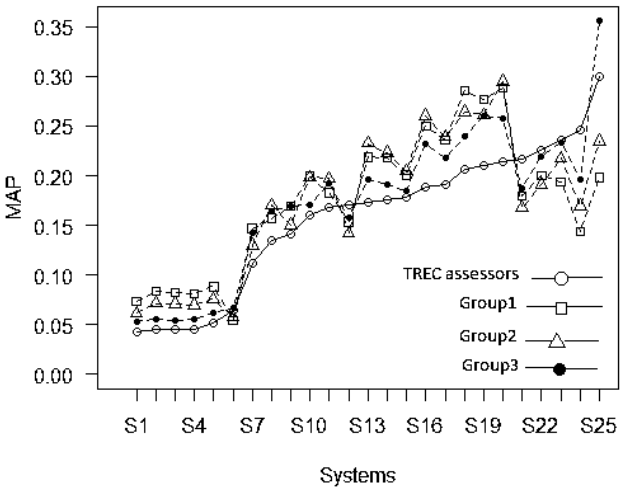


Figure 5.5: System rankings for groups; MAP (*k*=1000). The systems are sorted in ascending order of MAP (*k*=1000) scores using TREC assessors judgments.

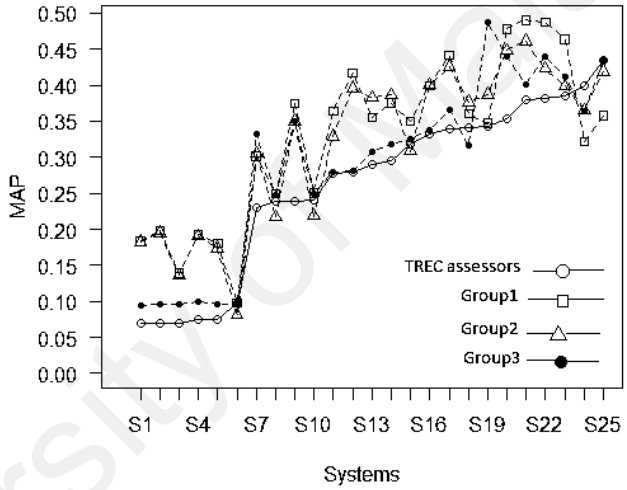


Figure 5.6: System rankings for groups; MAP (*k*=10). The systems are sorted in ascending order of MAP (*k*=10) scores using TREC assessors judgments.

Table 5.13: Kendall's tau correlation

Groups	Kendall's tau	
	MAP (*k*=1000)	MAP (*k*=10)
All workers	0.64	0.74
Group 1 (low scores)	0.51	0.55
Group 2 (moderate scores)	0.57	0.66
Group 3 (high scores)	0.76	0.73

The Kendall's tau correlation coefficients is also shows the highest correlation between system rankings based on Group 3 judgments and system rankings based on the TREC assessors judgments with a tau value of 0.76 for MAP(*k*=1000) and 0.73 for MAP

(10) (Table 5.13). The Kendall's tau correlation coefficients between Group 2 and the TREC assessors is 0.57 for MAP (k=1000) and 0.66 for MAP (k=10), which is slightly higher than that of for Group 1 (0.51 for MAP (k=1000) and 0.55 for MAP (k=10)).

According to the tau value, the system rankings based on relevance judgments set created by Group 1 and Group 2 were relatively similar whilst the system rankings based on Group 3 judgments were a little closer to the system rankings based on TREC assessors. Number of studies found that the variation in relevance judgments do not have a significant influence on system rankings. Harter (1996) summarized the studies which found that variation in relevance judgments do not significantly affect the system rankings. These studies investigated the effects of variations in relevance judgments on measures of retrieval effectiveness (Lesk & Salton, 1968; C. W. Cleverdon, 1970; Kazhdan, 1979; Burgin, 1992). Different types of judges including different groups of topic experts, librarians, topic originators and etc. in these studies. All of these studies found significant variation in relevance judgments among different judges whilst these variations have no considerable effect on system rankings. In a preliminary study (Voorhees, 2000), the relevance judgments of both NIST judges and University of Waterloo judges were compared for a TREC-6 dataset. The Kendall's tau correlation between these two groups showed 0.896 for 76 systems ranked by MAP. The study concluded that the variation in relevance judgments rarely influence the system rankings. In a separate study (Trotman & Jenkinson, 2007), the Spearman's r rank correlation between multiple judges and gold set for 64 systems was 0.95. They concluded that different judges have a little effect on system rankings.

In a recent study (Nowak & Rüger, 2010), the Kendall's tau test assigns a high correlation in ranking between the combined ground-truth of the workers and the combined ground-truth of the experts. To sum up, our results of system rankings are

consistent with the studies that reported that different judges have a little effect on system rankings.

5.5 Effect of Self-Reported Competence on Accuracy of Judgments

In this section, the effect of various self-reported competence about the workers including confidence in relevance judgments, difficulty of the relevance judgments and knowledge on the topic on crowdsourced judgment reliability was investigated. In each task and for every relevance judgment that they generated, the workers have to rate their confidence in their evaluation using a 4-point Likert scale, from not confident to very confident. Table 5.14 presents the ternary and binary accuracy for each level of confidence across the 4525 relevance judgments. As shown in Table 5.14, the binary accuracy is increasing as the level of confidence is increasing and it shows that workers who were more confident with their judgments obtained a higher binary accuracy whilst less confident workers achieved lower accuracy. Chi-square test shows that the relationship between confidence and ternary accuracy is significant ($\chi 2 = 12.382$, $p < 0.01$). The relationship between confidence and binary accuracy is also significant ($\chi 2 = 120.685$, $p < 0.001$).

Table 5.14: Self-reported competence and accuracy of judgments

	level	Judgments	Ternary correct judgments	Binary correct judgments	Ternary accuracy	Binary accuracy
Confidence in judgment	1	164	71	95	0.43	0.58
	2	703	330	423	0.47	0.60
	3	1593	722	1131	0.45	0.71
	4	2065	1047	1637	0.51	0.79
Difficulty of the judgment	1	2244	1131	1769	0.50	0.79
	2	1394	639	964	0.46	0.69
	3	736	334	469	0.45	0.64
	4	151	66	84	0.44	0.56
Knowledge on the topic	1	391	1955	1429	0.5	0.73
	2	238	1190	882	0.48	0.74
	3	205	1025	717	0.45	0.70
	4	71	355	258	0.47	0.73

Ratings are based on a 4-point Likert-type scale, confidence in judgment: Not confident 1 2 3 4 Very confident, difficulty of the judgment: Easy 1 2 3 4 Difficult, knowledge on the topic: Minimal 1 2 3 4 Extensive.

Workers also have to rate the difficulty of the evaluation using a 4-point Likert scale, from easy to difficult for each relevance judgment that they made. The binary and ternary accuracy are calculated across all the 4525 relevance judgment for each level of difficulty to see the effect of difficulty of judgment on accuracy of relevance judgments. The ternary and binary accuracy for each level of difficulty is shown in Table 5.14. The results show that while the difficulty of the judgment is increasing, the both ternary and binary accuracy is decreasing. A Chi-square test shows a relationship between difficulty and ternary accuracy to be significant ($\chi 2 = 10.925$, $p < 0.05$). Chi-square test also shows a significant relationship between difficulty and binary accuracy ($\chi 2 = 103.204$, $p < 0.001$). Workers also rated their knowledge of the given topic using a 4-point Likert scale, from minimal to extensive. The accuracy for both ternary and binary statistics reported in Table 5.14 but surprisingly the relationship between accuracy and knowledge on the topic is not significant under a Chi-square test for independence.

5.6 Effect of Demographics on Accuracy of Judgments

In this section, demographic information of the workers including age, gender, education, country, computer experience and Internet experience were examined to see how various demographics relate to the accuracy of relevance judgments. Table 5.15 shows the ternary and binary accuracy for each demographic information, across the 905 HITs. Chi-square tests found the relationship between demographic information and judgment reliability to be not significant and there is no relationship between demographic information and judgment reliability.

Table 5.15: Demographics and accuracy of judgments

			Correct judgments		Accuracy	
	Level	Judgments	Ternary	Binary	Ternary	Binary
Age	not yet 20	100	46	63	0.46	0.63
	in my 20's	1135	542	820	0.48	0.72
	in my 30's	1295	611	939	0.47	0.72
	in my 40's	1175	558	859	0.47	0.73
	in my 50's	450	224	331	0.50	0.74
	60+ years old	370	189	274	0.51	0.74
Gender	Male	2485	1196	1806	0.48	0.73
	Female	2040	974	1480	0.48	0.72
Education	no education	15	8	10	0.53	0.67
	primary school	125	60	91	0.48	0.73
	high school	1815	854	1302	0.47	0.72
	Bachelor degree	2020	999	1473	0.49	0.73
	master degree	440	189	323	0.43	0.73
	PhD or higher	110	60	87	0.54	0.79
Computer experience	1	10	4	7	0.40	0.70
	2	240	101	164	0.42	0.68
	3	1580	761	1134	0.48	0.72
	4	2695	1304	1981	0.48	0.73
Internet experience	1	10	4	8	0.40	0.80
	2	185	79	124	0.43	0.67
	3	1635	790	1197	0.48	0.73
	4	2695	1297	1957	0.48	0.73
Country	AUS	75	40	56	0.53	0.75
	BHS	15	5	8	0.33	0.53
	CAN	870	421	639	0.48	0.73
	GBR	1650	809	1200	0.49	0.73
	IRL	360	158	259	0.44	0.72
	NZL	35	14	28	0.40	0.80
	USA	1520	723	1096	0.48	0.72

5.7 Summary

This chapter presented findings from the general reasoning experiment as well as discussion on the results. The findings support that general reasoning skill of workers do influence the reliability of relevance judgments in crowdsourcing and those workers with higher levels of general reasoning skill appeared more reliable in performing relevance judgments. Our results showed that relevance judgments provided by workers with higher general reasoning skills are relatively more reliable for system rankings than that of made by workers with lower level of general reasoning skills. In the next chapter, the results of logical reasoning experiment are presented.

CHAPTER 6: LOGICAL REASONING EXPERIMENT

This chapter presents the results of logical reasoning experiment, looking into the filtering method used in this experiment and details of the collected data. The main goal of this experiment is to investigate the effect of logical reasoning skill on reliability of crowdsourced relevance judgments. The presentation and discussion of the results are provided in five main parts: (i) the effect of logical reasoning skill on the reliability of judgments, (ii) the effect of logical reasoning skill on system rankings, (iii) the effect of self-reported competence on accuracy of judgments and (iv) the effect of workers' demographics on accuracy of judgments. This experiment is provided in Appendix D.

6.1 Filtering Spam

Overall, 519 workers participated in the logical reasoning experiment. The number of reliable HIT assignments and the total number of performed HITs, including the rejected ones, are shown in Table 6.1. From 1000 HITs, 25 HITs were removed because of failing to answer the trap question. From 975 HITs, 186 HITs were considered as unreliable HITs because of the task completion time, which was less than 2 minute. Finally, from 1000 HITs, 789 HITs or 3945 judgments were recognized "reliable HITs".

6.2 Descriptive Statistics

Demographic data of workers in logical reasoning experiment is provided in Table 6.2 for "all HITs", "cleaned HITs" and "rejected HITs". Looking at the distribution of workers, the majority of workers were male (57.2% in "all HITs", 51.8% in "cleaned HITs" and 77.3% in "rejected HITs") from USA (45.5% in "all HITs", 44.2% in "cleaned HITs" and 50.2% in "rejected HITs"), aged 20-30 (36.7% in "all HITs", 30.8% in "cleaned HITs" and 58.8% in "rejected HITs"). Their educational level was mostly high school (44% in "all HITs", 45.2% in "cleaned HITs") with a high level of Computer knowledge (55.8% in "all HITs", 57.8% in "cleaned HITs" and 48.3% in "rejected HITs") and Internet experience (53.7% in "all HITs" and 55.6% in "cleaned HITs").

Table 6.1: Summary of HITs

HITs type	Number of HITs	Judgments	Workers
All HITs	1000	5000	519
Cleaned HITs (Reliable)	789	3945	478
Rejected HITs	211	1055	91

Table 6.2: Demographics of participant

	level	All HITs Percent	Cleaned HITs Percent	Rejected HITs Percent
Age	Less than 20	2.6	2.8	1.9
	in my 20's	36.7	30.8	58.8
	in my 30's	27.8	28.6	24.6
	in my 40's	17.4	19	11.4
	in my 50's	10.6	12.7	2.8
	60+ years old	4.9	6.1	0.5
Gender	Male	57.2	51.8	77.3
	Female	42.8	48.2	22.7
Education	No education	0.2	0	0.9
	Basic schooling	1.4	1.4	1.4
	High school	44	45.2	39.3
	Bachelor degree	41.4	40.3	45.5
	Master degree	11.5	11.8	10.4
	PhD or higher	1.5	1.3	2.4
Computer Experience	1	0.5	0.3	1.4
	2	11.9	8.1	26.1
	3	31.8	33.8	24.2
	4	55.8	57.8	48.3
Internet Experience	1	0.4	0.4	0.5
	2	5.7	5.7	5.7
	3	40.2	38.3	47.4
	4	53.7	55.6	46.4
Country	AUS	2.3	2.4	1.9
	BHS	0	0	0
	CAN	17.9	18.4	16.1
	GBR	29.6	30.4	26.5
	IRL	3.7	3.3	5.2
	NZL	1	1.3	0
	USA	45.5	44.2	50.2

Table 6.3 lists the number of HITs, minimum, maximum, mean score and standard deviation for each measure for "all HITs", "cleaned HITs" and "rejected HITs". All measures (logical reasoning score, ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness) were calculated for each HIT based on the formulas, which explained in Chapter 2 and 3. As logical reasoning score was calculated based on a test of 10 questions, the maximum score of logical reasoning is 10 in "all HITs" and "cleaned HITs". The mean of logical reasoning score for "all HITs" showed greater value (2.58) than the mean of logical reasoning score for "rejected HITs" (0.52) whilst the mean of logical reasoning score for cleaned data showed greatest value (3.13). The mean of ternary accuracy, binary accuracy, precision, NDCG, specificity and effectiveness for "rejected HITs" were less than that of parameters for "all HITs" and "cleaned HITs". In addition, mean of ternary accuracy, binary accuracy, precision, specificity and effectiveness were higher for "cleaned HITs" than that of parameters for "all HITs".

Table 6.4 shows descriptive statistics for self-reported competence of participants in doing their tasks. The most common responses reported by the workers were to state high confidence in their judgments (level 4) for "all HITs" (38.9%), "cleaned HITs" (41%) and "rejected HITs" (31.4%). Moreover, the majority of workers reported that the task was not difficult (level 1) for "all HITs" (42.1%), "cleaned HITs" (42.6%) and "rejected HITs" (40.1%). The level of knowledge on topic reported for the relevance judgment was 1 representing minimally familiar with the topic for "all HITs" (40.3%) and "cleaned HITs" (44.5%).

Table 6.3: Descriptive statistics for analysed measures

	Measures	HITs	Minimum	Maximum	Mean	SD
All HITs	Logical reasoning score	1000	-6	10	2.58	3.40
	Ternary accuracy	1000	0	1	0.46	0.23
	Binary accuracy	1000	0	1	0.70	0.22
	Precision	1000	0	1	0.75	0.24
	Recall	1000	0	1	0.87	0.23
	NDCG	1000	0.43	1	0.88	0.10
	Specificity	1000	0	1	0.37	0.43
	Effectiveness	1000	0	1	0.29	0.37
Cleaned HITs	Logical reasoning score	789	-6	10	3.13	3.20
	Ternary accuracy	789	0	1	0.47	0.23
	Binary accuracy	789	0	1	0.71	0.22
	Precision	789	0	1	0.76	0.24
	Recall	789	0	1	0.86	0.23
	NDCG	789	0.43	1	0.88	0.10
	Specificity	789	0	1	0.40	0.43
	Effectiveness	789	0	1	0.32	0.38
Rejected HITs	Logical reasoning score	211	-5	8	0.52	3.38
	Ternary accuracy	211	0	1	0.41	0.24
	Binary accuracy	211	0.2	1	0.65	0.21
	Precision	211	0	1	0.68	0.25
	Recall	211	0	1	0.89	0.20
	NDCG	211	0.43	1	0.86	0.11
	Specificity	211	0	1	0.24	0.38
	Effectiveness	211	0	1	0.19	0.33

Table 6.4: Descriptive statistics of self-reported competence

	level	All HITs Percent	Cleaned Percent	Rejected Percent
Confidence in judgment	1	3.9	4	3.3
	2	23.6	20.8	34
	3	33.6	34.2	31.3
	4	38.9	41	31.4
Difficulty of the judgment	1	42.1	42.6	40.1
	2	29.4	30	27.3
	3	24	22.5	29.4
	4	4.5	4.8	3.2
Knowledge on the topic	1	40.3	44.5	24.6
	2	20.4	21.7	15.6
	3	31	26.7	46.9
	4	8.3	7.1	12.8

Ratings are based on a 4-point Likert-type scale, confidence in judgment: Not confident 1 2 3 4 Very confident, difficulty of the judgment: Easy 1 2 3 4 Difficult, knowledge on the topic: Minimal 1 2 3 4 Extensive.

The number of HITs judged by each worker in logical reasoning experiment (cleaned data) is shown in Figure 6.1. Number of workers judged all of the HITs, with the most hard-working worker went through all 20 HITs, while a long tail of workers worked on a single task only.

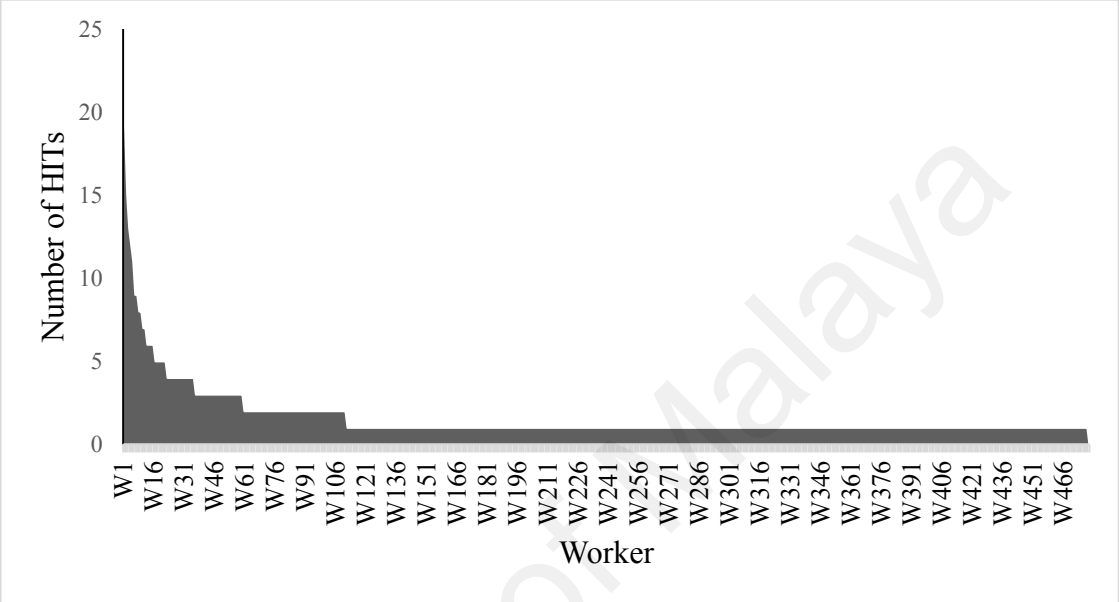


Figure 6.1: Number of HITs judged by each worker

6.3 Effect of Logical Reasoning Skill on Reliability of Judgments

Workers were divided into three groups according to their logical reasoning scores:

- Group 1- low scores: with the logical reasoning score less than or equal to two consisting of 272 HITs.
- Group 2- moderate scores: with the logical reasoning score between two and four, consisting of 305 HITs.
- Group 3- high scores: with the logical reasoning score more than four, consisting of 212 HITs.

Accordingly, this Section provides the results for correlation between workers' judgment reliability and logical reasoning score. In addition to the results for individual agreement and group agreement between workers and TREC judgments for relevance judgments, results for statistical differences among groups for reliability of relevance judgments is described in this Section. The intention behind these investigations is to find out whether there is any association between logical reasoning skill of the workers and their reliability of relevance judgments and whether the workers with high logical reasoning skill are more reliable in terms of creating relevance judgments.

6.3.1 Correlation Coefficient

The correlation matrix using Pearson for eight measures (ternary accuracy, binary accuracy, precision, recall, NDCG, specificity, effectiveness and logical reasoning score) is shown in Table 6.5. Logical reasoning score is positively correlated with other measures. Binary accuracy and logical reasoning has a moderate correlation (r=0.30) whilst there is a small but significant correlation between logical reasoning skill on the one hand, and ternary accuracy (r=0.24), precision (r=0.20), recall (r=0.20), NDCG (r=0.12), specificity (r=0.16) and effectiveness (r=0.20) on the other hand.

Table 6.5: Pearson correlation matrix for eight measures

Measures	1	2	3	4	5	6	7	8
1. Logical reasoning score	-							
2. Ternary accuracy	0.24**	-						
3. Binary accuracy	0.30**	0.64**	-					
4. Precision	0.20**	0.39**	0.73**	-				
5. Recall	0.20**	0.40**	0.59**	0.12**	-			
6. NDCG	0.12**	0.43**	0.34**	0.34**	0.14**	-		
7. Specificity	0.16**	0.39**	0.33**	0.42**	-0.21**	0.32**	-	
8.Effectiveness	0.20**	0.51**	0.55**	0.47**	0.15**	0.37**	0.89**	-

Note: $p<0.01$ (**)

Correlation of ternary accuracy with binary accuracy (r=0.64) and effectiveness (r=0.51) is strong whilst that of with precision (r=0.39), recall (r=0.40), NDCG (r=0.43) and specificity (r=0.39) is moderate. Binary accuracy is largely correlated with precision (r=0.73), recall (r=0.59) and effectiveness (r=0.55) and moderately correlated with NDCG (r=0.34) and specificity (r=0.33). Having a small correlation with recall (r=0.12), precision has a moderate correlation with NDCG (r=0.34), specificity (r=0.42) and effectiveness (r=0.47). Recall has a small positive correlation with NDCG (r=0.14) and small negative correlation with specificity (r=-0.21). Effectiveness has a small correlation with recall (r=0.15), moderate correlation with NDCG (r=0.37) and large correlation with specificity (r=0.89).

Logical reasoning score was found to be correlated with reliability of relevance judgments. Although, there is no previous research, which investigated the effect of logical reasoning skill on crowdsourced relevance judgment, there are a few studies, which investigated the effect of logical reasoning skill on search process. Logical reasoning skill of the user had been thought to be a characteristic of a successful information searcher (Teitelbaum-Kronish, 1984). Students of library school who had high logical reasoning skill performed better in doing online searches and the logical reasoning skill was introduced as a predictor to searching performance. In a separate study, C. S. Kim (2002) showed that logical reasoning ability was strongly correlated with search outcome. Allen (1994a) found that there is a relationship between information searching performance and logical reasoning skill. The current study indicated that workers' logical reasoning skill was correlated with reliability of relevance judgments. A moderate correlation between logical reasoning score and binary accuracy is consistent with (Teitelbaum-Kronish, 1984; Allen, 1994a; C. S. Kim, 2002) which found moderate correlation between logical reasoning skill and search performance.

6.3.2 Individual Agreement (Workers vs. TREC Assessors)

Table 6.6 shows ternary and binary agreement for relevance judgment between all of the workers and TREC assessors. There is a 47.81% (14.48% on "highly-relevant documents", 19.26% on "relevant documents" and 14.07% on "non-relevant documents") ternary agreement for relevance judgments between all of the workers and TREC assessors. The binary agreement between relevance judgments made by all of the workers and relevance judgments provided by TREC assessors is 71.59% (57.52% on "relevant documents" and 14.07% on "non-relevant documents").

Ternary agreement and binary agreement for relevance judgments between each group of workers (consisted of Group 1 for low scores, Group 2 for moderate scores and Group 3 for high scores) and the TREC assessors is presented in Table 6.7. There is 40.73% ternary agreement between relevance judgments made by Group 1 and relevance judgments provided by TREC assessors (11.47% on "highly-relevant documents", 19.48% on "relevant documents" and 9.78% on "non-relevant documents"). The ternary agreement for relevance judgments between Group 2 and TREC assessors is 51.21% (15.47% on "highly-relevant documents", 20% on "relevant documents" and 15.74% on "non-relevant documents"). Ternary agreement between relevance judgments made by Group 3 and relevance judgments provided by TREC assessors is 51.99% (16.89% on "highly-relevant documents", 17.93% on "relevant documents" and 17.17% on "non-relevant documents"). The binary agreement for relevance judgments between different groups of workers and the TREC assessors is also shown in Table 6.7. There is 63.53% binary agreement (53.75% on "relevant documents", 9.78% on "non-relevant documents") between relevance judgments generated by Group 1 and relevance judgments provided by TREC assessors whilst the binary agreement between Group 2 and TREC assessors is 74.04% (58.3% on "relevant documents", 15.74% on "non-relevant documents"). The binary agreement between relevance judgments made by

Group 3 and relevance judgments provided by TREC assessors is 78.4% (61.23% on "relevant documents" and 17.17% on "non-relevant documents").

Table 6.6: Agreement (Ternary and Binary) for all workers

		TREC assessors				
		Ternary agreement			Binary agreement	
		HR	R	NR	R	NR
All workers	HR	14.48%	11.89%	3.85%	-	-
	R	11.89%	19.26%	14.32%	57.52%	18.17%
	NR	4.06%	6.18%	14.07%	10.24%	14.07%

HR, highly-relevant; R, relevant; NR, non-relevant

Table 6.7: Agreement (Ternary and binary) for groups of workers

Workers		TREC assessors				
		Ternary agreement			Binary agreement	
		HR	R	NR	R	NR
Group 1 (low scores)	HR	11.47%	8.38%	3.68%	-	-
	R	14.41%	19.48%	18.24%	53.75%	21.91%
	NR	5%	9.56%	9.78%	14.56%	9.78%
Group 2 (moderate scores)	HR	15.47%	12.85%	4.13%	-	-
	R	9.97%	20%	13.12%	58.3%	17.24%
	NR	4.2%	4.52%	15.74%	8.72%	15.74%
Group 3 (high scores)	HR	16.89%	15%	3.68%	-	-
	R	11.41%	17.93%	11.04%	61.23%	14.71%
	NR	2.64%	4.24%	17.17%	6.89%	17.17%

HR, highly-relevant; R, relevant; NR, non-relevant

A summary of the individual agreement for relevance judgments between workers and the TREC assessors is presented in Table 6.8 providing Cohen's Kappa agreement for ternary agreement and binary agreement. Comparison between overall ternary agreement and overall binary agreement (both percentage and kappa), showed that the ternary agreement is lower since for ternary agreement both assessors have to agree on the exact level of relevancy. Group 3 has the highest agreement with TREC assessors for

relevance judgments for both ternary (51.99%) and binary agreement (78.4%). The binary agreement for relevance judgments between Group 3 and TREC assessors shows a moderate agreement with a kappa value of 0.47. The binary agreement for relevance judgments between Group 1 and TREC assessors is a slight agreement with kappa value of 0.10 whilst the binary agreement between Group 2 TREC assessors is a fair agreement (0.37) and lower than the binary agreement between Group 3 and TREC assessors (0.47).

Table 6.8: Summary of individual agreement

Workers	Ternary agreement		Binary agreement	
	Percentage	Kappa	Percentage	Kappa
All workers	47.81%	0.21	71.59%	0.30
Group 1(low scores)	40.73%	0.09	63.53%	0.10
Group 2 (moderate scores)	51.21%	0.26	74.04%	0.37
Group 3 (high scores)	51.99%	0.27	78.4%	0.47

6.3.3 Group Agreement (Workers vs. TREC Assessors)

Agreements for relevance judgments between the TREC assessors and the rest of relevance judgment sets including all workers, Group 1 (low logical reasoning skill), Group 2 (moderate logical reasoning skill) and Group 3 (high logical reasoning skill) is presented in Table 6.9. The level of agreement between workers and the TREC assessors on "relevant documents" is 63% for all of the workers. Group comparison showed that 61% of workers in Group 1, 61% in Group 2 and 65% in Group 3 have an agreement with TREC assessors on "relevant documents". The agreement between workers and the TREC assessors on "non-relevant documents" is 12%, 6%, 14% and 19% for all workers and Group 1-3, respectively. There is most disagreement between workers and TREC assessors on documents, which workers marked relevant and TREC assessors marked not relevant (21% for all of the workers, 27% for Group 1, 19% for Group 2, 14% for Group 3).

Table 6.10 summarizes the group agreement between relevance judgments made by crowdsourced workers and relevance judgments provided by TREC assessors using percentage agreement and Cohen's kappa for each group of workers. The kappa agreement for relevance judgments between Group 1 and TREC assessors shows slight agreement (0.11) whilst there is a fair agreement between Group 2 and TREC assessors (0.37). Relevance judgments made by Group 3 has the highest agreement with relevance judgments provided by TREC assessors and shows a moderate (nearly substantial) agreement (0.60). Indeed, the higher agreement both for individual and group agreement with relevance judgments provided by TREC assessors indicates that Group 3 is more reliable than other two groups and Group 2 is also more reliable than Group 1 in terms of creating relevance judgments.

Table 6.9: Group agreement (workers and TREC assessors)

		TREC assessors	
		R	NR
All workers	R	63%	21%
	NR	4%	12%
Group 1 (low scores)	R	61%	27%
	NR	6%	6%
Group 2 (moderate scores)	R	61%	19%
	NR	6%	14%
Group 3 (high scores)	R	65%	14%
	NR	2%	19%

R, relevant documents; NR, non-relevant documents

Table 6.10: Summary of group agreement

Workers	Group agreement	Kappa
All workers	75%	0.35
Group 1 (low scores)	67%	0.11
Group 2 (moderate scores)	75%	0.37
Group 3 (high scores)	84%	0.60

6.3.4 Difference of Reliability of Judgments among Groups

Table 6.11 presents means for ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness for each group. For all measures of judgment reliability, the mean values for Group 3 (high scores) is higher than Group 2 (moderate scores) and Group 1 (low scores) as this pattern is also noticeable in Figure 6.2. Comparing Group 1, Group 2 and Group 3, higher logical reasoning skill leads to higher ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness.

One-way statistical significance test was conducted to find statistically significant differences among these three groups (if any) for different measures including ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness. First, a test of homogeneity of variances was conducted for each measurement. The results of test of homogeneity of variances were negative for all measures except ternary accuracy and specificity. Therefore, the Welch's test was applied for binary accuracy, precision, recall and NDCG and the ANOVA test was used for ternary accuracy and specificity (Table 6.12).

There is a statistically significant difference between three groups for all measures (ternary accuracy, binary accuracy, precision, recall, NDCG, specificity and effectiveness) for p value less than 0.01. Post hoc comparisons using the Bonferroni test show that there is a significant difference between low logical reasoning skill workers on the one hand and moderate and high logical reasoning skill workers on the other hand in ternary accuracy and specificity. However, there is not any significant difference between workers with moderate and high logical reasoning skill in ternary accuracy and specificity.

Table 6.11: Mean of judgment reliability measures

Group	Ternary accuracy	Binary accuracy	Precision	Recall	NDCG	Specificity	Effectiveness
1	0.40	0.63	0.70	0.80	0.86	0.30	0.21
2	0.47	0.71	0.76	0.86	0.88	0.45	0.36
3	0.51	0.78	0.82	0.91	0.89	0.46	0.40

Table 6.12: ANOVA and Welch's test

Measure	Group	df	F	Test	p
Ternary accuracy	Between Groups	2	19.84	ANOVA test	0.000
	Within Groups	786			
Binary accuracy	Between Groups	2	28.75	Welch's test	0.000
	Within Groups	497.97			
Precision	Between Groups	2	14.51	Welch's test	0.000
	Within Groups	499.561			
Recall	Between Groups	2	11.58	Welch's test	0.000
	Within Groups	501.77			
NDCG	Between Groups	2	5.25	Welch's test	0.006
	Within Groups	507.93			
Specificity	Between Groups	2	11.249	ANOVA test	0.000
	Within Groups	786			
Effectiveness	Between Groups	2	19.75	Welch's test	0.000
	Within Groups	488.100			

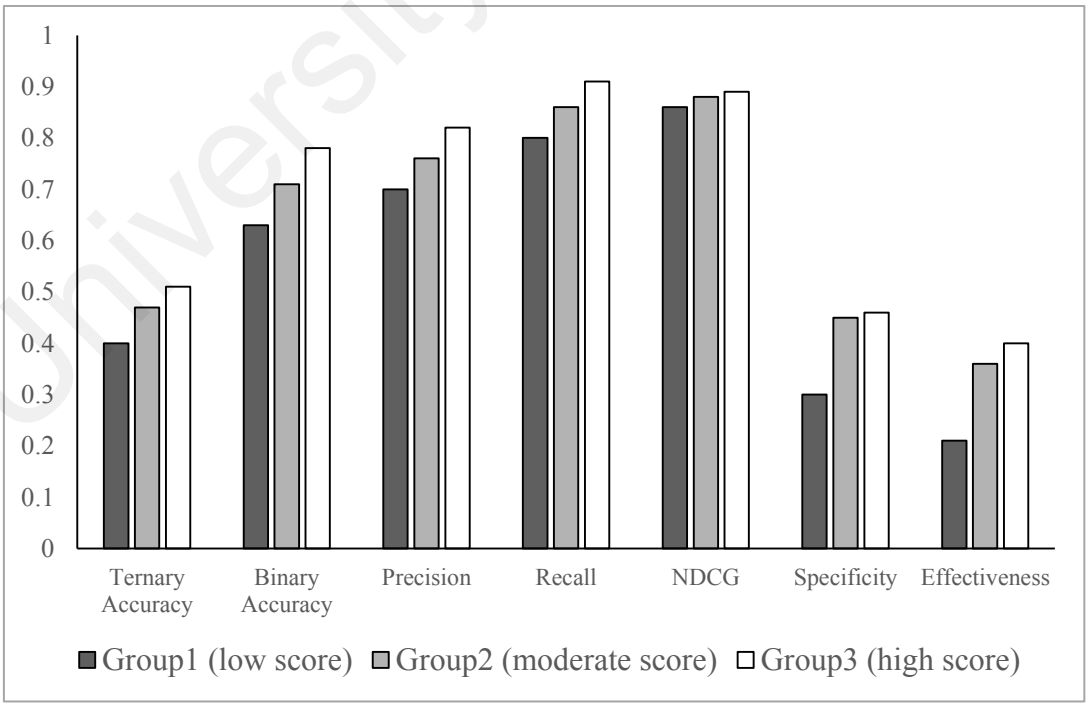


Figure 6.2: Mean of judgment reliability measures

Post hoc comparisons using the Games-Howell test reveal that workers with high logical reasoning skill showed a significantly greater binary accuracy in their relevance judgments than those workers with moderate logical reasoning skill or those workers with low logical reasoning skill. Workers with a moderate logical reasoning skill showed a significantly greater binary accuracy in their relevance judgments than those workers with low logical reasoning skill. In addition, Post hoc comparisons using the Games-Howell test reveal that workers with low logical reasoning skill showed a significantly lower precision, recall, NDCG and effectiveness in their relevance judgments than those workers with moderate logical reasoning skill or those workers with high logical reasoning skill. However, there is not any significant difference between workers with moderate and high logical reasoning skill in precision, recall, NDCG and effectiveness (Appendix G). Effect size value for binary accuracy (eta^2= 0.07) suggested a moderate significance among groups. Effect size value for ternary accuracy (eta^2= .05), precision (eta^2= 0.04), recall (eta^2= 0.03), specificity (eta^2= 0.03), effectiveness (eta^2= 0.04) and NDCG (eta^2= 0.01) showed a small practical significance among groups.

Indeed, the significance test showed that there is a significant difference between three groups. However, the Post-hoc test showed that there was not any significant difference between Group 2 and Group 3 in their ternary accuracy, precision, recall and NDCG. The Group 2 and Group 3 had a significant difference only in binary accuracy. The effect size showed that among the five measures used to find judgment reliability, the effect size was moderate for binary accuracy and small for other measures.

6.4 Effect of Logical Reasoning Skill on Rank Correlation

As explained earlier (see Section 3.5.2), there are five sets of relevance judgments: (i) a relevance judgment set provided by TREC assessors (a subset of *qrels*), (ii) a relevance judgment set created by all of the workers, (iii) a relevance judgment set created by Group 1 (low logical reasoning skill), (iv) a relevance judgment set created by Group 2 (moderate logical reasoning skill), and (v) a relevance judgment set created by Group 3 (high logical reasoning skill). System rankings based on relevance judgments set generated by all of the workers and relevance judgments set provided by TREC assessors using MAP (k=1000) is shown in Figure 6.3 and using MAP (k=10) in Figure 6.4.

Further, the 25 systems were scored and ranked using relevance judgments set created by Group 1 (low scores), Group 2 (moderate scores) and Group 3 (high scores). System rankings based on relevance judgments set created by Group 1, Group 2, Group 3 and TREC assessors using MAP (k=1000) is shown in Figure 6.5 and using MAP (k=10) in Figure 6.6. As it is observable in Figure 6.5 and 6.6, system rankings based on relevance judgments set made by Group 3 (high scores) is relatively closer to the system rankings based on TREC assessors judgments compared with two other system rankings (Group 1 and Group 2). System rankings based on relevance judgments made by Group 2 (moderate scores) is also a little closer to the system rankings based on TREC assessors judgments than Group 1. Besides, Table 6.13 shows Kendall's tau correlation that was computed for the rank comparison between different sets of relevance judgments and TREC assessors.

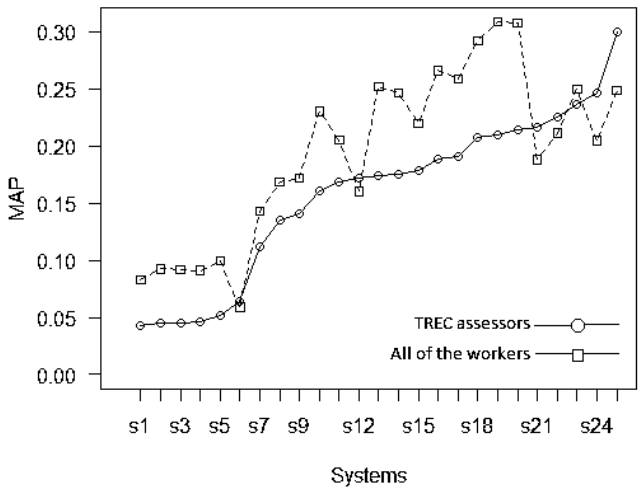


Figure 6.3: System rankings for all workers; MAP (*k*=1000). The systems are sorted in ascending order of MAP (*k*=1000) scores using TREC assessors judgments.

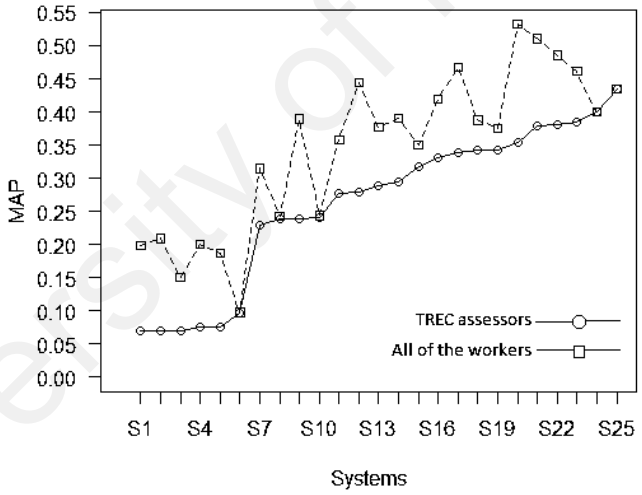


Figure 6.4: System rankings for all workers; MAP (*k*=10). The systems are sorted in ascending order of MAP (*k*=10) scores using TREC assessors judgments.

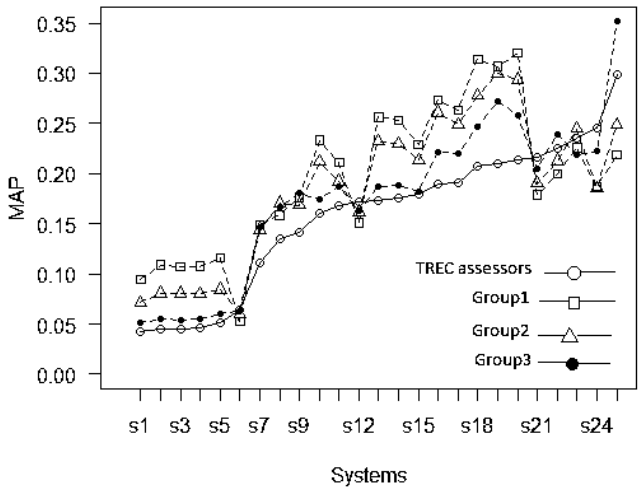


Figure 6.5: System rankings for groups; MAP (*k*=1000). The systems are sorted in ascending order of MAP (*k*=1000) scores using TREC assessors judgments.

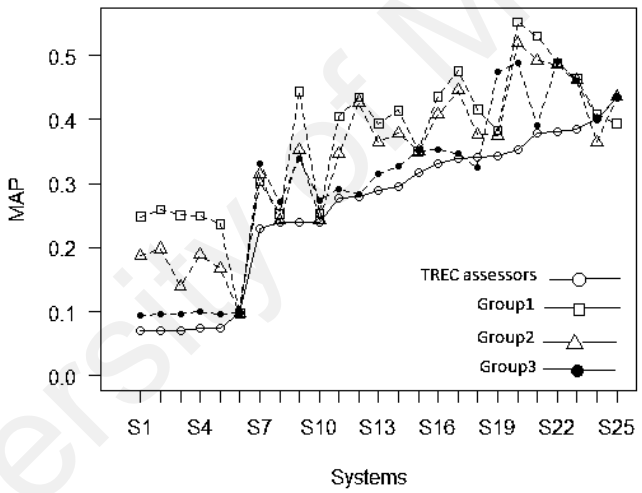


Figure 6.6: System rankings for groups; MAP (*k*=10). The systems are sorted in ascending order of MAP (*k*=10) scores using TREC assessors judgments.

Table 6.13: Kendall's tau correlation

Workers	Kendall's tau	
	MAP (*k*=1000)	MAP (*k*=10)
All workers	0.58	0.64
Group 1 (low scores)	0.54	0.52
Group 2 (moderate scores)	0.61	0.66
Group 3 (high scores)	0.81	0.75

Based on (Landis & Koch, 1977), there are high correlations between system rankings created by TREC assessors on one hand and the system rankings made by all of the workers, Group 1, Group 2 and Group3 on the other hand for both MAP (10) and MAP (1000). Besides, the Kendall's tau correlation coefficients between Group 3 and the TREC assessors is the highest where MAP (1000) and MAP (10) are 0.81 and 0.75, respectively, as it is also noticeable in Figure 6.5 and 6.6. The Kendall's tau correlation coefficients between Group 2 and the TREC assessors is 0.61 for MAP (1000) and 0.66 for MAP(k=10), which is slightly higher than that of for Group 1 (0.54 for MAP (1000) and 0.52 for MAP(k=10)). To sum up, the system rankings based on relevance judgments made by workers with higher logical reasoning skill were relatively more reliable than the system rankings based on relevance judgments made by workers with lower level of logical reasoning skill due to the higher correlation with system rankings based on TREC assessors judgments.

6.5 Effect of Self-Reported Competence on Accuracy of Judgments

In this section, the effect of various self-reported competence about the workers including confidence in relevance judgments, difficulty of the relevance judgments and knowledge on the topic on crowdsourced judgment reliability was investigated. Table 6.14 presents the ternary and binary accuracy for each level of confidence across the 3945 relevance judgments. The ternary and binary accuracy is increasing as the level of confidence is increasing and it shows that workers who were more confident with their judgments obtained a higher accuracy while less confident workers achiever lower accuracy. A Chi-square test shows a relationship between confidence and ternary accuracy to be significant ($\chi 2 = 117.65$, $p < 0.001$). The relationship between confidence and binary accuracy is also significant for chi-squared test ($\chi 2 = 25.27$, $p < 0.001$) as well.

Moreover, the binary and ternary accuracy are calculated across all the 3945 relevance judgment for each level of difficulty to see the effect of difficulty of judgment on accuracy of relevance judgments (Table 6.14). The results show that while the difficulty of the judgment is increasing, the both ternary and binary accuracy is decreasing. A Chi-square test shows the relationship between difficulty and binary accuracy ($\chi 2 = 95.65$, $p < 0.001$) to be significant as well as between difficulty and ternary accuracy ($\chi 2 = 31.71$, $p < 0.001$). For knowledge on the given topic, the results of binary and ternary accuracy is also shown in Table 6.14, but surprisingly the relationship between knowledge on the topic and accuracy of relevance judgments is not significant under a Chi-square test for independence.

Table 6.14: Self-reported competence and accuracy of judgments

	level	Judgments	Ternary correct judgments	Binary correct judgments	Ternary accuracy	Binary accuracy
Confidence in judgment	1	159	63	85	0.39	0.53
	2	821	349	505	0.42	0.61
	3	1349	622	949	0.46	0.70
	4	1616	852	1285	0.53	0.79
Difficulty of the judgment	1	1682	882	1331	0.52	0.79
	2	1183	555	818	0.47	0.69
	3	889	368	562	0.41	0.63
	4	191	81	113	0.42	0.59
Knowledge on the topic	1	351	700	1268	0.40	0.72
	2	171	318	616	0.37	0.72
	3	211	374	742	0.35	0.70
	4	56	97	198	0.35	0.71

Ratings are based on a 4-point Likert-type scale, confidence in judgment: Not confident 1 2 3 4 Very confident, difficulty of the judgment: Easy 1 2 3 4 Difficult, knowledge on the topic: Minimal 1 2 3 4 Extensive.

In the three experiments (verbal comprehension experiment, general reasoning experiment and logical reasoning experiment), we investigated how various self-reported competence about the workers relate to the reliability of their relevance judgments. The results of the three experiments showed a relationship between confidence in judgment

and reliability of relevance judgments. The workers who feel more confidence in their judgments were more reliable in their relevance judgments. These results were in line with a previously published work (Kinney, Huffman, & Zhai, 2008) that found lack of confidence enhances the possibility of incorrect judgments. In a separate study, Sormunen (2002) showed that the consistency of judgments was associated with the confidence of assessments and if assessors feel that the topics are ambiguous, it leads to inconsistency in assessments. In a study done by Ruthven, Baillie, and Elsweiler (2007), it was showed that the assessors' confidence in relevance assessment, interest in the search query and knowledge about the search query influence the number of documents assessed as relevant. A study conducted by Oyama, Baba, Sakurai, and Kashima (2013) used the confidence score for integration of crowdsourced labels as an aggregation approach and they showed that this approach can improve the accuracy of crowdsourced labels. In fact, the confidence score introduces as a useful information to estimate the quality of workers in different studies. Further, the results of the three experiments showed that there is a relationship between difficulty of judgment and reliability of relevance judgments. The workers who feel the judgements are easy; they were more accurate in their judgments. The results in accord with a previous study which found that the perception of task difficulty is an indicative of workers' performance (Kazai et al., 2013) and a clear drop was found in worker accuracy levels when workers reported the task is challenging. Panos Ipeirotis (2009) conducted experiments in which workers were asked to report the task difficulty and he showed that the reported difficulty was correlated with the probability of a correct answer. Therefore, difficulty of judgments can be introduced as a useful information to estimate the quality of workers.

Surprisingly, there is not any relationship between knowledge on the topic and reliability of relevance judgments. This trend may be in conflict with to what would be generally expected. Our results are consistent with a previous work which found the familiarity with the topic did not influence accuracy of relevance judgment (Kazai et al., 2013) while contrasting previous studies which found that unfamiliarity with task and topic plays an important role in accuracy of relevance judgment (Bailey et al., 2008; Kinney et al., 2008). In a separate study on Question Answering (QA) system to find the relationship between users and their knowledge of the search topic, Al-Maskari and Sanderson (2006) found that accuracy increase with query familiarity; the participants who are more familiar with a topic, they are more accurate in their answers.

There are several possibilities to justify our finding. Firstly, knowledge on the topic was self-reported and it may have induced to show their work in a better light. Secondly, their replies could refer to the workers' attitude and confidence in their tasks. Why workers with high self-reported knowledge are apparently less reliable may be that incompetent workers have an inflated sense of their own knowledge (Behrend, Sharek, Meade, & Wiebe, 2011); or, if self-reported knowledge was accurate, it may be those knowledgeable workers were more opinionated, and for that reason most likely to disagree with the original assessor on the relevance of an article to a topic.

6.6 Effect of Demographics on Accuracy of Judgments

In this study, some demographic information was acquired about the workers, which consists of age, gender, level of education, level of computer experience and level of Internet experience and their country as provided by Crowdflower. The demographics is assessed to find out how various demographics information about the workers is related to the reliability of their relevance judgments. Table 6.15 shows the ternary and binary accuracy for each demographic information, across the 789 HITs. Chi-square tests show

that there is not any relationship between demographic information and judgment reliability. Looking at the demographics, our results of the three experiments (verbal comprehension, general reasoning and logical reasoning) show no connection between demographics and judgment reliability of workers.

The findings for age are relatively in accord with a previous work (Kazai et al., 2012) which found a small correlation between age and accuracy over all data and no significant correlation between age and accuracy in a simple design HIT. This finding for gender supports the previously published work (Kazai et al., 2012) reporting no significant relationship between gender and accuracy of the results over all data. In term of education, the expectation was that more educated workers would be better in creating relevance judgments, however, the finding is in accord with a previous work (Kazai et al., 2012) which found no correlation between accuracy and education. Geographical location of the workers also showed no correlation with the judgment reliability. A previous study (Kazai et al., 2012) found that location has a very strong correlation with accuracy of judgments and the Asian workers had significantly lesser performance than American and European workers. However, as our HITs were limited to the English language countries mostly American and European workers, it is reasonable that no significant difference was found among different countries in their accuracy of relevance judgments in our study.

Table 6.15: Demographics and accuracy of judgments

	level	Judgment	Ternary Correct judgment	Binary Correct judgment	Ternary Accuracy	Binary Accuracy
Age	not yet 20	110	53	79	0.48	0.72
	in my 20's	1215	568	875	0.47	0.72
	in my 30's	1130	517	784	0.46	0.69
	in my 40's	750	368	557	0.49	0.74
	in my 50's	500	257	357	0.51	0.71
	60+ years old	240	123	172	0.51	0.72
Gender	Male	2045	934	1437	0.46	0.70
	Female	1900	952	1387	0.50	0.73
Education	no education	0	0	0	0.00	0.00
	primary school	55	31	40	0.56	0.73
	high school	1785	829	1256	0.46	0.70
	Bachelor degree	1590	775	1147	0.49	0.72
	master degree	465	221	338	0.48	0.73
	PhD or higher	50	30	43	0.60	0.86
Computer Experience	1	10	3	6	0.30	0.60
	2	320	128	213	0.40	0.67
	3	1335	633	928	0.47	0.70
	4	2280	1122	1677	0.49	0.74
Internet Experience	1	15	3	5	0.20	0.33
	2	225	109	158	0.48	0.70
	3	1510	687	1044	0.45	0.69
	4	2195	1087	1617	0.50	0.74
Country	AUS	95	41	62	0.43	0.65
	BHS	0	0	0	0.00	0.00
	CAN	725	349	511	0.48	0.70
	GBR	1200	592	864	0.49	0.72
	IRL	130	56	96	0.43	0.74
	NZL	50	26	33	0.52	0.66
	USA	1745	822	1258	0.47	0.72

6.7 Summary

The results of the logical reasoning experiment was presented in this chapter as well as discussion on the results. It was found that there is a relationship between logical reasoning skill of workers and reliability of relevance judgments and the workers who had a higher level of logical reasoning skill generated more accurate relevance judgments. System rankings, which were based on relevance judgments made by workers with higher level of logical reasoning skill, were relatively more reliable than system rankings, which were based on relevance judgments made by workers who had lower level of logical reasoning skill. In the next chapter, two proposed approaches to improve reliability of relevance judgments are explained and discussed.

CHAPTER 7: FILTERING AND AGGREGATION APPROACHES

In this chapter, two approaches are proposed for improving the reliability of relevance judgments generated through crowdsourcing provided by the three experiments (verbal comprehension, general reasoning and logical reasoning). As discussed before, the main objective of this study is to assess the reliability of crowdsourced relevance judgments according to the differences in cognitive abilities of workers. Our study shows that individual difference in cognitive abilities of crowdsourced workers are associated with the level of reliability of their relevance judgments. According to (Li et al., 2014), if certain features and characteristics of workers are associated with their quality, these characteristics can be considered to estimate the quality of their outcomes. In order to introduce an applicable solution to enhance the level of reliability of relevance judgment, in this chapter a filtering approach to filter out low quality judgments, and a judgment aggregation approach to effectively compute consensus judgment from individual judgments are discussed in details in this chapter.

The reliability of the proposed filtering approach was examined by comparing level of agreement for relevance judgments between filtered workers and the TREC assessors on one hand with that of between all workers (without filtering) and the TREC assessors on the other hand. Similarly, the reliability of the proposed filtering approach was assessed for system rankings by comparing system rankings obtained using relevance judgments made by filtered workers (filtered by using the proposed filtering approach) with system rankings obtained using relevance judgments made by all workers (without filtering). The reliability of the proposed judgment aggregation approach was evaluated by comparing the agreement between relevance judgments made by workers using the proposed aggregation approach and relevance judgments provided by TREC assessors on one hand with that of between relevance judgments made by workers using MV method

and TREC assessor's judgments on the other hand. The reliability of the judgment aggregation approach was evaluated for system rankings by comparing the judgment aggregation approach with MV method.

7.1 Filtering Approach

There is a relationship between the selected cognitive abilities (verbal comprehension, general reasoning and logical reasoning) and reliability of relevance judgments as discussed in previous chapters. Thus, workers with lower cognitive skills are less accurate (and reliable) than the others in creating relevance judgments. In fact, cognitive abilities of workers affect the reliability of relevance judgments. In this section, and on the basis of the findings of the three experiments, a filtering approach is proposed to IR practitioners to enhance the quality of relevance judgments in crowdsourcing.

The filtering approach suggests discriminating workers into various groups according to their cognitive abilities and to filter out (or to include) certain group(s) of workers. A requester submits his/her relevance judgment task at the crowdsourcing platform. The task is also comprises a test to evaluate cognitive abilities. A typical task provides an instruction about the task, and explains about the procedure that a worker should perform a cognitive abilities test prior to the actual relevance judgments task. The workers who achieve acceptable score in the cognitive abilities test may proceed to the next step, performing a relevance judgment task, otherwise they are not allowed to proceed and accomplish a relevance judgment task. This filtering approach is adjustable and the acceptable score in the cognitive abilities test can be changed. In this study, we suggests to recruit workers with high cognitive abilities by filtering out those whose cognitive abilities are low and/or moderate.

7.1.1 Reliability of Filtering Approach

The filtering approach was assessed for each experiment (verbal comprehension, general reasoning and logical reasoning experiment), separately. Individual and group agreements for relevance judgments between workers with high verbal comprehension skill (Group 3) and the TREC assessors were compared with that of agreements between all workers (without filtering workers with low and moderate verbal comprehension skill) and the TREC assessors (Table 7.1). In Chapter 4, differences between different groups in reliability of relevance judgments are discussed. In this Chapter, proceeding the filtering approach, the difference between workers with high cognitive abilities and all of the workers was examined for the level of reliability in relevance judgments.

In the general reasoning experiment, the workers with low and moderate general reasoning skill were filtered out and then the agreement for relevance judgments between Group 3 (high general reasoning score) and the TREC assessors were compared with the agreement between all workers and the TREC assessors for relevance judgments (Table 7.1). Similarly, in logical reasoning experiment, the agreement for relevance judgments between workers in Group 3 (high logical reasoning score) and the TREC assessors were compared to the agreement between all workers and the TREC assessors as summarized in Table 7.1.

Table 7.1: Agreement (workers and TREC assessors)

Experiment	Workers	Kappa (ternary)	Kappa (binary)	Kappa (group)
Verbal comprehension experiment	All workers	0.17	0.33	0.42
	Proposed filtering approach (using high score workers)	0.30	0.57	0.66
General reasoning experiment	All of the workers	0.21	0.34	0.33
	Proposed filtering approach (using high score workers)	0.28	0.49	0.53
Logical reasoning experiment	All of the workers	0.21	0.30	0.35
	Proposed filtering approach (using high score workers)	0.27	0.47	0.60

Applying the filtering approach for verbal comprehension experiment, ternary agreements for relevance judgments between workers and the TREC assessors is fair (0.30) but ternary agreements between all workers and the TREC assessors is a slight agreement (0.17). Using the filtering approach, binary agreement for relevance judgments between workers and the TREC assessors is moderate (0.57) while that of between all workers and the TREC assessors is a fair agreement (0.33). Similarly, group agreement can highlight these differences when we applied the filtering approach, showing a substantial agreement (0.66) between relevance judgments set made by workers and relevance judgments set provided by TREC assessors while it is a moderate agreement (0.42) between all workers and the TREC assessors.

Proceeding the application of filtering approach for general reasoning experiment, ternary agreement and binary agreement for relevance judgments between filtered workers and the TREC assessors are higher than those of for relevance judgments between all workers and the TREC assessors. The filtering approach, filters out those workers whose general reasoning skills are either low or moderate, as a result group agreement for relevance judgments between workers and the TREC assessors is moderate (0.53), higher than that of between all workers and the TREC assessors which shows a fair agreement (0.33).

When the filtering approach is applied to the logical reasoning experiment, group agreement for relevance judgments between workers and the TREC assessors is substantial (0.60) while it appears a fair agreement (0.35) between all workers and the TREC assessors. Similarly, ternary agreement and binary agreement between filtered workers (using the filtering approach) and the TREC assessors are higher than the agreements between all workers and the TREC assessors.

Our results for the three experiments show that filtered outcomes have a higher level of agreement with the TREC assessors than that of between all workers and TREC assessors. The proposed filtering approach appears a suitable and adjustable filtering approach (as we applied against workers with low and moderate cognitive ability) to improve the reliability of relevance judgments.

7.1.2 Reliability of Filtering Approach in System Rankings

In this Section, the system rankings for relevance judgments made by all of the workers are compared with those of made by workers with high cognitive abilities using the proposed filtering approach. Through this procedure, we can understand whether the filtering approach improves system rankings or not. Regarding the verbal comprehension experiment, system rankings for relevance judgments provided by the TREC assessors, provided by all workers, and provided by the filtered workers whose verbal comprehension scores are high are shown in Figure 7.1 for MAP (k=1000) and in Figure 7.2 for MAP (k=10) . According to Figure 7.1 and 7.2, the system rankings for relevance judgments made by those workers with high verbal comprehension skill, identified by filtering approach is relatively similar to the system rankings provided by TREC assessors, rather than that of produced by all workers without filtering approach.

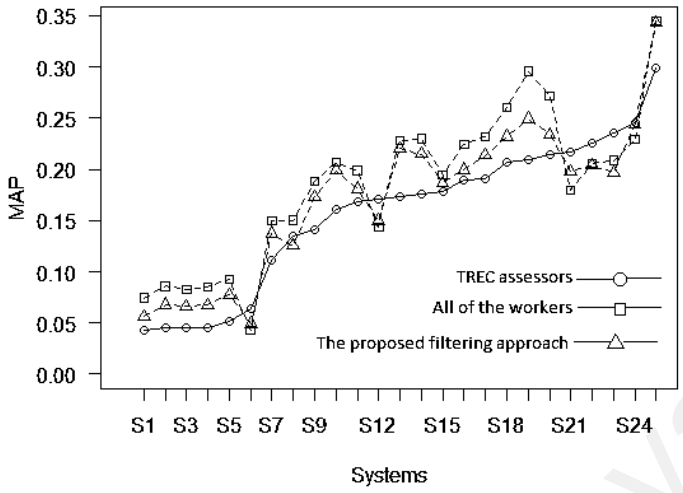


Figure 7.1: System rankings MAP (*k*=1000); verbal comprehension. System rankings for relevance judgments are provided for TREC assessors, all workers and workers with high verbal comprehension scores.

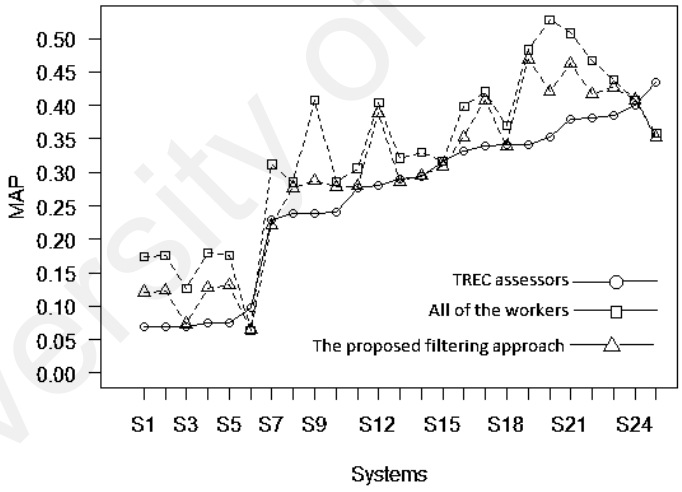


Figure 7.2: System rankings MAP (*k*=10); verbal comprehension. System rankings for relevance judgments are provided for TREC assessors, all workers and workers with high verbal comprehension scores.

In general reasoning experiment, system rankings for relevance judgments provided by TREC assessors, all workers and workers with high general reasoning scores using the filtering approach are shown in Figure 7.3 for MAP (k=1000) and Figure 7.4 for MAP (k=10). For logical reasoning experiment, the comparison between the system rankings for relevance judgments made by the TREC assessors, high logical reasoning skill workers (who identified by the filtering approach) and by all of the workers are shown in Figure 7.5 for MAP (k=1000) and in Figure 7.6 for MAP (k=10). The system rankings for relevance judgments created by workers with high logical reasoning skill as assigned by filtering approach is relatively more similar to the system rankings produced by the TREC assessors than that of by all workers (without using the filtering approach).

In this study, Kendall's tau correlation was applied to find out to what extend the filtering approach improves system rankings. As such, the system rankings for relevance judgments generated by all workers are compared with that of created by the filtered workers. Each of these system rankings is compared with the system ranking generated by the TREC assessors and respective Kendall' tau correlations with TREC assessors are shown in Table 7.2.

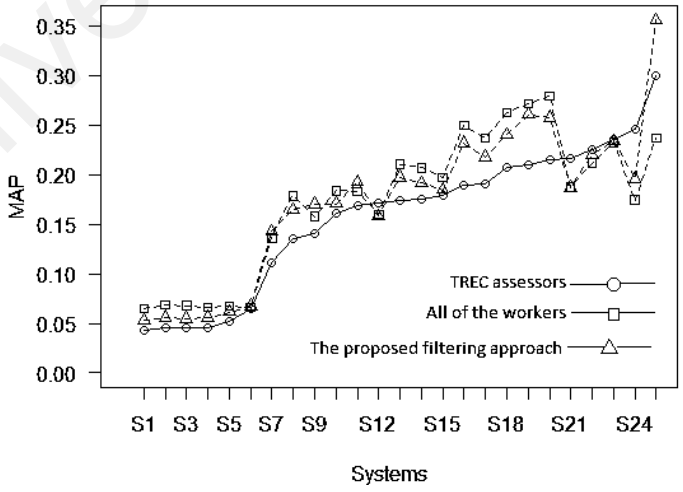


Figure 7.3: System rankings MAP (k=1000); general reasoning. System rankings for relevance judgments are provided for TREC assessors, all workers and workers with high general reasoning scores.

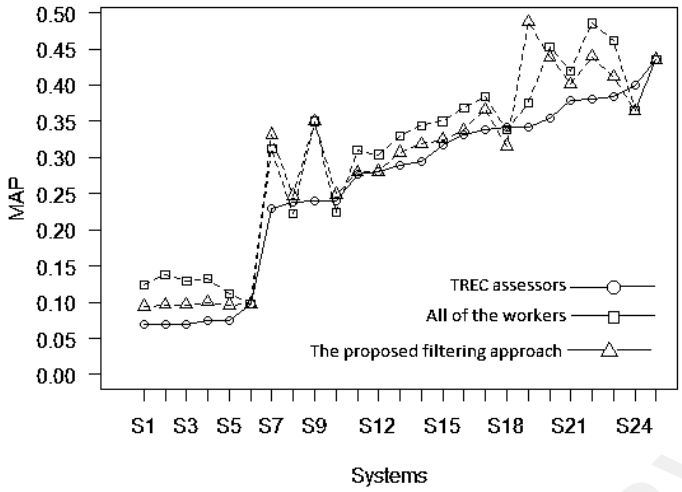


Figure 7.4: System rankings MAP (*k*=10); general reasoning. System rankings for relevance judgments are provided for TREC assessors, all workers and workers with high general reasoning scores.

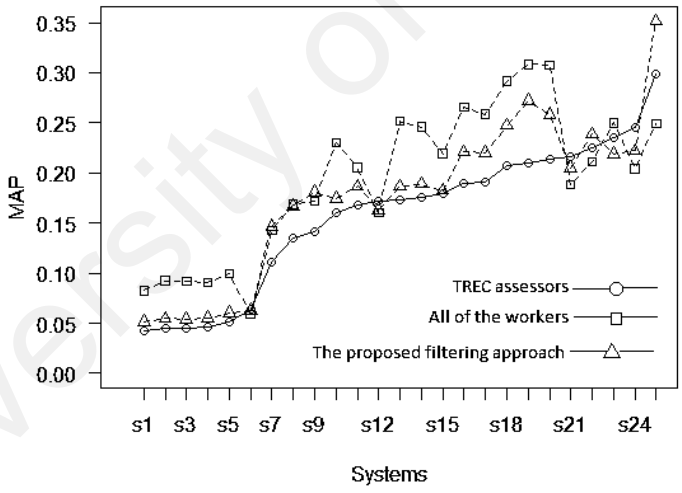


Figure 7.5: System rankings MAP (*k*=1000); logical reasoning. System rankings for relevance judgments are provided for TREC assessors, all workers and workers with high logical reasoning scores.

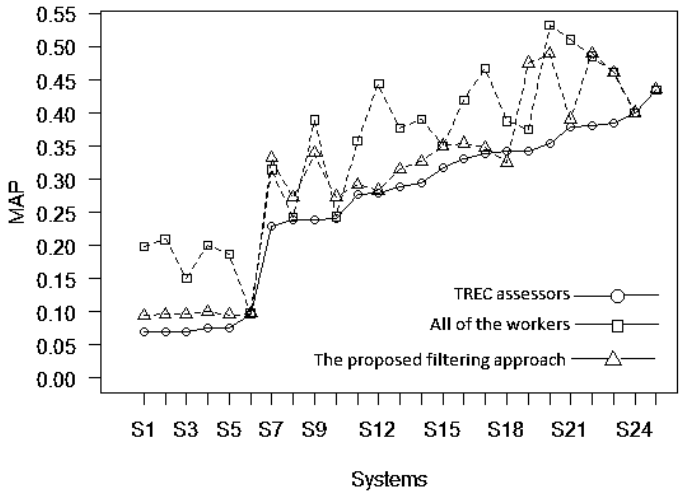


Figure 7.6: System rankings MAP (*k*=10); logical reasoning. System rankings for relevance judgments are provided for TREC assessors, all workers and workers with high logical reasoning scores.

Table 7.2: Kendall's tau correlation (workers and TREC assessors)

Experiment	Workers	Kendall's tau	
		MAP(*k*=1000)	MAP(*k*=10)
Verbal comprehension experiment	All workers	0.66	0.65
	Proposed filtering approach (using high score workers)	0.69	0.75
General reasoning experiment	All of the workers	0.64	0.74
	Proposed filtering approach (using high score workers)	0.76	0.73
Logical reasoning experiment	All of the workers	0.58	0.64
	Proposed filtering approach (using high score workers)	0.81	0.75

In verbal comprehension experiment, Kendall's tau correlations for system rankings between relevance judgments produced by filtered workers (high scores workers) and that of reported by the TREC assessors are higher for MAP (*k*=1000)=0.69 and MAP (*k*=10)=0.75, than the correlations between all workers and the TREC assessors for MAP (*k*=1000)=0.66 and MAP (*k*=10)=0.65. Therefore, the proposed filtering approach relatively improves system rankings. The use of proposed filtering approach in general reasoning experiment is not significantly influential as Kendall's tau correlation value for MAP(*k*=10) is (0.73) shows similar correlation value to tau MAP(*k*=10) for all

workers (without using the filtering approach) which is 0.74. Kendall's tau correlation for MAP (*k*=1000) between filtered workers using the proposed approach and the TREC assessors is 0.76, which is higher than that of between all workers and the TREC assessors judgments for a value of 0.64. According to the provided results in Table 7.2 for logical reasoning experiment, Kendall's tau correlation for system rankings for MAP (*k*=10) and MAP (*k*=1000), between filtered workers (using the proposed approach) and the TREC assessors is higher 0.75 and 0.81 respectively, as compared with Kendall's tau correlation between all workers and the TREC assessors (0.64 and 0.58, respectively). Based on the results provided by the three experiments, the proposed filtering approach can enhance system rankings.

7.2 Judgment Aggregation Approach

The three experiments show that the cognitive abilities of the workers affect the reliability of relevance judgments. Workers with higher level of cognitive abilities are more reliable for creating relevance judgments. The proposed judgment aggregation approach suggests to weight judgments provided by the workers based on their cognitive abilities. According to our study, cognitive abilities of the workers is associated with the quality of their judgments. Therefore, relevance judgments created by high cognitive ability worker acquire higher weights in the aggregating judgments. We define three steps for judgment aggregation approach as followed. Supposing there are *n* topics and *n* documents, therefore;

Step 1: Sum of the scores for the cognitive abilities of workers who judge a given topic and a given document as "relevant" is;

$$Sum\ of\ cognitive\ abilities\ Scores\ "vote\ relevant" = \sum_{1}^{n*n} RCAScore \quad (7.1)$$

Step 2: Sum of the scores for the cognitive abilities of workers who judge the topic and document as "non-relevant" is;

$$Sum\ of\ cognitive\ abilitis\ Scores\ vote\ "non-relevant" = \sum_{1}^{n*n} NRCAScore \quad (7.2)$$

Step 3: Comparing the sum of the cognitive abilities scores for "relevant" (Equation 7.1) and "non-relevant" judgements (Equation 7.2). If sum of cognitive ability scores vote "relevant" is higher than sum of cognitive ability scores vote "non-relevant", the topic and document is considered "relevant", otherwise the topic and document is "non-relevant".

Figure 7.7 gives an example for the judgment aggregation approach. Five workers judged the topic and document <100, 2>, three workers judged the topic and document as "non-relevant" (W1, W2 and W5) and the other two workers judged the topic and document as "relevant" (W3 and W4). "Cognitive ability scores" column shows the score that each worker gained for cognitive abilities, such as verbal comprehension scores. According to the three steps of the judgment aggregation approach, the sum of cognitive ability scores of the workers who judged the topic and the document as "relevant" (step 1) is compared with the sum of cognitive skill scores of the workers who considered the topic and the document "not-relevant" (step 2). Because the sum of the first step is greater than that of the second step, the topic and document are assigned "relevant" according to the proposed judgment aggregation approach. Figure 7.7 also provides the result for MV method, a common method used for aggregating judgments. Using MV method, since majority of workers judged the topic and document as "non-relevant", the topic and document is considered "non-relevant".

7.2.1 Reliability of Judgment Aggregation Approach

The judgment aggregation approach was tested for every experiment. Group agreement between relevance judgments made by workers aggregated by MV method and relevance judgments provided by TREC is compared with that of aggregated by the proposed judgment aggregation approach as shown in Table 7.3.

Topic	Doc	Worker ID	Cognitive ability score	Binary worker Judgment	MV method
100	2	W1	10	NR	NR
100	2	W2	12	NR	
100	2	W3	16	R	
100	2	W4	18	R	
100	2	W5	10	NR	

Step1: Sum of cognitive ability score vote relevant= 16+18=34
Step2: Sum of cognitive ability score vote not-relevant= 10+12+10=32
Step3: 34>32 ➔ Relevant

Figure 7.7: Example of the proposed judgment aggregation approach

Table 7.3: Group agreement (workers and TREC assessors)

Experiments	Aggregating method	Percentage Agreement	Kappa
Verbal comprehension experiment	MV method	77%	0.42
	Proposed judgment aggregation approach	79%	0.47
General reasoning experiment	MV method	74%	0.33
	Proposed judgment aggregation approach	77%	0.41
Logical reasoning experiment	MV method	75%	0.35
	Proposed judgment aggregation approach	79%	0.47

The results of verbal comprehension experiment show that the proposed judgment aggregation approach outperforms MV method by 2%. Using the proposed judgment aggregation approach, group agreement for relevance judgments between all workers and the TREC assessors is 79% for a kappa value of 0.47 while the group agreement using MV method is 77% with kappa value of 0.42. Similarly, in general reasoning experiment, the proposed judgment aggregation approach outperforms the MV method. Group agreement for relevance judgments between all workers and TREC assessor is 77% for the proposed judgment aggregation approach and 74% for MV method. In the logical reasoning experiment, group agreement for relevance judgments between all workers and the TREC assessors shows 79% for the proposed judgment aggregation approach higher than that of MV method for 75%.

7.2.2 Judgment Aggregation Approach in System Rankings

Comparing the proposed judgment aggregation approach with MV method in system rankings, in this section, the impact of the proposed judgment aggregation approach on system rankings is discussed. We also examine whether the proposed judgment aggregation approach provides similar system rankings to that of produced by the TREC assessors. This comparison was performed for each of the three experiments, including three system rankings as followed:

i. System rankings derived using relevance judgments provided by the TREC assessors.
ii. System rankings obtained using relevance judgments made by all workers using MV method for judgment aggregation.
iii. System rankings obtained using relevance judgments made by all workers using the proposed judgment aggregation approach.

In verbal comprehension experiment, system rankings comparisons for MAP (k=1000) and for MAP (k=10) are illustrated in Figure 7.8 and Figure 7.9. The system rankings using the proposed judgment aggregation approach is relatively similar to that of using MV method. As it is presented in Figure 7.10 for MAP (k=1000) and Figure 7.11 for MAP (k=10), in general reasoning experiment, system rankings comparison between MV method and the proposed judgment aggregation approach are quite comparable. Figure 7.12 and Figure 7.13 demonstrate system rankings comparisons (MAP (k=1000) and MAP (k=10), respectively), for the logical reasoning experiment. System rankings generated using workers' relevance judgments aggregated by the proposed judgment aggregation approach is relatively more similar to that of provided by TREC assessments than system rankings generated using workers' relevance judgments aggregated by MV method.

Table 7.4 summarizes differences between system rankings of all workers (using either MV method or the proposed judgment aggregation approach) and the TREC assessors providing Kendall's tau correlation for MAP (k=1000) and MAP (k=10).

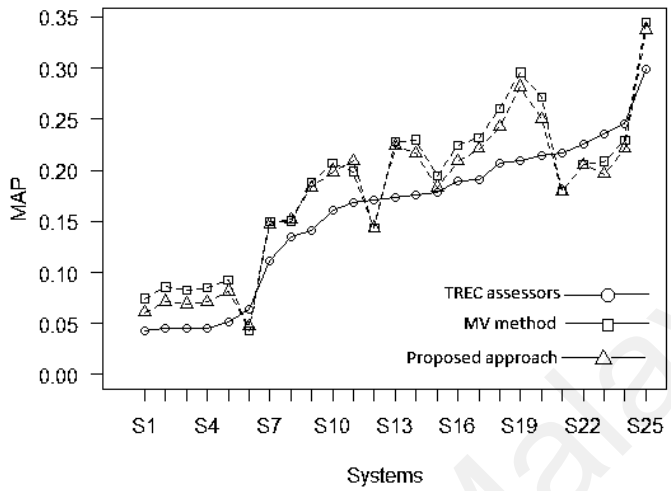


Figure 7.8: System rankings MAP (k=1000); verbal comprehension. Comparison of system rankings between MV method and the proposed judgment aggregation approach.

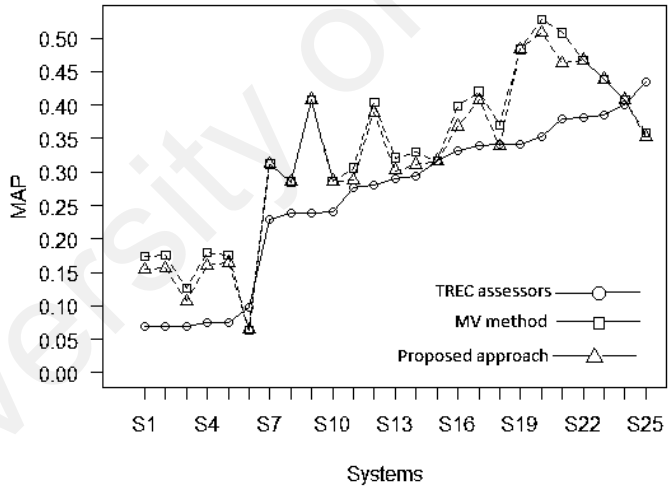


Figure 7.9: System rankings MAP (k=10); verbal comprehension. Comparison of system rankings between MV method and the proposed judgment aggregation approach.

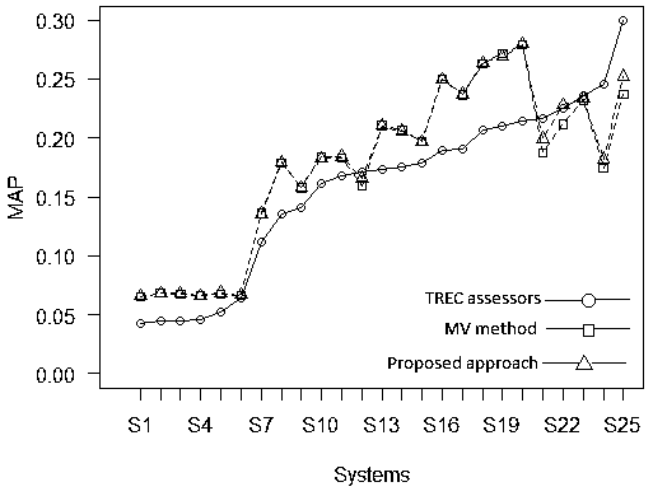


Figure 7.10: System rankings MAP (*k*=1000); general reasoning. Comparison of system rankings between MV method and the proposed judgment aggregation approach.

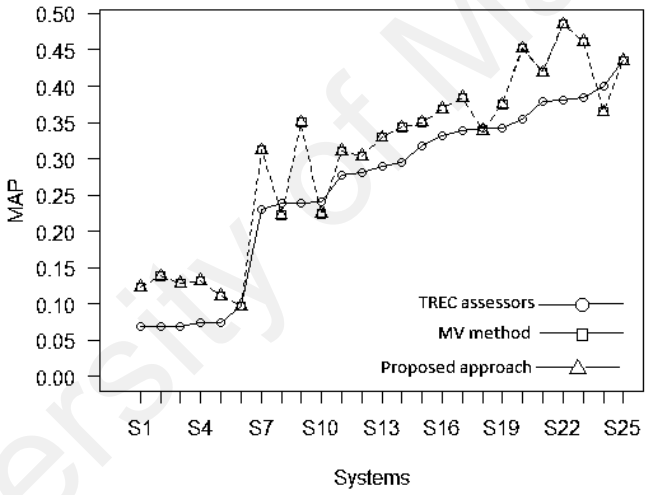


Figure 7.11: System rankings MAP (k=10); general reasoning. Comparison of system rankings between MV method and the proposed judgment aggregation approach.

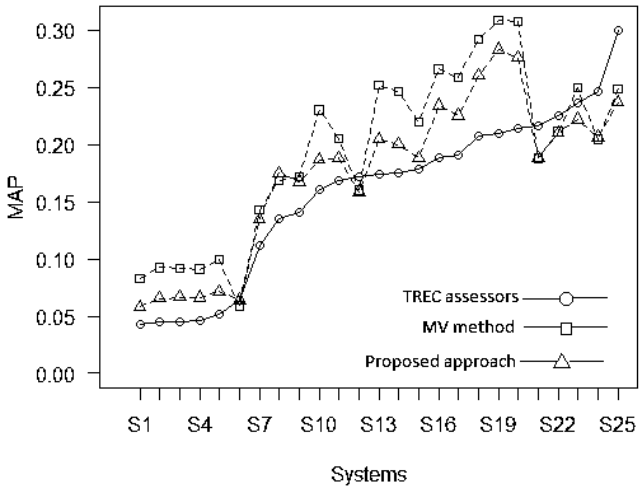


Figure 7.12: System rankings MAP (k=1000); logical reasoning. Comparison of system rankings between MV method and the proposed judgment aggregation approach.

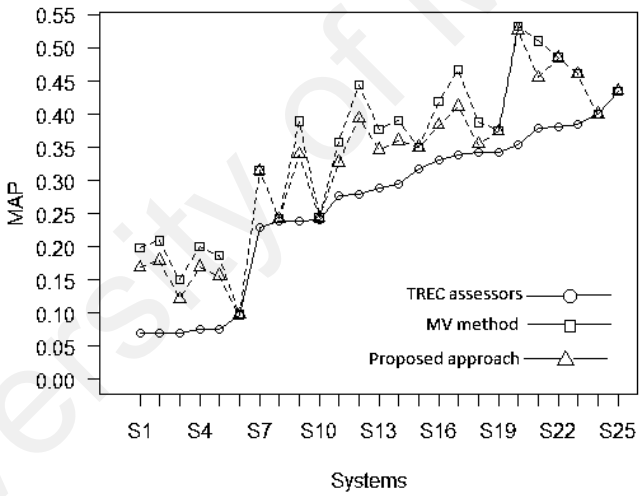


Figure 7.13: System rankings MAP (k=10); logical reasoning. Comparison of system rankings between MV method and the proposed judgment aggregation approach.

Table 7.4: Kendall's tau correlation (workers and TREC assessors)

Experiments	Aggregating	Kendall's tau	
		MAP(*k*=1000)	MAP(*k*=10)
Verbal comprehension experiment	MV method	0.66	0.65
	Proposed judgment aggregation approach	0.63	0.66
General reasoning experiment	MV method	0.64	0.74
	Proposed judgment aggregation approach	0.69	0.74
Logical reasoning experiment	MV method	0.58	0.64
	Proposed judgment aggregation approach	0.72	0.74

In verbal comprehension experiment, Kendall's tau correlation (Table 7.4) between system rankings derived using relevance judgments made by workers (using MV method) and system rankings derived using TREC assessors judgments is 0.66 for MAP (*k*=1000) and 0.65 for MAP (*k*=10), which are relatively similar to that of between workers (using proposed judgment aggregation approach) and the TREC assessors showing 0.63 and 0.66 for MAP(*k*=1000) and MAP(*k*=10), respectively (Figure 7.8 and 7.9).

In general reasoning experiment, Kendall's tau correlation between system rankings derived using relevance judgments made by workers (using MV method for judgment aggregation) and system rankings derived using TREC assessors judgments for MAP (*k*=10) is 0.74, which is exactly as the same as that of for MAP (*k*=10) between workers (using proposed judgment aggregation approach) and the TREC assessors (Figure 7.11). For MAP (*k*=1000), the correlation between system rankings derived using relevance judgments made by workers (using proposed judgment aggregation approach) and system rankings derived using TREC assessors judgments is 0.69, which is a little higher than that of seen between workers (using MV method) and the TREC assessors (0.64).

As illustated in Figure 7.12 and Figure 7.13, MAP (k=1000) for 0.72 and MAP (k=10) for 0.74 are Kendall's tau correlations between system rankings derived using relevance judgments made by workers (using proposed judgment aggregation approach) and system rankings derived using TREC assessors judgments. These values are relatively lower than the correlation between workers (using MV method) and the TREC assessors which are 0.58 and 0.64, respectively (Table 7.4). Overall, using the judgment aggregating approach in relevance judgment set produces relatively similar system rankings to that of produced by MV method for aggregation.

7.3 Summary

Findings from the three experiments (as discussed in previous chapters) provide scientific evidences to propose the two approaches, filtering approach and judgment aggregation approach, which can contribute in improving the reliability of relevance judgments in crowdsourcing. The two approaches may provide means to IR practitioners to utilize cognitive abilities of workers to enhance the quality of their outcomes. Practical uses of these approaches need further investigation. The next chapter provides a conclusion of this research study, which includes a briefing about some implications of the findings of this research providing further suggestions for future work.

CHAPTER 8: CONCLUSION

Test collections are applied to evaluate IR techniques in system-based retrieval evaluation. A main pitfall for test collections is its cost to create relevance assessments, which are usually conducted by human expert assessors. Crowdsourcing provides an affordable platform to produce relevance judgment sets. Nevertheless, the quality of outputs from crowdsourcing needs to be assessed precisely. According to our results from the three experiments named verbal comprehension experiment, general reasoning experiment and logical reasoning experiment, this work provides a number of contributions. Our results show the impact of workers' cognitive abilities on reliability of relevance judgments in crowdsourcing, highlighting an association between cognitive abilities and reliability of relevance judgments, *i.e.* the higher cognitive abilities a worker has the more reliable judgments can be produced. This association convey an idea to propose two approaches for improving the reliability of relevance judgments: a filtering approach and a judgment aggregation approach. This chapter provides a summary and conclusion about the research study presented through the course of this thesis. Furthermore, a section is assigned to present limitations of this work providing some suggestions for future Work.

8.1 Significance of the Study

The main hypothesis of this study is that crowdsourced workers with different levels of cognitive abilities have a positive correlation with their reliability of relevance judgments. We also hypothesized that when the assessments of workers with higher cognitive abilities are used to evaluate and rank retrieval systems by effectiveness, they provide a similar ranking to that of the TREC expert assessments. Verbal comprehension skill is a cognitive ability, which has an effect on reliability of relevance judgments as shown in our study. The workers with higher level of verbal comprehension skill appear more accurate in creating relevance judgments. Agreement with the relevance judgments

is considered to evaluated accuracy as it is compared with the official TREC assessors (gold standard dataset). Similarly, when the assessments of workers with high verbal comprehension skill are used to evaluate and rank retrieval systems by effectiveness, they give a ranking more similar to that of the official TREC assessments than when the assessments of workers with low and moderate verbal comprehension are so used. General reasoning skill is of cognitive abilities. Our study shows that, this skill is associated with reliability of relevance judgments. In the same way, when the assessments of workers with high general reasoning skill are used to evaluate and rank retrieval systems by effectiveness, they give a ranking more similar to that of the official TREC assessments than when the assessments of workers with low and moderate general reasoning skill are so used.

Association between logical reasoning skill and reliability of relevance judgments shows that workers who have higher logical reasoning skills are more accurate in their relevance judgments as compared with the other groups. Moreover, when the relevance judgments set made by workers with high logical reasoning skill are used to rank retrieval systems by effectiveness, they give a ranking more similar to that of the official TREC assessments than when the relevance judgments of workers with low and moderate logical reasoning skill are so used.

Based on the findings of this study, and in order to improve the reliability of crowdsourced relevance judgments, two approaches were proposed, filtering approach and judgment aggregation approach. Former approach is a filtering technique for recruiting workers in crowdsourcing. This approach provides a possibility for the requesters to discriminate workers into various groups according to their cognitive abilities and to filter out (or to include) certain group(s) of workers. For instance, in our study, workers with high scores in the test were considered to accomplish the relevance

judgment task by filtering workers with either low or moderate scores out. The intention behind this approach is that workers with higher cognitive abilities are probably more accurate for making relevance judgments.

We examined this approach with statistical test to find out the level of agreement for relevance judgments between filtered workers and the TREC assessors on one hand and that of between all workers (without filtering) and the TREC assessors on the other hand. As a result, the filtering approach improves the reliability of relevance judgments whilst it has a minor effect on system rankings for the three experiments.

In crowdsourcing, the level of cognitive abilities of workers may help to estimate the reliability of crowdsourced relevance judgments. Therefore, the second approach was proposed, judgment aggregation approach for integration process. Judgment aggregation approach is to consider cognitive abilities of workers during aggregating process. In this study, results from judgment aggregation approach were compared to MV results (a common method of judgment aggregation). One possible weakness of MV method is that it computes consensus by equally weighting each worker's judgment and the assumption is all workers are equally good; however, worker's qualities may be dynamic over time. The proposed judgment aggregation approach outperforms the MV method in each of the three experiments. Applying the proposed judgment aggregation approach, the level of agreement for relevance judgment between workers and TREC assessors is higher than when MV method is used. However, the effect of this approach on system rankings for the three experiments are heterogeneous. In the logical reasoning experiment, the judgment aggregation approach relatively improves the system rankings but in the general reasoning experiment and the verbal comprehension experiment, the judgment aggregation approach and MV method are quite similar. Overall, using the judgment

aggregating approach in relevance judgment set produces relatively similar system rankings to that of produced by MV method for aggregation.

This study highlights the importance of deeming cognitive characteristics as imperative factors during a relevance judgment process in order to produce outcomes that are more reliable. This work has implications for IR evaluation design through crowdsourcing. Showing the association between cognitive abilities and the reliability of relevance judgments in crowdsourcing, IR practitioners are suggested to consider these elements in designing IR evaluation experimentations.

8.2 Limitations and Future Work

In this work, several issues were addressed in the context of crowdsourced IR evaluation; however, there are several interesting topics in the field that need further investigations. The effects of cognitive abilities on reliability of relevance judgments were investigated, assessing three cognitive abilities. However, there are other cognitive abilities as well as cognitive style that are suggested to be included in future experimental designs. Moreover, there are other tests to evaluate cognitive abilities, which are suggested to be applied in future work. First, this is to evaluate the outcomes of the other test and second, to compare them with our findings in this work. We used the FRCT in this study, but it is important to assess the consistency of our findings through other tests. In this work, two approaches are proposed to improve the reliability of relevance judgments, tested by the three selected cognitive abilities. Further assessments are required to investigate the effectiveness of the proposed approaches using various cognitive abilities. It seems an interesting topic if researchers include a wider range of factors in task design and assess the impacts of various psychological factors such as emotion and other personality traits on reliability of relevance judgments for their future work. This assessment will help categorize the crowdsourced workers and to investigate

their effects on crowdsourcing outcome. The scalability of crowdsourcing for large-scale IR evaluation is a thrilling area of research that is highly recommended for future assessments.

The output of crowdsourcing is often noisy. Therefore, quality control methods are suggested especially in designing tasks and/or after completion of tasks. As discussed before in detail, workers' characteristics and behaviors have a great impact on the quality of crowdsourcing outcome. Therefore, a comprehensive understanding about workers' behavior will enhance the reliability of crowdsourcing output.

The interesting results of relationship between self-reported competence (confidence and difficulty) and reliability of relevance judgments motivate the need for further investigation to utilize this competence in filtering and judgment aggregation approaches. In fact, the self-reported competence (confidence and difficulty) can be introduced as a useful information to estimate the quality of workers. Lastly, assessments about the relationship between cognitive abilities and crowdsourcing outcome for different tasks are suggested for future studies to find out whether that association is stable in various types of tasks.

REFERENCES

Al-Maskari, A., & Sanderson, M. (2006). *The Effects of Topic Familiarity on User Search Behavior in Question Answering Systems.* Paper presented at the Proceedings of the LWA 2006 Workshop.

Al-Maskari, A., & Sanderson, M. (2011). The Effect of User Characteristics on Search Effectiveness in Information Retrieval. *Information Processing & Management,* 47(5), 719-729.

Al-Maskari, A., Sanderson, M., & Clough, P. (2008). *Relevance Judgments Between Trec and Non-Trec Assessors.* Paper presented at the Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Al-Maskari, A., & Sanderson, M. (2010). A Review of Factors Influencing User Satisfaction in Information Retrieval. *Journal of the American Society for Information Science & Technology,* 61(5), 859-868.

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

Allen, B. (1992). *Cognitive Differences in End User Searching of a CD-ROM Index.* Paper presented at the Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Allen, B. (1994a). Cognitive Abilities and Information System Usability. *Information Processing & Management,* 30(2), 177-191.

Allen, B. (1994b). *Perceptual Speed, Learning and Information Retrieval Performance.* Paper presented at the SIGIR'94.

Allen, B., & Allen, G. (1993). Cognitive Abilities of Academic Librarians and their Patrons. *College & research libraries,* 54(1), 67-73.

Alonso, O. (2012). Implementing Crowdsourcing-Based Relevance Experimentation: an Industrial Perspective. *Information Retrieval*, 1-20. doi:10.1007/s10791-012-9204-1

Alonso, O., & Baeza-Yates, R. (2011). Design and Implementation of Relevance Assessments Using Crowdsourcing *Advances in Information Retrieval* (pp. 153-164): Springer.

Alonso, O., & Mizzaro, S. (2009). *Can We Get Rid of TREC Assessors? Using Mechanical Turk for Relevance Assessment.* Paper presented at the Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation.

Alonso, O., & Mizzaro, S. (2012). Using Crowdsourcing for TREC RELEVANCE ASSESSMENT. *Information Processing & Management,* 48(6), 1053-1066. doi:http://dx.doi.org/10.1016/j.ipm.2012.01.004

Alonso, O., Rose, D. E., & Stewart, B. (2008). *Crowdsourcing for Relevance Evaluation.* Paper presented at the ACM SIGIR Forum.

Ambati, V., Vogel, S., & Carbonell, J. (2010). Active Learning and Crowd-Sourcing for Machine Translation. *Language Resources and Evaluation (LREC),* 7, 2169-2174.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search*: Addison Wesley Professional.

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., & Yilmaz, E. (2008). *Relevance Assessment: are Judges Exchangeable and Does It Matter.* Paper presented at the Proceedings of the 31st Annual International ACM SIGIR Conference on Research And Development in Information Retrieval.

Barhydt, G. C. (1964). *A Comparison of Relevance Assessments by Three Types of Evaluator.* Paper presented at the Proceedings of the American Documentation Institute.

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The Roles of Associative and Executive Processes in Creative Cognition. *Memory & cognition,* 42(7), 1186-1197.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The Viability of Crowdsourcing for Survey Research. *Behavior research methods,* 43(3), 800-813.

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., & Tran Duc, T. (2011). *Repeatable and reliable search system evaluation using crowdsourcing.* Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., & Tran, T. (2013). Repeatable and reliable semantic search evaluation. *Web Semantics: Science, Services and Agents on the World Wide Web,* 21, 14-29.

Borgatta, E. F., & Corsini, R. J. (1964). Quick word test manual. *New York*.

Borgman, C. L. (1989). All users of information retrieval systems are not created equal: An exploration into individual differences. *Information Processing & Management,* 25(3), 237-251.

Brabham, D. C. (2009). Crowdsourcing the public participation process for planning projects. *Planning Theory,* 8(3), 242-262.

Brennan, K., Kelly, D., & Arguello, J. (2014). *The effect of cognitive abilities on information search for tasks of varying levels of complexity.* Paper presented at the Proceedings of the 5th Information Interaction in Context Symposium.

Brew, A., Greene, D., & Cunningham, P. (2010). *Using crowdsourcing and active learning to track sentiment in online media.* Paper presented at the Proceedings of the 2010 conference on ECAI.

Burgin, R. (1992). Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management,* 28(5), 619-627.

Callison-Burch, C. (2009). *Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.

Carpenter, B. (2008). Multilevel bayesian models of categorical data annotation. *Unpublished manuscript*.

Carroll, J. B. (1974). *Psychometric Tests as Cognitive Tasks: A New'Structure of Intellect'*. Retrieved from

Carterette, B., Bennett, P. N., Chickering, D. M., & Dumais, S. T. (2008). Here or there *Advances in Information Retrieval* (pp. 16-27): Springer.

Carterette, B., & Soboroff, I. (2010). *The effect of assessor error on IR system evaluation.* Paper presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

Carterette, B., & Voorhees, E. M. (2011). Overview of information retrieval evaluation *Current challenges in patent information retrieval* (pp. 69-85): Springer.

Charness, N., Kelley, C. L., Bosman, E. A., & Mottram, M. (2001). Word-processing training and retraining: effects of adult age, experience, and interface. *Psychology and aging,* 16(1), 110.

Choffnes, D. R., Bustamante, F. E., & Ge, Z. (2010). *Crowdsourcing service-level network event monitoring.* Paper presented at the ACM SIGCOMM Computer Communication Review.

Cleverdon, C. (1967). *The Cranfield tests on index language devices.* Paper presented at the Aslib proceedings.

Cleverdon, C. W. (1970). The effect of variations in relevance assessments in comparative experimental tests of index languages.

Clough, P., Sanderson, M., Tang, J., Gollins, T., & Warner, A. (2012). Examining the limits of crowdsourcing for relevance assessment. *Internet Computing, IEEE,* PP(99), 1-1. doi:10.1109/MIC.2012.95

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement,* 20(1), 37-46.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (rev*: Lawrence Erlbaum Associates, Inc.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*: Academic press.

Collins-Thompson, K., Bennett, P., Diaz, F., Clarke, C. L., & Voorhees, E. M. (2014). TREC 2013 Web track overview. *MICHIGAN UNIV ANN ARBOR, Tech. Rep*.

Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and aging,* 21(2), 333.

Czaja, S. J., Sharit, J., Ownby, R., Roth, D. L., & Nair, S. (2001). Examining age differences in performance of a complex information search and retrieval task. *Psychology and aging,* 16(4), 564.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 20-28.

De Alfaro, L., Kulshreshtha, A., Pye, I., & Adler, B. T. (2011). Reputation systems for open collaboration. *Communications of the ACM,* 54(8), 81-87.

Difallah, D. E., Demartini, G., & Cudré-Mauroux, P. (2012). *Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms.* Paper presented at the CrowdSearch 2012 workshop at WWW 2012, Lyon, France.

Dow, S., Kulkarni, A., Bunge, B., Nguyen, T., Klemmer, S., & Hartmann, B. (2011). *Shepherding the crowd: managing and providing feedback to crowd workers.* Paper presented at the Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems.

Eickhoff, C., & de Vries, A. P. (2012). Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 1-17.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. *Princeton, NJ: Educational Testing Service*.

Ford, N., Miller, D., & Moss, N. (2001). The role of individual differences in Internet searching: An empirical study. *Journal of the American Society for Information Science and Technology,* 52(12), 1049-1066. doi:10.1002/asi.1165

Ford, N., Wilson, T., Ellis, D., Foster, A., & Spink, A. (2000). *Individual Differences in Information Seeking: An Empirical Study.* Paper presented at the Proceedings of the ASIS Annual Meeting.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PloS one,* 8(1), e54651.

Goker, A., & Davies, J. (2009). *Information retrieval: Searching in the 21st century*: John Wiley & Sons.

Grady, C., & Lease, M. (2010). *Crowdsourcing document relevance assessment with Mechanical Turk.* Paper presented at the Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk.

Hammon, L., & Hippner, H. (2012). Crowdsourcing. *WIRTSCHAFTSINFORMATIK,* 54(3), 165-168. doi:10.1007/s11576-012-0321-7

Harter, S. P. (1990). Search term combinations and retrieval overlap: A proposed methodology and case study. *Journal of the American Society for Information Science,* 41(2), 132-146.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science (1986-1998),* 43(9), 602.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *JASIS,* 47(1), 37-49.

Heer, J., & Bostock, M. (2010). *Crowdsourcing graphical perception: using mechanical turk to assess visualization design.* Paper presented at the Proceedings of the 28th international conference on Human factors in computing systems.

Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2012). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*(0). doi:http://dx.doi.org/10.1016/j.mcm.2012.01.006

Holley, R. (2009). Crowdsourcing and social engagement: potential, power and freedom for libraries and users.

Hosseini, M., Cox, I. J., Milić-Frayling, N., Kazai, G., & Vinay, V. (2012). On aggregating labels from multiple crowd workers to infer relevance of documents *Advances in Information Retrieval* (pp. 182-194): Springer.

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

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of vocational behavior,* 29(3), 340-362.

Ipeirotis, P. (2009). Turker demographics vs. Internet demographics. *A Computer Scientist in a Business School*.

Ipeirotis, P. (2010). Demographics of mechanical turk. *CeDER-10–01 working paper, New York University*.

Järvelin, K., & Kekäläinen, J. (2000). *IR evaluation methods for retrieving highly relevant documents.* Paper presented at the Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS),* 20(4), 422-446.

Joachims, T., & Radlinski, F. (2007). Search engines that learn from implicit feedback. *Computer*(8), 34-40.

John, O., Naumann, L., & Soto, C. (2008). Paradigm Shift to the Integrative Big Five Trait Taxonomypp. 114—158: Handbook of Personality: Theory and Research, New York: Guilford Press.

Jones, K. S. (1981a). *Information Retrieval Experiment*: Butterworth-Heinemann.

Jones, K. S. (1981b). Retrieval system tests 1958-1978.

Jung, H. J., & Lease, M. (2012). *Improving Quality of Crowdsourced Labels via Probabilistic Matrix Factorization.* Paper presented at the Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

Karahoca, D., Karahoca, A., & Güngör, A. (2008). *Assessing effectiveness of the cognitive abilities and individual differences on e-learning portal usability evaluation.* Paper presented at the Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing.

Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation *Advances in information retrieval* (pp. 165-176): Springer.

Kazai, G. (2014). *Information retrieval evaluation with humans in the loop*. Paper presented at the Proceedings of the 5th Information Interaction in Context Symposium, Regensburg, Germany.

Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011). *Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking.* Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2011). *Worker types and personality traits in crowdsourcing relevance labels.* Paper presented at the Proceedings of the 20th ACM international conference on Information and knowledge management.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2012). *The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy.* Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval,* 16(2), 138-178. doi:10.1007/s10791-012-9205-0

Kazhdan, T. (1979). Effects of subjective expert evaluation of relevance on the performance parameters of document-based information retrieval system. *Nauchno-Tekhnicheskaya Informatsiya, Seriya,* 2(13), 21-24.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika,* 30(1/2), 81-93.

Khanna, S., Ratan, A., Davis, J., & Thies, W. (2010). *Evaluating and improving the usability of Mechanical Turk for low-income workers in India.* Paper presented at the Proceedings of the first ACM symposium on computing for development.

Kim, C. S. (2002). *Predicting information searching performance with measures of cognitive diversity*: University of Wisconsin--Madison.

Kim, K. S., & Allen, B. (2002). Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology,* 53(2), 109-119.

Kinney, K. A., Huffman, S. B., & Zhai, J. (2008). *How evaluator domain expertise affects search result relevance judgments.* Paper presented at the Proceedings of the 17th ACM conference on Information and knowledge management.

Kittur, A., Chi, E. H., & Suh, B. (2008). *Crowdsourcing user studies with Mechanical Turk.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Kulkarni, A., Can, M., & Hartmann, B. (2012). *Collaboratively crowdsourcing workflows with turkomatic.* Paper presented at the Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics,* 33(1), 159-174.

Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010). *Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution.* Paper presented at the SIGIR 2010 workshop on crowdsourcing for search evaluation.

Lease, M., & Kazai, G. (2011). *Overview of the trec 2011 crowdsourcing track (conference notebook).* Paper presented at the Text Retrieval Conference Notebook.

Lease, M., & Yilmaz, E. (2013). Crowdsourcing for information retrieval: introduction to the special issue. *Information Retrieval*, 1-10. doi:10.1007/s10791-013-9222-7

Lesk, M. E., & Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information storage and retrieval,* 4(4), 343-359.

Li, H., Zhao, B., & Fuxman, A. (2014). *The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing.* Paper presented at the Proceedings of the 23rd international conference on World wide web.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods,* 44(1), 1-23.

Mason, W., & Watts, D. J. (2010). Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter,* 11(2), 100-108.

Mayer, R. E. (1992). *Thinking, problem solving, cognition*: WH Freeman/Times Books/Henry Holt & Co.

McCreadie, R. M., Macdonald, C., & Ounis, I. (2010). *Crowdsourcing a news query classification dataset.* Paper presented at the Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010).

Moghadasi, S. I., Ravana, S. D., & Raman, S. N. (2013). Low-cost evaluation techniques for information retrieval systems: A review. *Journal of Informetrics,* 7(2), 301-312.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., . . . Tily, H. (2010). *Crowdsourcing and language studies: the new generation of linguistic data.* Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.

Nowak, S., & Rüger, S. (2010). *How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation.* Paper presented at the Proceedings of the international conference on Multimedia information retrieval.

Oyama, S., Baba, Y., Sakurai, Y., & Kashima, H. (2013). *Accurate integration of crowdsourced labels using workers' self-reported confidence scores.* Paper

presented at the Proceedings of the Twenty-Third international joint conference on Artificial Intelligence.

Pallant, J. (2001). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows (Versions 10 and 11): SPSS Student Version 11.0 for Windows*: Open University Press Milton Keynes, UK, USA.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making,* 5(5), 411-419.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). *An evaluation framework for plagiarism detection.* Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters.

Quinn, A. J., & Bederson, B. B. (2011). *Human computation: a survey and taxonomy of a growing field.* Paper presented at the Proceedings of the 2011 annual conference on Human factors in computing systems.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *The Journal of Machine Learning Research,* 99, 1297-1322.

Rees, A. M., & Schultz, D. G. (1967). A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Final Report to the National Science Foundation. Volume I.

Reis, H. T., & Judd, C. M. (2000). *Handbook of research methods in social and personality psychology*: Cambridge University Press.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). *Who are the crowdworkers?: shifting demographics in mechanical turk.* Paper presented at the Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems.

Ruthven, I., Baillie, M., & Elsweiler, D. (2007). The relative effects of knowledge, interest and confidence in assessing relevance. *Journal of Documentation,* 63(4), 482-504.

Sakai, T., & Kando, N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval,* 11(5), 447-470.

Salakhutdinov, R., & Mnih, A. (2008). Probabilistic matrix factorization. *Advances in neural information processing systems,* 20, 1257-1264.

Salthouse, T. A. (2014). Frequent assessments may obscure cognitive decline. *Psychological assessment,* 26(4), 1063.

Saracevic, T. (1991). *Individual Differences in Organizing, Searching and Retrieving Information.* Paper presented at the Proceedings of the ASIS annual meeting.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology,* 58(13), 2126-2144.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science,* 39(3), 197-216.

Scekic, O., Truong, H.-L., & Dustdar, S. (2012). Modeling rewards and incentive mechanisms for social BPM *Business Process Management* (pp. 150-155): Springer.

Schamber, L. (1994). Relevance and Information Behavior. *Annual review of information science and technology (ARIST),* 29, 3-48.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition∗. *Information Processing & Management,* 26(6), 755-776.

Scholer, F., Kelly, D., Wu, W.-C., Lee, H. S., & Webber, W. (2013). *The effect of threshold priming and need for cognition on relevance calibration and assessment.* Paper presented at the Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.

Scholer, F., Turpin, A., & Sanderson, M. (2011). *Quantifying test collection quality based on the consistency of relevance judgements.* Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

Schulze, T., Seedorf, S., Geiger, D., Kaufmann, N., & Schader, M. (2011). *Exploring task properties in crowdsourcing-an empirical study on mechanical turk.* Paper presented at the ECIS.

Sharit, J., Czaja, S. J., Hernandez, M., Yang, Y., Perdomo, D., Lewis, J. E., . . . Nair, S. (2004). An evaluation of performance by older persons on a simulated telecommuting task. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences,* 59(6), P305-P316.

Sharit, J., Czaja, S. J., Nair, S., & Lee, C. C. (2003). Effects of age, speech rate, and environmental support in using telephone voice menu systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 45(2), 234-251.

Sharit, J., Hernández, M. A., Czaja, S. J., & Pirolli, P. (2008). Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the web. *ACM Transactions on Computer-Human Interaction (TOCHI),* 15(1), 3.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). *Get another label? improving data quality and data mining using multiple, noisy labelers.* Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

Smithson, S. (1994). Information retrieval evaluation in practice: A case study approach. *Information Processing & Management,* 30(2), 205-221.

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.* Paper presented at the Proceedings of the conference on empirical methods in natural language processing.

Soboroff, I., Nicholas, C., & Cahan, P. (2001). *Ranking retrieval systems without relevance judgments.* Paper presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.

Sormunen, E. (2002). *Liberal relevance criteria of TREC-: Counting on negligible documents?* Paper presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

Stolee, K. T., & Elbaum, S. (2010). *Exploring the use of crowdsourcing to support empirical studies in software engineering.* Paper presented at the Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement.

Sutcliffe, A., & Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with computers,* 10(3), 321-351.

Swanson, D. R. (1977). Information retrieval as a trial-and-error process. *The Library Quarterly*, 128-148.

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *The Library Quarterly*, 389-398.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics.

Tague-Sutcliffe, J. M. (1996). Some perspectives on the evaluation of information retrieval systems. *JASIS,* 47(1), 1-3.

Tang, W., & Lease, M. (2011). *Semi-supervised consensus labeling for crowdsourcing.* Paper presented at the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR).

Teitelbaum-Kronish, P. (1984). *Relationship of selected cognitive aptitudes and personality characteristics of the online searcher to the quality of performance in online bibliographic retrieval*: New York University.

Trotman, A., & Jenkinson, D. (2007). IR Evaluation Using Multiple Assessors per Topic. *Proceedings of ADCS*.

Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003). CAPTCHA: Using hard AI problems for security *Advances in Cryptology—EUROCRYPT 2003* (pp. 294-311): Springer.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-based character recognition via web security measures. *Science,* 321(5895), 1465-1468.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management,* 36(5), 697-716.

Voorhees, E. M. (2002). *The philosophy of information retrieval evaluation.* Paper presented at the Evaluation of cross-language information retrieval systems.

Voorhees, E. M., & Harman, D. K. (2005). *TREC: Experiment and evaluation in information retrieval* (Vol. 63): MIT press Cambridge.

Welinder, P., & Perona, P. (2010, 13-18 June 2010). *Online crowdsourcing: Rating annotators and obtaining cost-effective labels.* Paper presented at the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops.

Wilson, P. (1973). Situational relevance. *Information storage and retrieval,* 9(8), 457-471.

Wonderlic, E. F. (1961). *Wonderlic personnel test manual*: EF Wonderlic [& Associates].

Xia, T., Zhang, C., Li, T., & Xie, J. (2011). BUPT_WILDCAT at TREC Crowdsourcing Track.

Xia, T., Zhang, C., Xie, J., & Li, T. (2012, 21-23 Sept. 2012). *Real-time quality control for crowdsourcing relevance evaluation.* Paper presented at the 2012 3rd IEEE International Conference on Network Infrastructure and Digital Content.

Zhao, Y., & Zhu, Q. (2012). Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 1-18. doi:10.1007/s10796-012-9350-4

Zhu, D., & Carterette, B. (2010). *An analysis of assessor behavior in crowdsourced preference judgments.* Paper presented at the SIGIR 2010 workshop on crowdsourcing for search evaluation.

Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J., & Azzopardi, L. (2012). Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval*, 1-39. doi:10.1007/s10791-012-9206-z

LIST OF PUBLICATIONS

Article in Academic Journals

Samimi, P., & Ravana, S. D. (2014). Creation of Reliable Relevance Judgments in Information Retrieval Systems Evaluation Experimentation through Crowdsourcing: A Review. *The Scientific World Journal, 2014* (Chapter 2 of thesis- ISI-Indexed).

Samimi, P., Ravana, S. D., & Koh, Y. S. (2016). Effect of verbal comprehension skill and self-reported features on reliability of crowdsourced relevance judgments. Computers in Human Behavior, 64, 793-804 (Chapter 4 of thesis, ISI-Indexed-Q1 Journal).

Samimi, P., & Ravana, S. D. (2015) Agreement of Relevance Assessment between Human Assessors and Crowdsourced Workers in Information Retrieval Systems Evaluation Experimentation. *Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE) Vol. 2, Issue 2.*

Accepted Paper in Journal

Effects of Objective and Subjective Competence on Reliability of Crowdsourced Relevance Judgments (The paper is accepted to publish in Information Research Journal, Chapter 3 of this thesis, ISI-Indexed).

Proceeding

Samimi, P., & Ravana, S. D. (2014). Agreement between Crowdsourced Workers and Expert Assessors in Making Relevance Judgment for System Based IR Evaluation. *Recent Advances on Soft Computing and Data Mining (399-407): Springer International Publishing* (SCOPUS-Indexed).

Samimi, P., & Ravana, S. D. (2016). Effect of Cognitive Ability on Reliability of Crowdsourced Relevance Judgments. *Paper presented at the Third International Conference on Information Retrieval and Knowledge Management (CAMP16)* (Chapter 5 of thesis, SCOPUS-Indexed).

Appendix A: Pilot Study

Document Relevance Evaluation

Instructions ▲

Step 1: There are four different topics and documents. For each topic and document, evaluate the relevance of each document to the given topic. Then answer another four questions about your evaluation.

Step 2: Answer two questions about your first language and field of interest.

Step 3: Complete vocabulary test

STEP 1

Topic 1: parkinson's disease

Document

The treatment of essential tremor or tremor due to Parkinson's disease is one of the first therapies developed by our Neurostimulation ventures group. Other initiatives currently under investigation include: Neurostimulation for advanced Parkinson's disease-the subthalamic nucleus (STN) and the internal portion of the globus pallidus (GPi)-two parts of the brain that when stimulated may affect Parkinson's disease symptoms other than tremor. Clinical trials for this indication are under way in Europe, North America, and Australia, and early results show promise.

Please rate the above document according to its relevance to the [parkinson's disease] as follows:

- ○ Relevant
- ○ Not Relevant
- ○ Don't know

Please justify your answer (we appreciate your comments):

Rate your knowledge on the topic [parkinson's disease]:

1 2 3 4

Minimal	○	○	○	○	Extensive

How difficult was this evaluation?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your evaluation?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Topic 2: what is the composition of zirconium

Document

Refractory & Reactive Metals Specialists What are "refractory" metals? One definition of "refractory" is obstinate; unmanageable; difficult to melt or work Do your production requirements include metals such as: • Titanium • Hafnium • Zirconium These are the metals demanded by many of today's industries - some because of their high heat strength and ability to survive elevated temperatures; others because of their resistance to corrosion. Whatever your need, you've probably discovered that each has production peculiarities which require special attention.

Please rate the above document according to its relevance to the [what is the composition of zirconium] as follows:

- ○ Relevant
- ○ Not Relevant
- ○ Don't know

Please justify your answer (we appreciate your comments):

Rate your knowledge on the topic [what is the composition of zirconium]:

	1	2	3	4	
Minimal	○	○	○	○	Extensive

How difficult was this evaluation?

	1	2	3	4

Easy	○	○	○	○	Difficult

How confident were you in your evaluation?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Topic 3: Jennifer Aniston

Document

"The More You Know" commercial announcements feature talent from NBC's top series along with music and an animation tag. Participating stars include David Schwimmer, Jennifer Aniston and Lisa Kudrow from "Friends", Julianna Margulies, Eriq LaSalle, Noah Wyle and Gloria Reuben from "E.R."; David Hyde-Pierce and even Eddie the dog from "Frasier"; Steven Weber from "Wings"; Joey, Matthew and Andy Lawrence from the new Sunday night comedy series "Brotherly Love"; LL Cool J from "In The House"; Jonathan Silverman and Ming-Na Wen from the new Thursday night comedy series "The Single Guy".

Please rate the above document according to its relevance to the [Jennifer Aniston] as follows:

- ○ Relevant
- ○ Not Relevant
- ○ Don't know

Please justify your answer (we appreciate your comments):

Rate your knowledge on the topic [Jennifer Aniston]:

	1	2	3	4	
Minimal	○	○	○	○	Extensive

How difficult was this evaluation?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your evaluation?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Topic 4: fasting

Document

St. Chrysostom, Homilies on Genesis HOMILY 12 On the sequel to creation: "This is the book about the origins of heaven and earth when they were created, on the day God made heaven and earth." [Gen 2:4b] COME NOW, TODAY let us fulfil our promise and move on to the accustomed instruction, connecting what we are about to say with the thread of the sermons given so far. (98c) You remember, of course, that when we were all set on one or two occasions and quite intent on following that course, concern for our brethren changed the direction of our speech towards encouragement of them.

Please rate the above document according to its relevance to the [fasting] as follows:

- ○ Relevant
- ○ Not Relevant
- ○ Don't know

Please justify your answer (we appreciate your comments):

Rate your knowledge on the topic [fasting]:

	1	2	3	4	
Minimal	○	○	○	○	Extensive

How difficult was this evaluation?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your evaluation?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

STEP 2

Answer the following two questions(Please be honest, these questions will not affect your pay).

Is English your first language?

- ○ Yes
- ○ No

Which of these subjects are you interested?

- ☐ Medicine
- ☐ Chemistry
- ☐ Entertainment
- ☐ Religion and History
- ☐ Others

STEP 3

In the following ten questions select the word that means the same or has the closest meaning to the word in capitals (Please be honest, these questions will not affect your pay).

1. AIRTIGHT

- ○ firm
- ○ light
- ○ hermetically sealed
- ○ plane sick
- ○ Don't know

2. FELINE

- ○ guileless
- ○ fabulous
- ○ equine
- ○ catlike
- ○ Don't know

3. EXCERPT

- ○ accept
- ○ extract
- ○ curtail
- ○ deprive
- ○ Don't know

4. GLOAMING

- ○ autumn

- midnight
- twilight
- daybreak
- Don't know

5. IMPLICATE

- involve
- remove
- retaliate
- exaggerate
- Don't know

6. CHEF

- cheese
- style
- head cook
- candle
- Don't know

7. MILESTONE

- marker
- plant
- soft music
- grindstone
- Don't know

8. EMERGENCE

- laziness
- identity
- contrast
- coming forth
- Don't know

9. CHOWDER

- dog
- chemical
- pigment
- stew
- Don't know

10. UNGAINLY

- cheap
- stupid
- clumsy

- hazardous
- Don't know

Test Validators

Appendix B: Verbal Comprehension Experiment

Relevance Evaluation

Instructions ▲

Step 1: There is a topic given for each task. First, answer a question about familiarity with the given topic.

Step 2: Imagine you are searching for some information about a topic in the internet. The searching engine lists five webpages for you. Which of them meet your information needs? Here we give you a topic, for the five documents below please judge their relevance with the given topic. Then answer another two questions about your judgment.

Step 3: Complete vocabulary test consists of 24 items.

Step 4: Complete six questions about yourself.

Consent Form

My name is Parnia and we are conducting a research project at the University of Malaya, Malaysia. I would appreciate your time and precise attention participating in this research study.

Herewith, we will provide you (participant) a set of 45-question which need your precise answers. Your answers will recorded anonymously and might be considered for future researches and publications.

Participant Qualifications: This task is for those whose English is their first language, and their educational grades have reached ninth grade or higher [according to the US system or the US-equivalent systems].

Participation is voluntary and you as a participant are free to refuse participation and discontinue participation now. If you, as a participant, select 'Accept' below, you agree to the terms and the conditions explained in the "Instruction" section and in the "Consent Form", and you consider yourself eligible (according to the above mentioned criteria) to take part in this study.

Please feel free to contact me at parniasamimi@siswa.um.edu.my if you need further inquiries.

○ Accept

STEP 1: Rate your familiarity with the given topic

Topic: vice president richard nixon

Rate your familiarity with the topic [vice president richard nixon] (how much do you know about this topic?):

	1	2	3	4	
Minimal	○	○	○	○	Extensive

STEP 2: For the five documents below please judge their relevance with the given topic. Then answer another two questions about your judgment.

Document 1:

Vice President of the United States From Wikipedia, the free encyclopedia. The Vice President of the United States is the holder of a public office in the United States of America created by the United States Constitution. The Vice President is the first person in the presidential line of succession, becoming the new President of the United States upon the death, resignation, or removal of the president. He or she also serves as the President of the Senate, but can only cast a vote in the event of a tie. United States Vice Presidents' tie-breaking votes happen very rarely. In recent times, the President has assigned the Vice President additional duties that fall outside the Vice President's constitutional duties. The Vice President, however, only performs such duties as an agent of and at the discretion of the President.

Please rate the document 1 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Document 2:

HARPER'S MAGAZINE Nixon, Richard M. (Richard Milhous) (19131994) Article (3) Quotation (2)

Please rate the document 2 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant

○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Document 3:

Richard Nixon Thirty-Seventh President of the United States From Martin Kelly, About.com Richard Nixon Information, Richard Nixon Biography, Richard Nixon Quotes, Watergate Scandal

Please rate the document 3 according to its relevance to the [vice president richard nixon] as follows:

○ Not Relevant
○ Relevant
○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Document 4:

2008 president richardson white house president schedule president trivia answers youngest president of us being president qualification u s vice commander in chief after president kenya president election violence president hoover greatdepression of 1929 candidates for president 1860 abama president elections mike cassady president med assets party president us 50th anniversary letter from president bush biography of president truman president bush signing statements machiavelli

Please rate the document 4 according to its relevance to the [vice president richard nixon] as follows:

○ Not Relevant
○ Relevant

○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Document 5:

Richard Nixon From Wikipedia, the free encyclopedia Richard Milhous Nixon (January 9, 1913April 22, 1994) was the 37th President of the United States (1969 1974) and the only president to resign the office. He was also the 36th Vice President of the United States (1953 1961). Nixon was born in Yorba Linda, California. After completing undergraduate work at Whittier College, he graduated from Duke University School of Law in 1937 and returned to California to practice law in La Mirada. After the attack on Pearl Harbor, he joined the United States Navy and rose to the rank of Lieutenant Commander during World War II.

Please rate the document 5 according to its relevance to the [vice president richard nixon] as follows:

○ Not Relevant
○ Relevant
○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

STEP 3: Complete vocabulary test consists of 24 items

In the following 24 questions select the word that means the same or has the closest meaning to the given word (Please be honest, these questions will not affect your pay).

1. cottontail

○ squirrel

- ○ poplar
- ○ boa
- ○ marshy plant
- ○ rabbit
- ○ Don't know

2. marketable

- ○ partisan
- ○ jocular
- ○ marriageable
- ○ salable
- ○ essential
- ○ Don't know

3. boggy

- ○ afraid
- ○ false
- ○ marshy
- ○ dense
- ○ black
- ○ Don't know

4. gruesomeness

- ○ blackness
- ○ falseness
- ○ vindictiveness
- ○ drunkenness
- ○ ghastliness
- ○ Don't know

5. loathing

- ○ diffidence
- ○ laziness
- ○ abhorrence
- ○ cleverness
- ○ comfort
- ○ Don't know

6. bantam

- ○ fowl
- ○ ridicule
- ○ cripple
- ○ vegetable
- ○ ensign

- Don't know

7. evoke

- wake up
- surrender
- reconnoiter
- transcend
- call forth
- Don't know

8. unobtrusive

- unintelligent
- epileptic
- illogical
- lineal
- modest
- Don't know

9. terrain

- ice cream
- final test
- tractor
- area of ground
- weight
- Don't know

10. capriciousness

- stubbornness
- courage
- whimsicality
- amazement
- greediness
- Don't know

11. maelstrom

- slander
- whirlpool
- enmity
- armor
- majolica
- Don't know

12. tentative

- critical

- ○ conclusive
- ○ authentic
- ○ provisional
- ○ apprehensive
- ○ Don't know

13. placate

- ○ rehabilitate
- ○ plagiarize
- ○ depredate
- ○ apprise
- ○ conciliate
- ○ Don't know

14. surcease

- ○ enlightenment
- ○ cessation
- ○ inattention
- ○ censor
- ○ substitution
- ○ Don't know

15. apathetic

- ○ wandering
- ○ impassive
- ○ hateful
- ○ prophetic
- ○ overflowing
- ○ Don't know

16. paternoster

- ○ paternalism
- ○ patricide
- ○ malediction
- ○ benediction
- ○ prayer
- ○ Don't know

17. opalescence

- ○ opulence
- ○ senescence
- ○ bankruptcy
- ○ iridescence
- ○ assiduity

- Don't know

18. lush

- stupid
- luxurious
- hazy
- putrid
- languishing
- Don't know

19. curtailment

- expenditure
- abandonment
- abridgment
- improvement
- forgery
- Don't know

20. perversity

- adversity
- perviousness
- travesty
- waywardness
- gentility
- Don't know

21. calumnious

- complimentary
- analogous
- slanderous
- tempestuous
- magnanimous
- Don't know

22. illiberality

- bigotry
- imbecility
- illegibility
- cautery
- immaturity
- Don't know

23. clabber

- rejoice

- gossip
- curdle
- crow
- hobble
- Don't know

24. sedulousness

- diligence
- credulousness
- seduction
- perilousness
- frankness
- Don't know

STEP 4: complete six questions about yourself

Please tell us a bit about yourself

I am ...

- Female
- Male

I am ...

- not yet 20
- in my 20's
- in my 30's
- in my 40's
- in my 50's
- 60+ years old

I have ...

- No education
- Basic schooling
- High school
- University degree
- Master degree
- PhD or higher

To be sure that you are paying attention, please select 'Neither Agree nor Disagree' for this item.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree

- agree
- Strongly agree

How do you rate your level of experience in using computer:

	1	2	3	4	
No previous experience	○	○	○	○	extensive experience

How do you rate your level of experience in online-search in the Internet using common search engines:

	1	2	3	4	
No previous experience	○	○	○	○	extensive experience

Test Validators

Appendix C: General Reasoning Experiment

Search Relevance

Instructions ▲

Important notice: You will receive 10¢ for completion of this task and 20¢ bonus if you qualify and complete the task precisely.

Step 1: There is a search query given for each task. First, answer a question about familiarity with the given search query.

Step 2: Imagine you are searching for some information about a topic in the internet. The search engine lists five webpages for you. Which of the listed five webpages relevant to the given topic? Here we give you a search query (topic), for the five webpages below please judge their relevance with the given search query. Then answer another two questions about your judgment.

Step 3: Complete Arithmetic Operations test consists of 10 items.This test consists of problems in mathematics. However, instead of solving the problems and finding an answer, your task will be to indicate which arithmetic operations could be used to solve the problems.

Example: If a man earns $5.75 an hour, how many hours should he work each day in order to make an average of $46.00 per day?

1- subtract

2- divide

3- add

4- multiply

Solution: In order to solve the problem you should divide $46.00 by $5.75; therefore, you should select through 2.

Step 4: Complete six questions about yourself.

Consent Form

My name is Parnia and we are conducting a research project at the University of Malaya, Malaysia. I would appreciate your time and precise attention participating in this research study.

Herewith, we will provide you (participant) a set of 32-question which need your precise answers. Your answers will be recorded anonymously and might be considered for future researches and publications.

Participant Qualifications: This task is for those whose English is their first language, and their educational grades have reached ninth grade or higher [according to the US system or the US-equivalent systems].

Participation is voluntary and you as a participant are free to refuse participation and discontinue participation now. If you, as a participant, select 'Accept' below, you agree to the terms and the conditions explained in the "Instruction" section and in the "Consent Form", and you consider yourself eligible (according to the above mentioned criteria) to take part in this study.

Please feel free to contact me at parniasamimi@siswa.um.edu.my if you need further inquiries.

○ Accept

STEP 1: Rate your familiarity with the given search query

Search query: vice president richard nixon

Description: Information about Richard Nixon as Vice President.

Rate your familiarity with the search query [vice president richard nixon] (how much do you know about this search query?):

	1	2	3	4	
Minimal	○	○	○	○	Extensive

STEP 2: For the five webpages below please judge their relevance with the given search query. Then answer another two questions about your judgment.

Webpage 1:

Vice President of the United States

From Wikipedia, the free encyclopedia
(Redirected from List of US Presidents who have also been Vice President)

This article needs additional citations for verification. Please help improve this article by adding reliable references. Unsourced material may be challenged and removed. (April 2007)

The **Vice President of the United States**[1] is the holder of a public office in the United States of America created by the United States Constitution.

The Vice President is the first person in the presidential line of succession, becoming the new President of the United States upon the death, resignation, or removal of the president.

He or she also serves as the President of the Senate, but can only cast a vote in the event of a tie. United States Vice Presidents' tie-breaking votes[2] happen very rarely.

In recent times, the President has assigned the Vice President additional duties that fall outside the Vice President's constitutional duties. The Vice President, however, only performs such duties as an agent of and at the discretion of the President.[3]

Contents

1 Eligibility
2 Oath
3 Election of the Vice President
3.1 Original Constitution and reform
3.2 Residency limitations
3.3 Nominating process
4 Role of the Vice President
4.1 Duties
4.2 President of the Senate
5 Growth of the office
6 Succession and the 25th Amendment
7 Salary
8 Vice Presidents of the United States
9 Former Vice Presidents
10 Records
11 Notes and references
12 Further reading
13 External links

Vice President of the United States

Official seal

Incumbent
Joseph Robinette Biden, Jr.
since January 20, 2009

Residence	Number One Observatory Circle
Term length	Four years
Inaugural holder	John Adams April 21, 1789
Formation	U.S. Constitution March 4, 1789
Succession	First
Website	www.whitehouse.gov/vicepresident

Eligibility [edit]

The Twelfth Amendment states that "no person constitutionally ineligible to the office of President shall be eligible to that of Vice-President of the United States."[4] Thus, to serve as Vice President, an individual must:

- Be a natural-born U.S. citizen;
- Not be younger than 35 years old; and
- Have lived in the U.S. for at least 14 years.[5]

Under the Twenty-second Amendment, the President of the United States may not be elected to more than two terms. Scholars dispute whether a former President barred from election to the Presidency is also ineligible to be elected Vice President as suggested by the Twelfth Amendment.[6][7] However, there is no similar such limitation as to how many times one can be elected Vice President.

Oath [edit]

Unlike the president, the United States Constitution does not specify an oath of office for the Vice President. Several variants of the oath have been used since 1789; the current form, which is also recited by Senators, Representatives and other government officers, has been used since 1884:

I, [illegible], do solemnly swear (or affirm) that I will

United States

This article is part of the series:
Politics and government of the United States

Federal government
Legislature
Presidency
Judiciary

Please rate the webpage 1 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

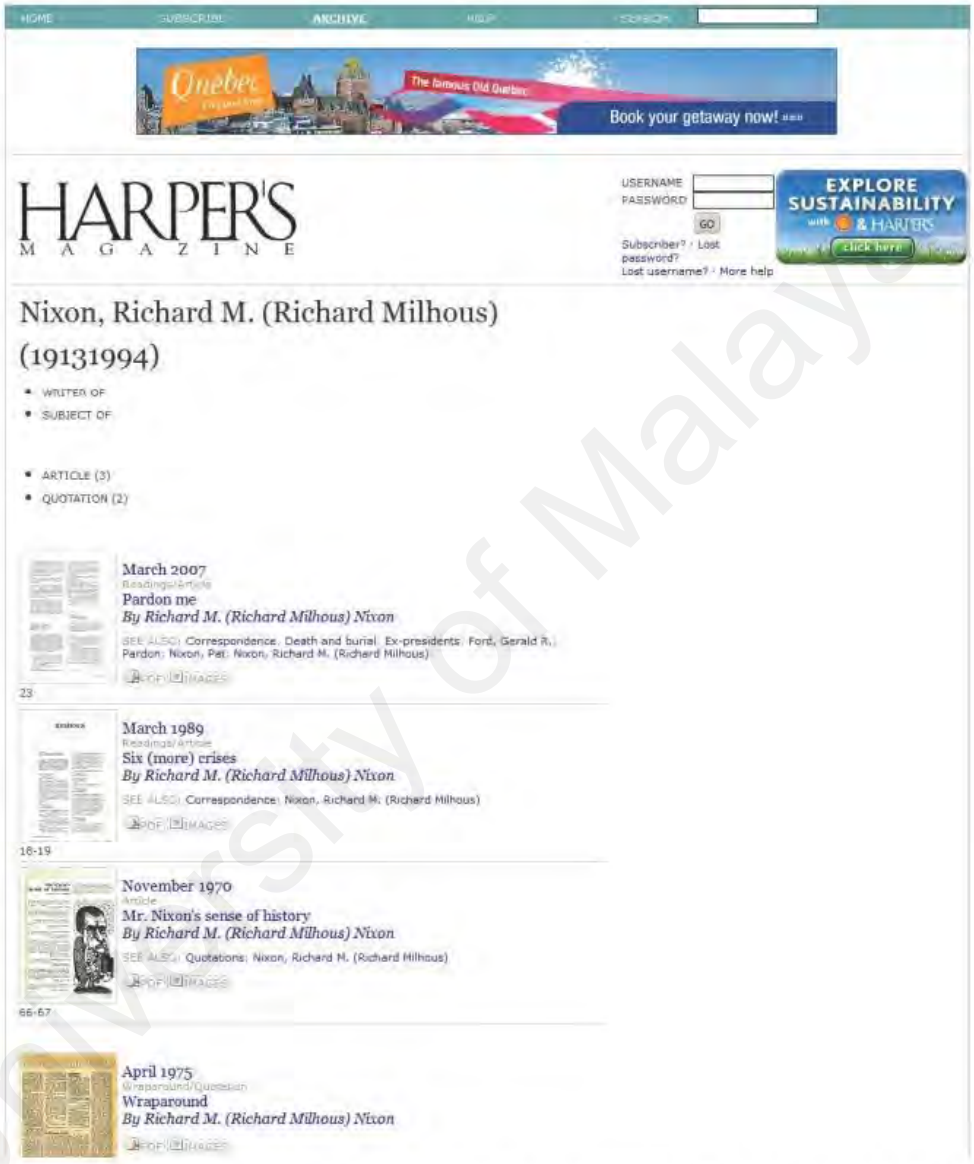
How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 2:



Please rate the Webpage 2 according to its relevance to the [vice president richard nixon] as follows:

○ Not Relevant

- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 3:

Please rate the Webpage 3 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

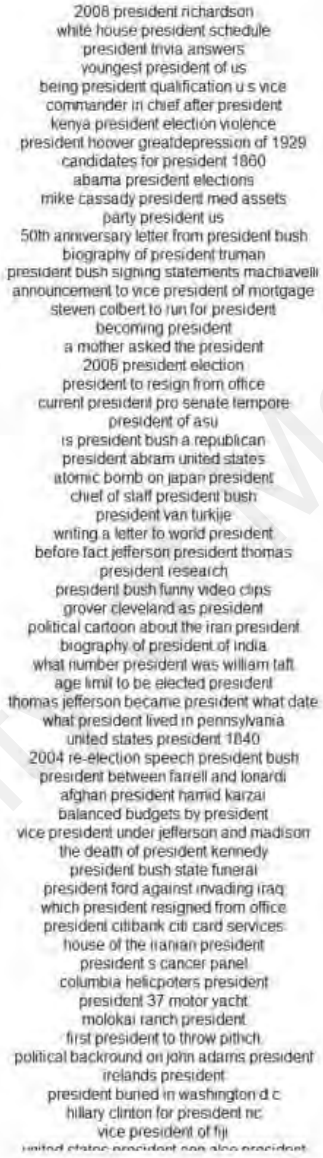
How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 4:



Please rate the Webpage 4 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 5:

Richard Nixon

From Wikipedia, the free encyclopedia
(Redirected from President Richard M. Nixon)

"Nixon" redirects here. For other uses, see Nixon (disambiguation).

Richard Milhous Nixon (January 9, 1913 □□□ April 22, 1994) was the 37th President of the United States (1969□□□1974) and the only president to resign the office. He was also the 36th Vice President of the United States (1953□□□1961).

Nixon was born in Yorba Linda, California. After completing undergraduate work at Whittier College, he graduated from Duke University School of Law in 1937 and returned to California to practice law in La Mirada. After the attack on Pearl Harbor, he joined the United States Navy and rose to the rank of Lieutenant Commander during World War II. He was elected in 1946 as a Republican to the House of Representatives representing California's 12th Congressional district, and in 1950 to the United States Senate. He was chosen by Republican Party nominee Dwight D. Eisenhower to be his running mate in 1952 and served as vice president from 1953 until 1961. Despite announcing his retirement from politics after losing the 1960 presidential election and 1962 California gubernatorial election, Nixon was elected to the presidency in 1968.

The most immediate task facing President Nixon was the Vietnam War. He initially escalated the conflict, overseeing secret bombing campaigns, but soon withdrew American troops and successfully negotiated a ceasefire with North Vietnam, effectively ending American involvement in the war. His foreign policy was largely successful; he opened relations with the People's Republic of China and initiated d□□tente with the Soviet Union. Domestically, he implemented new economic policies which called for wage and price control and the abolition of the gold standard. He was reelected by a landslide in 1972. In his second term, the nation was afflicted with economic difficulties. In the face of likely impeachment for his role in the Watergate scandal,[1] Nixon resigned on August 9, 1974. His successor, Gerald Ford, issued a pardon for any federal crimes Nixon may have committed while in office.

In his retirement, Nixon became a prolific author and undertook many foreign trips. Though far from universally popular, he gained respect as an elder statesman. He suffered a stroke on April 18, 1994, and died four days later at the age of 81.

Contents

1 Early life
2 Law practice
3 Marriage
4 World War II
5 Congressional career
  5.1 House of Representatives
  5.2 Senate
6 Vice Presidency (1953-1961)
  6.1 1960 presidential election
7 Wilderness years
8 1968 presidential election
9 Presidency (1969□□□1974)
  9.1 First term
  9.2 Second term
  9.3 Watergate
  9.4 Judicial appointments
  9.5 Pardons
10 Later life
11 Death and funeral
12 Public perception
13 Legacy
14 Bibliography
15 Notes
16 References
17 External links

Richard M. Nixon

37th President of the United States

In office
January 20, 1969 □□□ August 9, 1974

Vice President	Spiro Agnew (1969□□□1973) vacant (Oct. □□□Dec. 1973) Gerald Ford (1973□□□1974)
Preceded by	Lyndon B. Johnson
Succeeded by	Gerald Ford

36th Vice President of the United States

In office
January 20, 1953 □□□ January 20, 1961

President	Dwight D. Eisenhower
Preceded by	Alben W. Barkley
Succeeded by	Lyndon B. Johnson

United States Senator from California

In office
December 1, 1950 □□□ January 1, 1953

Preceded by	Sheridan Downey
Succeeded by	Thomas Kuchel

Member of the United States House of Representatives from California's 12th congressional district

In office
January 2, 1947 □□□ December 1, 1950

Preceded by	Jerry Voorhis
Succeeded by	Patrick J. Hillings

Born	January 9, 1913 Yorba Linda, California
Died	April 22, 1994 (aged 81) New York City
Political party	Republican
Spouse	Thelma Catherine "Pat" Ryan

Please rate the Webpage 5 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

STEP 3: Complete Arithmetic Operations test consists of 10 items.This test consists of problems in mathematics. However, instead of solving the problems and finding an answer, your task will be to indicate which arithmetic operations could be used to solve the problems.

1. There are 4 quarts in a gallon and 4 cups in a quart. How many cups are there in a gallon?

- ○ add
- ○ subtract
- ○ multiply
- ○ divide
- ○ Don't know

2. An electrical planer is set to remove .02 of an inch each time a piece of wood is passed through it. If a board is put through 7 times, how much wood will have been removed?

- ○ multiply
- ○ subtract
- ○ divide
- ○ add
- ○ Don't know

3. There are 54 children at a summer camp. If there are 33 boys attending the camp, how many campers are girls?

- ○ add
- ○ multiply
- ○ subtract
- ○ divide
- ○ Don't know

4. A man wants to seed a lawn around his new home. His lot is 120 feet by 90 feet (10,800 sq. feet). His house is centered on the lot and occupies 2,785 square feet. What is the greatest number of square feet of ground that could possibly be put into lawn?

- ○ add
- ○ divide
- ○ multiply
- ○ subtract
- ○ Don't know

5. A wholesale meat dealer sells sirloin steak for $3.19 per pound and chuck steak for $1.89 per pound. One day the meat dealer sold 76 pounds of each. How much money was taken in?

- add and divide
- add and multiply
- multiply and subtract
- divide and divide
- Don't know

6. A cyclist in an international bicycle race covered an average of 9 miles every 20 minutes. If she maintained the same average speed, how long did it take her to cycle the remaining 84 miles of the race?

- divide and multiply
- subtract and divide
- add and subtract
- divide and add
- Don't know

7. A grocer sells oranges for $1.50 a dozen. The oranges cost 75 cents a dozen. How much profit is there from each orange?

- subtract and multiply
- divide and subtract
- add and divide
- subtract and divide
- Don't know

8. A boy works in a store after school for a total of 10 hours a week He also works 8 hours on Saturdays. How much is he being paid per hour, if he makes $72.00 per week?

- multiply and subtract
- add and divide
- divide and subtract
- add and multiply
- Don't know

9. A college student takes a part-time job which pays $65.00 per week. After withholding and other taxes she is left with 76% of her salary, and each week she spends a total of $6.00 on lunches and bus fares. How much money does she make from her job each week after the taxes, lunches, and bus fares have been paid?

- divide and subtract
- subtract and multiply
- add and divide
- multiply and subtract
- Don't know

10. A rectangular underground reservoir is 15 feet deep and contains 2,000,000 gallons of water, when it is full. Spring rains filled the reservoir, but a summer drought caused the water level to drop 8 feet. How many gallons of water were consumed during the drought?

- subtract and divide
- add and subtract
- divide and multiply
- subtract and multiply
- Don't know

STEP 4: complete six questions about yourself

Please tell us a bit about yourself

I am ...

- Female
- Male

I am ...

- not yet 20
- in my 20's
- in my 30's
- in my 40's
- in my 50's
- 60+ years old

I have ...

- No education background
- Basic schooling
- High school
- Bachelor degree
- Master degree
- PhD or higher

To be sure that you are paying attention, please select 'Neither Agree nor Disagree' for this item.

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly agree

How do you rate your level of experience in using computer:

	1	2	3	4	
No previous experience	○	○	○	○	extensive experience

How do you rate your level of experience in online-search in the Internet using common search engines:

	1	2	3	4	
No previous experience	○	○	○	○	extensive experience

Test Validators

Appendix D: Logical Reasoning Experiment

Search Relevance

Instructions▲

Step 1: There is a search query given for each task. First, answer a question about familiarity with the given search query.

Step 2: Imagine you are searching for some information about a topic in the internet. The search engine lists five webpages for you. Which of the listed five webpages relevant to the given topic? Here we give you a search query (topic), for the five webpages below please judge their relevance with the given search query. Then answer another two questions about your judgment.

Step 3: Complete a test consists of 10 items. This is a test of your ability to tell whether the conclusion drawn from certain statements is correct or incorrect. Although all of the statements are really nonsense, you are to assume that the first two statements in each problem are correct. The conclusion drawn from them may or may not show good reasoning. You are to think only about the reasoning.

If the conclusion drawn from the statements shows good reasoning, select "True" otherwise select "False". Example:

All trees are fish. All fish are horses Therefore all trees are horses.

1) True 2) False

Solution: you should select True.

Step 4: Complete six questions about yourself.

Consent Form

My name is Parnia and we are conducting a research project at the University of Malaya, Malaysia. I would appreciate your time and precise attention participating in this research study.

Herewith, we will provide you (participant) a set of 32-question which need your precise answers. Your answers will be recorded anonymously and might be considered for future researches and publications.

Participant Qualifications: This task is for those whose English is their first language, and their educational grades have reached ninth grade or higher [according to the US system or the US-equivalent systems].

Participation is voluntary and you as a participant are free to refuse participation and discontinue participation now. If you, as a participant, select 'Accept' below, you agree to the terms and the conditions explained in the "Instruction" section and in the "Consent Form", and you consider

yourself eligible (according to the above mentioned criteria) to take part in this study.

Please feel free to contact me at parniasamimi@siswa.um.edu.my if you need further inquiries.

○ Accept

STEP 1: Rate your familiarity with the given search query

Search query: vice president richard nixon

Description: Information about Richard Nixon as Vice President.

Rate your familiarity with the search query [vice president richard nixon] (how much do you know about this search query?):

	1	2	3	4	
Minimal	○	○	○	○	Extensive

STEP 2: For the five webpages below please judge their relevance with the given search query. Then answer another two questions about your judgment.

Webpage 1:

Vice President of the United States

From Wikipedia, the free encyclopedia
(Redirected from List of US Presidents who have also been Vice President)

This article needs additional citations for verification. Please help improve this article by adding reliable references. Unsourced material may be challenged and removed. (April 2007)

The **Vice President of the United States**[1] is the holder of a public office in the United States of America created by the United States Constitution.

The Vice President is the first person in the presidential line of succession, becoming the new President of the United States upon the death, resignation, or removal of the president.

He or she also serves as the President of the Senate, but can only cast a vote in the event of a tie. United States Vice Presidents' tie-breaking votes[2] happen very rarely.

In recent times, the President has assigned the Vice President additional duties that fall outside the Vice President's constitutional duties. The Vice President, however, only performs such duties as an agent of and at the discretion of the President.[3]

Contents

1 Eligibility
2 Oath
3 Election of the Vice President
3.1 Original Constitution and reform
3.2 Residency limitations
3.3 Nominating process
4 Role of the Vice President
4.1 Duties
4.2 President of the Senate
5 Growth of the office
6 Succession and the 25th Amendment
7 Salary
8 Vice Presidents of the United States
9 Former Vice Presidents
10 Records
11 Notes and references
12 Further reading
13 External links

Vice President of the United States

Official seal

Incumbent
Joseph Robinette Biden, Jr.
since January 20, 2009

Residence	Number One Observatory Circle
Term length	Four years
Inaugural holder	John Adams April 21, 1789
Formation	U.S. Constitution March 4, 1789
Succession	First
Website	www.whitehouse.gov/vicepresident

Eligibility [edit]

The Twelfth Amendment states that "no person constitutionally ineligible to the office of President shall be eligible to that of Vice-President of the United States."[4] Thus, to serve as Vice President, an individual must:

- Be a natural-born U.S. citizen;
- Not be younger than 35 years old; and
- Have lived in the U.S. for at least 14 years.[5]

Under the Twenty-second Amendment, the President of the United States may not be elected to more than two terms. Scholars dispute whether a former President barred from election to the Presidency is also ineligible to be elected Vice President as suggested by the Twelfth Amendment.[6][7] However, there is no similar such limitation as to how many times one can be elected Vice President.

Oath [edit]

Unlike the president, the United States Constitution does not specify an oath of office for the Vice President. Several variants of the oath have been used since 1789; the current form, which is also recited by Senators, Representatives and other government officers, has been used since 1884:

I, [illegible], do solemnly swear (or affirm) that I will

United States

This article is part of the series:
Politics and government of the United States

Federal government
Legislature
Presidency
Judiciary

Please rate the webpage 1 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

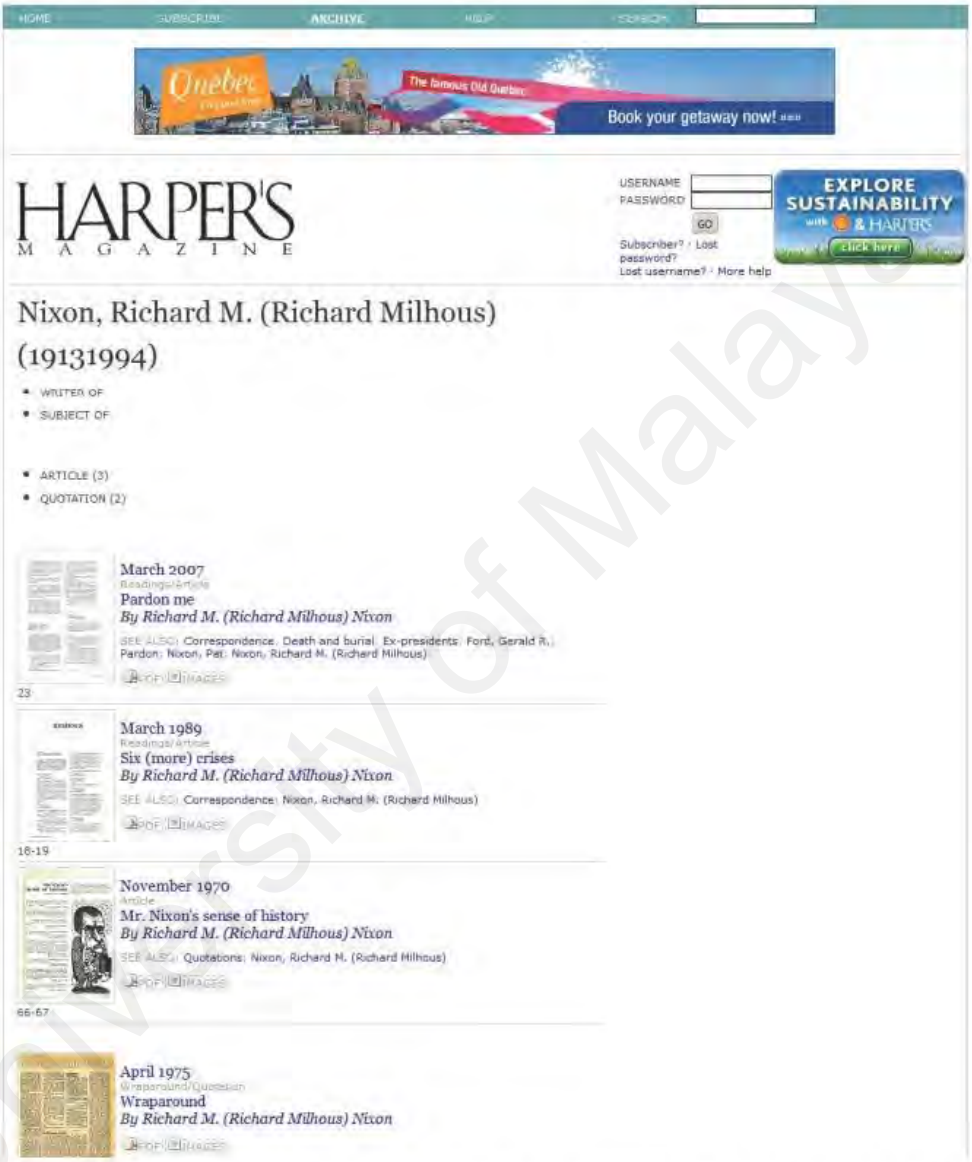
How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 2:



Please rate the Webpage 2 according to its relevance to the [vice president richard nixon] as follows:

○ Not Relevant

- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 3:

Please rate the Webpage 3 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 4:

2008 president richardson
white house president schedule
president trivia answers
youngest president of us
being president qualification u s vice
commander in chief after president
kenya president election violence
president hoover greatdepression of 1929
candidates for president 1860
obama president elections
mike cassady president med assets
party president us
50th anniversary letter from president bush
biography of president truman
president bush signing statements machiavelli
announcement to vice president of mortgage
steven colbert to run for president
becoming president
a mother asked the president
2008 president election
president to resign from office
current president pro senate tempore
president of asu
is president bush a republican
president abram united states
atomic bomb on japan president
chief of staff president bush
president van turkije
writing a letter to world president
before fact jefferson president thomas
president research
president bush funny video clips
grover cleveland as president
political cartoon about the iran president
biography of president of india
what number president was william taft
age limit to be elected president
thomas jefferson became president what date
what president lived in pennsylvania
united states president 1840
2004 re-election speech president bush
president between farrell and lonardi
afghan president hamid karzai
balanced budgets by president
vice president under jefferson and madison
the death of president kennedy
president bush state funeral
president ford against invading iraq
which president resigned from office
president citibank citi card services
house of the iranian president
president s cancer panel
columbia helicpoters president
president 37 motor yacht
molokai ranch president
first president to throw pithch
political backround on john adams president
irelands president
president buried in washington d c
hillary clinton for president nc
vice president of fiji
united states president can also president

Please rate the Webpage 4 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

Webpage 5:

Richard Nixon

From Wikipedia, the free encyclopedia
(Redirected from President Richard M. Nixon)

"Nixon" redirects here. For other uses, see Nixon (disambiguation).

Richard Milhous Nixon (January 9, 1913 □□□ April 22, 1994) was the 37th President of the United States (1969□□□1974) and the only president to resign the office. He was also the 36th Vice President of the United States (1953□□□1961).

Nixon was born in Yorba Linda, California. After completing undergraduate work at Whittier College, he graduated from Duke University School of Law in 1937 and returned to California to practice law in La Mirada. After the attack on Pearl Harbor, he joined the United States Navy and rose to the rank of Lieutenant Commander during World War II. He was elected in 1946 as a Republican to the House of Representatives representing California's 12th Congressional district, and in 1950 to the United States Senate. He was chosen by Republican Party nominee Dwight D. Eisenhower to be his running mate in 1952 and served as vice president from 1953 until 1961. Despite announcing his retirement from politics after losing the 1960 presidential election and 1962 California gubernatorial election, Nixon was elected to the presidency in 1968.

The most immediate task facing President Nixon was the Vietnam War. He initially escalated the conflict, overseeing secret bombing campaigns, but soon withdrew American troops and successfully negotiated a ceasefire with North Vietnam, effectively ending American involvement in the war. His foreign policy was largely successful; he opened relations with the People's Republic of China and initiated d□□tente with the Soviet Union. Domestically, he implemented new economic policies which called for wage and price control and the abolition of the gold standard. He was reelected by a landslide in 1972. In his second term, the nation was afflicted with economic difficulties. In the face of likely impeachment for his role in the Watergate scandal,[1] Nixon resigned on August 9, 1974. His successor, Gerald Ford, issued a pardon for any federal crimes Nixon may have committed while in office.

In his retirement, Nixon became a prolific author and undertook many foreign trips. Though far from universally popular, he gained respect as an elder statesman. He suffered a stroke on April 18, 1994, and died four days later at the age of 81.

Contents

1 Early life
2 Law practice
3 Marriage
4 World War II
5 Congressional career
  5.1 House of Representatives
  5.2 Senate
6 Vice Presidency (1953-1961)
  6.1 1960 presidential election
7 Wilderness years
8 1968 presidential election
9 Presidency (1969□□□1974)
  9.1 First term
  9.2 Second term
  9.3 Watergate
  9.4 Judicial appointments
  9.5 Pardons
10 Later life
11 Death and funeral
12 Public perception
13 Legacy
14 Bibliography
15 Notes
16 References
17 External links

Richard M. Nixon

37th President of the United States

In office
January 20, 1969 □□□ August 9, 1974

Vice President: Spiro Agnew (1969□□□1973)
vacant (Oct. □□□Dec. 1973)
Gerald Ford (1973□□□1974)

Preceded by: Lyndon B. Johnson
Succeeded by: Gerald Ford

36th Vice President of the United States

In office
January 20, 1953 □□□ January 20, 1961

President: Dwight D. Eisenhower
Preceded by: Alben W. Barkley
Succeeded by: Lyndon B. Johnson

United States Senator from California

In office
December 1, 1950 □□□ January 1, 1953

Preceded by: Sheridan Downey
Succeeded by: Thomas Kuchel

Member of the United States House of Representatives from California's 12th congressional district

In office
January 2, 1947 □□□ December 1, 1950

Preceded by: Jerry Voorhis
Succeeded by: Patrick J. Hillings

Born: January 9, 1913
Yorba Linda, California

Died: April 22, 1994 (aged 81)
New York City

Political party: Republican

Spouse: Thelma Catherine "Pat" Ryan

Please rate the Webpage 5 according to its relevance to the [vice president richard nixon] as follows:

- ○ Not Relevant
- ○ Relevant
- ○ Highly Relevant

How difficult was this judgment?

	1	2	3	4	
Easy	○	○	○	○	Difficult

How confident were you in your judgment?

	1	2	3	4	
Not confident	○	○	○	○	Very confident

STEP 3: Complete this test consists of 10 items. This is a test of your ability to tell whether the conclusion drawn from certain statements is correct or incorrect. Although all of the statements are really nonsense, you are to assume that the first two statements in each problem are correct. The conclusion drawn from them may or may not show good reasoning. You are to think only about the reasoning.

1. All birds have purple tails. All cats are birds. Therefore all cats have purple tails.

- ○ True
- ○ False
- ○ Don't know

2. No singer is a pogo stick. All pogo sticks are movie stars. Therefore no singer is a movie star.

- ○ True
- ○ False
- ○ Don't know

3. All cars have sails. Some swimming pools are cars. Therefore some swimming pools have sails.

- ○ True
- ○ False
- ○ Don't know

4. No chipmunks are clowns. Some mushrooms are chipmunks. Therefore some mushrooms are not clowns.

- ○ True
- ○ False
- ○ Don't know

5. No skunks have green toes. All skunks are pigs. Therefore no pig has green toes.

- ○ True
- ○ False
- ○ Don't know

6. All horses have wings. No turtle has wings. Therefore no turtle is a horse.

- ○ True
- ○ False

- Don't know

7. No hummingbirds fly. Some tractors fly. Therefore some tractors are not hummingbirds.

- True
- False
- Don't know

8. All apes are houseflies. Some houseflies are not snails. Therefore some apes are not snails.

- True
- False
- Don't know

9. Some dogs like to sing. All dogs are snowdrifts. Therefore some snowdrifts like to sing.

- True
- False
- Don't know

10. All doctors are sea horses. Some doctors are tornadoes. Therefore some tornadoes are sea horses.

- True
- False
- Don't know

STEP 4: complete six questions about yourself

Please tell us a bit about yourself

I am ...

- Female
- Male

I am ...

- not yet 20
- in my 20's
- in my 30's
- in my 40's
- in my 50's
- 60+ years old

I have ...

- ○ No education background
- ○ Basic schooling
- ○ High school
- ○ Bachelor degree
- ○ Master degree
- ○ PhD or higher

To be sure that you are paying attention, please select 'Neither Agree nor Disagree' for this item.

- ○ Strongly Disagree
- ○ Disagree
- ○ Neither Agree nor Disagree
- ○ Agree
- ○ Strongly agree

How do you rate your level of experience in using computer:

	1	2	3	4	
No previous experience	○	○	○	○	extensive experience

How do you rate your level of experience in online-search in the Internet using common search engines:

	1	2	3	4	
No previous experience	○	○	○	○	extensive experience

Test Validators

Appendix E: Verbal Comprehension (Post-Hoc)

Multiple Comparisons

Games-Howell

Dependent Variable	(I) Verbal Comprehension Score (Binned)	(J) Verbal Comprehension Score (Binned)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Ternary Accuracy	1	2	-.07926*	.01744	.000	-.1202	-.0383
		3	-.16531*	.01847	.000	-.2087	-.1219
	2	1	.07926*	.01744	.000	.0383	.1202
		3	-.08605*	.01878	.000	-.1302	-.0419
	3	1	.16531*	.01847	.000	.1219	.2087
		2	.08605*	.01878	.000	.0419	.1302
Binary Accuracy	1	2	-.12594*	.01788	.000	-.1680	-.0839
		3	-.20729*	.01721	.000	-.2477	-.1668
	2	1	.12594*	.01788	.000	.0839	.1680
		3	-.08135*	.01633	.000	-.1197	-.0430
	3	1	.20729*	.01721	.000	.1668	.2477
		2	.08135*	.01633	.000	.0430	.1197
Precision	1	2	-.07084*	.01978	.001	-.1173	-.0243
		3	-.13999*	.01894	.000	-.1845	-.0955
	2	1	.07084*	.01978	.001	.0243	.1173
		3	-.06916*	.01767	.000	-.1107	-.0276
	3	1	.13999*	.01894	.000	.0955	.1845
		2	.06916*	.01767	.000	.0276	.1107
Recall	1	2	-.12571*	.02085	.000	-.1747	-.0767
		3	-.15227*	.01993	.000	-.1991	-.1054
	2	1	.12571*	.02085	.000	.0767	.1747
		3	-.02657	.01637	.237	-.0650	.0119
	3	1	.15227*	.01993	.000	.1054	.1991
		2	.02657	.01637	.237	-.0119	.0650
NDCG	1	2	-.02865*	.00967	.009	-.0514	-.0059
		3	-.06094*	.00912	.000	-.0824	-.0395
	2	1	.02865*	.00967	.009	.0059	.0514
		3	-.03229*	.00836	.000	-.0519	-.0127
	3	1	.06094*	.00912	.000	.0395	.0824
		2	.03229*	.00836	.000	.0127	.0519
Specificity	1	2	-.10169*	.03406	.008	-.1817	-.0059
		3	-.26291*	.03498	.000	-.3451	-.0395
	2	1	.10169*	.03406	.008	.0217	.0514

		3	-.16122*	.03603	.000	-.2459	-.0127
	3	1	.26291*	.03498	.000	.1807	.0824
		2	.16122*	.03603	.000	.0766	.0519
Effectiveness	1	2	-.14459*	.02723	.000	-.2086	-.0806
		3	-.32644*	.02968	.000	-.3962	-.2567
	2	1	.14459*	.02723	.000	.0806	.2086
		3	-.18185*	.03264	.000	-.2585	-.1052
	3	1	.32644*	.02968	.000	.2567	.3962
		2	.18185*	.03264	.000	.1052	.2585

*. The mean difference is significant at the 0.05 level.

Appendix F: General Reasoning (Post-Hoc)

Multiple Comparisons

Bonferroni

Dependent Variable	(I) ThreeGroups	(J) ThreeGroups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Ternary Accuracy	Low General Reasoning Skill	Moderate General Reasoning Skill	-.10915*	.02044	.000	-.1582	-.0601
		High General Reasoning Skill	-.12167*	.02001	.000	-.1697	-.0737
	Moderate General Reasoning Skill	Low General Reasoning Skill	.10915*	.02044	.000	.0601	.1582
		High General Reasoning Skill	-.01252	.01926	1.000	-.0587	.0337
	High General Reasoning Skill	Low General Reasoning Skill	.12167*	.02001	.000	.0737	.1697
		Moderate General Reasoning Skill	.01252	.01926	1.000	-.0337	.0587
Precision	Low General Reasoning Skill	Moderate General Reasoning Skill	-.09466*	.01933	.000	-.1410	-.0483
		High General Reasoning Skill	-.12010*	.01892	.000	-.1655	-.0747
	Moderate General Reasoning Skill	Low General Reasoning Skill	.09466*	.01933	.000	.0483	.1410
		High General Reasoning Skill	-.02544	.01822	.489	-.0691	.0183
	High General Reasoning Skill	Low General Reasoning Skill	.12010*	.01892	.000	.0747	.1655
		Moderate General Reasoning Skill	.02544	.01822	.489	-.0183	.0691

*. The mean difference is significant at the 0.05 level.

Multiple Comparisons

Games-Howell

Dependent Variable	(I) ThreeGroups	(J) ThreeGroups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Binary Accuracy	Low General Reasoning Skill	Moderate General Reasoning Skill	-.12742*	.01825	.000	-.1703	-.0845
		High General Reasoning Skill	-.16168*	.01756	.000	-.2029	-.1204
	Moderate General Reasoning Skill	Low General Reasoning Skill	.12742*	.01825	.000	.0845	.1703
		High General Reasoning Skill	-.03425	.01576	.077	-.0713	.0028
	High General Reasoning Skill	Low General Reasoning Skill	.16168*	.01756	.000	.1204	.2029
		Moderate General Reasoning Skill	.03425	.01576	.077	-.0028	.0713
Recall	Low General Reasoning Skill	Moderate General Reasoning Skill	-.09540*	.02099	.000	-.1448	-.0460
		High General Reasoning Skill	-.11147*	.02002	.000	-.1586	-.0644
	Moderate General Reasoning Skill	Low General Reasoning Skill	.09540*	.02099	.000	.0460	.1448
		High General Reasoning Skill	-.01607	.01594	.572	-.0535	.0214
	High General Reasoning Skill	Low General Reasoning Skill	.11147*	.02002	.000	.0644	.1586
		Moderate General Reasoning Skill	.01607	.01594	.572	-.0214	.0535
NDCG	Low General Reasoning Skill	Moderate General Reasoning Skill	-.02612*	.00911	.012	-.0475	-.0047
		High General Reasoning Skill	-.03849*	.00893	.000	-.0595	-.0175
	Moderate General Reasoning Skill	Low General Reasoning Skill	.02612*	.00911	.012	.0047	.0475
		High General Reasoning Skill	-.01238	.00728	.206	-.0295	.0047
	High General Reasoning Skill	Low General Reasoning Skill	.03849*	.00893	.000	.0175	.0595
		Moderate General Reasoning Skill	.01238	.00728	.206	-.0047	.0295

Specificity	Low General Reasoning Skill	Moderate General Reasoning Skill	-.10867*	.03586	.008	-.1919	-.235
		High General Reasoning Skill	-.21322*	.03485	.000	-.2951	-.1313
	Moderate General Reasoning Skill	Low General Reasoning Skill	.10867*	.03583	.008	.0235	.1919
		High General Reasoning Skill	-.10555	.03533	.008	-.1885	-.0225
	High General Reasoning Skill	Low General Reasoning Skill	.21322*	.03485	.000	.1313	.2951
		Moderate General Reasoning Skill	.10555	.03533	.008	.0225	.1885
Effectiveness	Low General Reasoning Skill	Moderate General Reasoning Skill	-.11215*	.03036	.001	-.1835	-.0408
		High General Reasoning Skill	-.23154*	.03049	.000	-.3032	-.1599
	Moderate General Reasoning Skill	Low General Reasoning Skill	.11215*	.03036	.001	.0408	.1835
		High General Reasoning Skill	-.11939*	.03175	.001	-.1940	-.0448
	High General Reasoning Skill	Low General Reasoning Skill	.23154*	.03049	.000	.1599	.3032
		Moderate General Reasoning Skill	.11939*	.03175	.001	.0448	.1940

*. The mean difference is significant at the 0.05 level.

Appendix G: Logical Reasoning (Post-Hoc)

Multiple Comparisons

Bonferroni

Dependent Variable	(I) ThreeGroups	(J) ThreeGroups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Ternary Accuracy	Low logical reasoning	Moderate logical reasoning	-.10478*	.01911	.000	-.1506	-.0589
		High logical reasoning	-.11246*	.02099	.000	-.1628	-.0621
	Moderate logical reasoning	Low logical reasoning	.10478*	.01911	.000	.0589	.1506
		High logical reasoning	-.00768	.02049	1.000	-.0568	.0415
	High logical reasoning	Low logical reasoning	.11246*	.02099	.000	.0621	.1628
		Moderate logical reasoning	.00768	.02049	1.000	-.0415	.0568
Specificity	Low logical reasoning	Moderate logical reasoning	-.14817*	.03609	.000	-.2348	-.0616
		High logical reasoning	-.16087*	.03964	.000	-.2560	-.0658
	Moderate logical reasoning	Low logical reasoning	.14817*	.03609	.000	.0616	.2348
		High logical reasoning	-.01270	.03869	1.000	-.1055	.0801
	High logical reasoning	Low logical reasoning	.16087*	.03964	.000	.0658	.2560
		Moderate logical reasoning	.01270	.03869	1.000	-.0801	.1055

Multiple Comparisons

Games-Howell

Dependent Variable	(I) ThreeGroups	(J) ThreeGroups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Binary Accuracy	Low logical reasoning	Moderate logical reasoning	-.10503*	.01867	.000	-.1489	-.0612
		High logical reasoning	-.14867*	.02001	.000	-.1957	-.1016
	Moderate logical reasoning	Low logical reasoning	.10503*	.01867	.000	.0612	.1489
		High logical reasoning	-.04363*	.01757	.036	-.0850	-.0023
	High logical reasoning	Low logical reasoning	.14867*	.02001	.000	.1016	.1957
		Moderate logical reasoning	.04363*	.01757	.036	.0023	.0850
Precision	Low logical reasoning	Moderate logical reasoning	-.08322*	.02064	.000	-.1317	-.0347
		High logical reasoning	-.11569*	.02191	.000	-.1672	-.0642
	Moderate logical reasoning	Low logical reasoning	.08322*	.02064	.000	.0347	.1317
		High logical reasoning	-.03247	.01913	.207	-.0774	.0125
	High logical reasoning	Low logical reasoning	.11569*	.02191	.000	.0642	.1672
		Moderate logical reasoning	.03247	.01913	.207	-.0125	.0774
Recall	Low logical reasoning	Moderate logical reasoning	-.07327*	.02115	.002	-.1230	-.0235
		High logical reasoning	-.10400*	.02163	.000	-.1549	-.0531
	Moderate logical reasoning	Low logical reasoning	.07327*	.02115	.002	.0235	.1230
		High logical reasoning	-.03072	.01724	.177	-.0712	.0098
	High logical reasoning	Low logical reasoning	.10400*	.02163	.000	.0531	.1549
		Moderate logical reasoning	.03072	.01724	.177	-.0098	.0712

NDCG	Low logical reasoning	Moderate logical reasoning	-.02237*	.00887	.032	-.0432	-.0015
		High logical reasoning	-.02832*	.00914	.006	-.0498	-.0068
	Moderate logical reasoning	Low logical reasoning	.02237*	.00887	.032	.0015	.0432
		High logical reasoning	-.00595	.00847	.762	-.0259	.0140
	High logical reasoning	Low logical reasoning	.02832*	.00914	.006	.0068	.0498
		Moderate logical reasoning	.00595	.00847	.762	-.0140	.0259
Effectiveness	Low logical reasoning	Moderate logical reasoning	-.15251*	.03002	.000	-.2231	-.0820
		High logical reasoning	-.18836*	.03464	.000	-.2698	-.1069
	Moderate logical reasoning	Low logical reasoning	.15251*	.03002	.000	.0820	.2231
		High logical reasoning	-.03585	.03557	.572	-.1195	.0478
	High logical reasoning	Low logical reasoning	.18836*	.03464	.000	.1069	.2698
		Moderate logical reasoning	.03585	.03557	.572	-.0478	.1195

*. The mean difference is significant at the 0.05 level.