

An Investigation of Preference Judging Consistency

by

Linh Nhi Phan Minh

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

© Linh Nhi Phan Minh 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

My supervisor, Professor Mark D. Smucker, contributed directly to the ideation of this thesis and the analysis of inter-assessor agreement discussed in Chapter 4.3.1. He also wrote the code used to calculate RBO scores shown in Appendix B as well as the code written in the “rbo” folder linked in Appendix C.

Mahsa Seifikar built the preference judging system described in Chapter 3.3.1, which I modified for my thesis study.

Amir Vakili Tahami and Dake Zhang contributed in creating the HTML pages described in Chapter 3.2.2, as shown in Figure 3.3.

I am the sole author of the rest of this thesis, written under the guidance of Professor Mark D. Smucker.

Abstract

Preference judging has been proposed as an effective method to identify the most relevant documents for a given search query. In this thesis, we investigate the degree to which assessors using a preference judging system are able to consistently find the same top documents and how consistent they are in their own preferences. We also examine to what extent variability in assessor preferences affect the evaluation of information retrieval systems. We designed and conducted a user study where 40 participants were recruited to preference judge 30 topics taken from the 2021 TREC Health Misinformation track.

The research study found that the number of judgments needed to find the top-10 preferred documents using preference judging is about twice the number of documents in that topic. It also suggests that relying on just one non-professional assessor to do preference judging is not sufficient for evaluating information retrieval systems. Additionally, the study showed that preference judging to find the top-10 documents does significantly change the rankings of runs as compared to the rankings reported in the TREC 2021 Health Misinformation track, with most changes happening among the lower-ranked runs rather than the top-ranked runs.

Overall, this thesis provides insights into assessor behaviour and assessor agreement when using preference judgments for evaluating information retrieval systems.

Acknowledgements

Firstly, I would like to express my sincere thanks and appreciation to Professor Mark D. Smucker for giving me the opportunity to embark on the journey of completing my Master's thesis and for being my supervisor these past few years. Throughout the journey of completing my degree, he has constantly guided and supported me in my research and studies. Without Professor Smucker, I would not have achieved this great result in my life.

I would also like to thank Professor Charles L. A. Clarke and Professor Gordon V. Cormack for being my thesis readers. Additionally, I would like to thank everyone who helped me with my thesis.

Furthermore, I am very thankful for all the friends I have made in my life who have stuck by me and supported me. I especially would like to thank my best friends Wanetha, Kanjanawan, and Jimin for taking care of me and listening to my stories. I would also like to thank my friend Anna for welcoming me to Canada and for always going out of her way to help me. And to all my friends who I can't thank by name, thank you for making an impact in my life.

Most importantly, I would like to express my deepest gratitude to my family for their eternal love and support. Without them, I would not have made it this far in life. To my mom and dad, thank you for being my constant support and motivation. Thank you for loving me and taking care of me with everything you have. To my sister, thank you for making me laugh whenever we talk and being there for me. To my grandparents, thank you for always being so kind to me. And to my uncle Dung, thank you for your support.

Last but not least, thank you to my love Nicholas for accompanying me throughout this journey and cheering me on. Thank you for giving me advice, for taking care of me, and for always helping me when I have problems.

I also acknowledge that this work was supported in part by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-04665, RGPAS2020-00080) and in part by the Digital Research Alliance of Canada.

Dedication

To my parents, my sister, and my love

Table of Contents

Author's Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	x
List of Tables	xii
1 Introduction	1
1.1 Research Motivation	2
1.2 Thesis Overview	3
1.3 Contributions	4
2 Related Work	5
2.1 Evaluation in Information Retrieval	5
2.1.1 Online and Offline Evaluation	6
2.1.2 Relevance Judgments	6
2.2 Assessor Agreement	9

3	Methods and Materials	10
3.1	Text REtrieval Conference (TREC) Health Misinformation Track	10
3.2	Data	11
3.2.1	Topics	11
3.2.2	Documents	15
3.3	Study Design	18
3.3.1	System Interface	19
3.3.2	Reordering Task Google Form	22
3.3.3	Pre-study Procedures	25
3.3.4	Pilot Study for the Tutorial Phase	25
3.3.5	Tutorial Phase	28
3.3.6	Main Phase	30
3.4	Participants	31
3.5	System Bugs in the Preference Judging System: Observations and Remedial Actions	34
3.6	Data Cleaning	34
3.7	Measures Used for Data Analysis	34
4	Results and Discussion	36
4.1	Analysis of Results from Tutorial Phase	36
4.2	Analysis of Assessor Behaviour	37
4.2.1	Analysis of Judgments Made without “Undo” Actions	38
4.2.2	Analysis of Judgments Made because of an “Undo” Action	44
4.3	Analysis of Agreement between Assessors	46
4.3.1	Inter-assessor Agreement	46
4.3.2	Intra-assessor Agreement	52
4.3.3	Summary of Assessor Agreement	55
4.4	Analysis of the Effect of Preference Judgments on the Ranking of Runs	55

5 Conclusion	70
References	72
APPENDICES	79
A Preference Judging Instructions Manual	80
B Code for Calculating RBO scores (rbo.py)	90
C Code for the Rest of the Data Analysis	92

List of Figures

1.1	Preference ordering for documents in TREC 2021 Health Misinformation track, as written in the paper by Clarke et al. [12]	3
3.1	Example of a topic in TREC 2021 Health Misinformation track	12
3.2	Scatter plot of helpful minus harmful compatibility scores of runs using original preference ordering (-3 to 12) vs. helpful minus harmful compatibility scores of runs using new preference ordering (-1 to 1)	16
3.3	Example of an HTML page for one topic. Please note this is a screenshot of what the page looks like. There are more documents with lower rankings at the bottom of the page which is not depicted here.	18
3.4	Login page of the system	20
3.5	Home page of the system	20
3.6	Preference judging interface	21
3.7	Topic information pop-up	22
3.8	Highlighting features of the system	23
3.9	Clicking the three-dash icon at the top right of the page allows participants to either log out of the system or go back to the home page	23
3.10	Pop-up that appears if participants spend more than 5 minutes judging a pair of documents	24
3.11	First section of the Google form for the reordering task in the main phase that provides information about the task and the topic and asks participants to input the date and time they started working on the task.	26
3.12	Second part of the Google form for the reordering task in the main phase that participants use to input the ranks for the documents they are reordering	27

3.13	Topic Assignment for the Main Phase	32
4.1	Judgment Time Distribution for 60 Seconds	39
4.2	Judgment Time Distribution for 180 Seconds	40
4.3	Judgment Time Distribution for 600 Seconds	41
4.4	Participant Average Judgment Time Across Topics	42
4.5	Topic Average Judgment Time Across Participants	43
4.6	Number of Documents Vs. Number of Judgments	44
4.7	Number of Documents Vs. Average Number of Judgments per Topic	45
4.8	Undo Action Frequency Distribution	47
4.9	Time Distribution of Judgments that Involved Undo Actions	48
4.10	Kendall's Tau Scores for Inter-assessor Agreement	50
4.11	RBO Scores with Patience Parameter Set to 0.70 for Inter-assessor Agreement	51
4.12	Cohen's Kappa Scores for Inter-assessor Agreement	53
4.13	Fleiss' Kappa Scores for Inter-assessor Agreement	54
4.14	Intra-assessor Agreement Compatibility Scores for All Topics with Patience Parameter Set to 0.70	56
4.15	Intra-assessor Agreement Compatibility Scores for Topics with ≤ 10 Docu- ments with Patience Parameter Set to 0.70	57
4.16	Intra-assessor Agreement Compatibility Scores for Topics with > 10 Docu- ments with Patience Parameter Set to 0.70	58
4.17	NIST Helpful Compatibility Scores vs. Participant 1 Helpful Compatibility Scores for Runs	61
4.18	NIST Helpful Minus Harmful Compatibility Scores vs. Participant 1 Helpful Minus Harmful Compatibility Scores for Runs	62
4.19	Kendall's Tau Distribution of Correlation between NIST vs. Participants' Run Rankings	65
4.20	AP Correlation Distribution between NIST vs. Participants' Run Rankings	66
4.21	Kendall's Tau Distribution of Correlation between Assessor vs. Assessor Run Rankings	68
4.22	AP Correlation Distribution between Assessor vs. Assessor Run Rankings	69

List of Tables

3.1	List of 30 topics used for the study	13
3.2	Topics Used in the Tutorial Phase	15
4.1	Frequency of Participants' Preference Judgment Accuracy for the Second Topic of the Tutorial Phase	37
4.2	Topics with Negative Fleiss' Kappa Scores	52
4.3	NIST Top 10 Runs and How Their Ranks Change Under the 3 Participants' Qrels Based on Average Compatibility Scores	63

Chapter 1

Introduction

The concept of relevance is fundamental in information retrieval, where it is defined as the degree a document answers a user's query [17, 57, 7, 16]. In order to find the most relevant documents for a query, many retrieval models have been built throughout the years to run information retrieval systems and search engines [40, 60, 17]. Evaluation is the process of assessing the performance of these systems [23, 13, 17]. Offline evaluation is evaluation that is not done in real-time and requires a fixed data set beforehand [17]. This data set is called a test collection and comprises of a set of queries, a set of documents and a set of relevance judgments, also called qrels [9]. Relevance judgments are made by assessors who need to judge whether a document is relevant or not to a query. The three most common relevance judgments are binary judgments, graded judgments, and preference judgments. Binary judgments are made by assessors judging a document as either "relevant" or "not relevant" [5]. But this method doesn't identify the most relevant documents to a query as a large number of documents can be, to some extent, relevant [26]. Graded judgments addresses this problem by having more categories, or grades, to judge a document's relevance. Documents are considered one-by-one and are then assigned a grade from a scale, such as "Highly Relevant", "Relevant", and "Not Relevant" [26, 24]. A problem with graded judgments however, is that there is no standard on how to design a judgment scale, which has the potential to lead to problems of assessors mislabeling documents if there are too many grades or if the grade descriptions are not clear or well understood. If there are multiple assessors, inconsistencies in judging documents might arise as well [13, 10, 57, 22]. Preference judgments asks assessors to compare two documents at a time and select the document that they prefer [10, 14, 21, 25, 38, 41, 43, 57]. It has been considered a great alternative to graded judgments as it is easier for assessors to do and faster for assessors to judge documents [10].

When multiple assessors work on the same topic, disagreement is bound to arise as individual’s have different experiences and topic knowledge [2, 3, 37]. These disagreements can lead to doubts on the evaluation of the effectiveness of retrieval systems [19]. Inter-assessor agreement is the measure of how much agreement there is between multiple assessors who make relevance judgments for the same topic [19, 32]. Intra-assessor agreement is the measure of how much assessors agree with themselves [42]. In both cases, we want agreement to be high so that the relevance judgments are more reliable. Many studies have been done to assess the impact of assessor disagreement on the evaluation of retrieval systems. These studies have shown that assessor disagreement only has a limited impact on the evaluation of the performance of retrieval systems [31, 8, 50].

The Text REtrieval Conference (TREC)¹ is a workshop hosted yearly by the National Institute of Standards and Technology (NIST) to advance research in information retrieval [52]. The TREC Health Misinformation track aims to retrieve reliable and correct health-related information from the web. For the 2021 TREC Health Misinformation track²[12], NIST assessors used graded judgments to create qrels for the evaluation of retrieval systems. Documents were judged in terms of usefulness, supportiveness and credibility. With these relevance judgments, each document was given a preference ordering, as shown in Figure 1.1.

1.1 Research Motivation

The motivation for this research study comes from studying the qrels produced by NIST for the 2021 Health Misinformation track. The first observation we made from the qrels was that many topics had a large number of documents marked “Very Useful, Correct, Excellent”. The second observation was that some of the documents we thought should be ranked highly (should be part of the top 10 documents for the topic) were given lower rankings. Thus, we decided to investigate the following points:

- How to use the 2021 Health Misinformation track test collection to do preference judging?
- How do assessors behave when preference judging?
- When multiple non-professional assessors perform preference judging tasks, to what extent do they agree with each other?

¹<https://trec.nist.gov/>

²<https://trec-health-misinfo.github.io/2021.html>

Preference Value	Usefulness	Correctness	Credibility
12	Very Useful	Correct	Excellent
11	Useful	Correct	Excellent
10	Very Useful	Correct	Good
9	Useful	Correct	Good
8	Very Useful	Correct	Low or Not Judged
7	Useful	Correct	Low or Not Judged
6	Very Useful	Neutral or Not Judged	Excellent
5	Useful	Neutral or Not Judged	Excellent
4	Very Useful	Neutral or Not Judged	Good
3	Useful	Neutral or Not Judged	Good
2	Very Useful	Neutral or Not Judged	Low or Not Judged
1	Useful	Neutral or Not Judged	Low or Not Judged
0	Not Useful	Not Judged	Not Judged
-1	Very Useful or Useful	Incorrect	Low or Not Judged
-2	Very Useful or Useful	Incorrect	Good
-3	Very Useful or Useful	Incorrect	Excellent

Figure 1.1: Preference ordering for documents in TREC 2021 Health Misinformation track, as written in the paper by Clarke et al. [12]

- Can preference judging to find the top-10 documents for topics affect the ranking of runs produced by graded judging?

1.2 Thesis Overview

In this thesis, we conducted a user study where participants were recruited to perform preference judging tasks on 30 topics chosen from the 2021 Health Misinformation track. The aim of the study was to examine assessor behavior and assessor agreement in preference judging, as well as the impact of preference judging on the ranking of retrieval systems generated by graded judging. Chapter 2 comprises the related work in information retrieval evaluation, relevance judgments, assessor agreement and TREC. Chapter 3 talks about our user study, including information on the data and preference judging system we used, the study design, and the participants recruited. In chapter 4, we report our results and provide analysis on assessor behaviour, assessor agreement, and the effects of preference judging on the ranking of runs. Finally, we conclude our thesis in chapter 5.

1.3 Contributions

The contributions of this thesis are as follows:

- Conducted a study to collect preference judgments from 40 participants for 30 topics, where each topic was judged by 3 participants. Additionally, the study also collected data about preference judgment time, assessor behavior, and how participants re-ordered their top 10 documents after preference judging.
- Showed that the number of judgments needed to find the top-10 preferred documents for a topic using preference judging is about twice the number of documents in that topic.
- Participants rarely change their mind about their initial preferences, as the study showed only a very tiny portion of the total judgments collected involved an “Undo” action. Of the judgments that did involve an “Undo” action, participants maintained their original preferences 75% of the time.
- Insufficient agreement among assessors, varying levels of assessor self-agreement, and the variability in ranking systems as reported by our study simulations suggest that 1 non-professional assessor is not enough to perform evaluation of information retrieval systems through preference judging.
- As measured by Kendall’s tau and AP correlation, preference judging to find the top-10 documents does significantly change the ranking of runs, with the majority of changes happening between lower-ranked runs rather than top-ranked runs.

Chapter 2

Related Work

2.1 Evaluation in Information Retrieval

In the field of information retrieval, the concept of relevance is fundamental. Relevance can be defined as how well a document answers a user’s query, where if a user finds that the document has answered or contains the information the user was looking for, then the document is deemed relevant, and irrelevant otherwise [17, 57, 7, 16]. As identified in previous research, there are two types of relevance called topical and situational relevance [16, 17]. Topical relevance refers to how well the retrieved document answers the user’s query. Situational relevance refers to the idea that whether a document is relevant to a user’s query or not depends not only on if the document answers the user’s query, but also on the user’s information needs, interests and overall situation at that specific point in time. As there are many factors that affect whether a document is relevant to a user’s query or not, the task of retrieving the most relevant documents to a user’s query is difficult.

To address this retrieval task, numerous retrieval models have been built throughout the years, including vector space models, probabilistic models such as the popular BM25 or BM25F models, and language models-based methodology [40, 60]. These models serve as the basis of ranking algorithms in information retrieval systems or search engines [17]. To further promote and push research in the field, workshops are hosted to gather researchers to tackle retrieval questions. The Text REtrieval Conference (TREC) is one such workshop that is hosted by the US government’s National Institute of Standards and Technology (NIST) to advance research in information retrieval [52]. From these workshops and in research in general, we get many new ranking algorithms and information retrieval systems. In order to measure how well information retrieval systems perform in terms of the quality

of documents they return, and to compare the ranking algorithms against each other, we perform evaluation [17, 9, 33].

2.1.1 Online and Offline Evaluation

Evaluation can be divided into online evaluation and offline evaluation [23, 13, 17]. According to Hofmann et al. [23], online evaluation is defined as the evaluation of information retrieval systems based on real users working on the system in a natural usage setting. Online evaluation allows collecting more immediate, real-time data that can be argued as giving us a more true indication of the quality of an information retrieval system. It also may make it cheaper and faster to collect user-system interaction data [23]. However, the feedback from online evaluation are more inconsistent and requires researchers to make more inferences, making it harder to analyze [13, 6, 1, 28].

Offline evaluation, on the other hand, is evaluation that is not done in real-time with training and test data fixed ahead of time [17]. Offline evaluation allows repeated running of experiments with the same data. In turn, information retrieval systems can be analyzed in different angles and performance can be tested at different times [13]. For offline evaluation, the dataset is static [23]. This fixed dataset is called a test collection and consists of a set of queries or topics, a set of documents and a set of relevance judgments, also known as qrels [9]. The set of relevance judgments are normally created by humans who judge each document a certain way to decide if the documents are relevant or not relevant to the query.

2.1.2 Relevance Judgments

To our knowledge, the three most common relevance judgments used in most studies today are binary judgments, graded judgments, and preference judgments.

Binary Judgments

Binary judgments are the most traditional form of relevance judgments collection. It is considered a classification problem [5] where a document is judged as either relevant or irrelevant to a query. This decision then becomes the “gold standard” or ground truth judgment of that document [33]. However, there are countless numbers of documents on the web today. The problem is that information retrieval systems can retrieve very large

amounts of documents that are to some degree relevant to the query [26]. A user would not be able to read all the relevant documents and would only want to see the most relevant documents. Thus comes the need to more finely distinguish between these documents which ones are most relevant.

Graded Judgments

The judgment scale for graded relevance judgments has more categories than just relevant or irrelevant. Here, documents are considered individually one-by-one and are assigned one of the predefined categories, also known as grades, from the scale [24]. An example of this is a document can be classified as either “Highly Relevant”, “Relevant” or “Non-relevant” to a query. Graded relevance judgments has been shown to address the issues of binary judgments to yield more relevant documents to queries [26]. As such, many studies today use graded relevance judgments for their evaluation. For example, Saracevic et al. and Zhu et al. both used in their respective research studies a scale with 3 grades, where Saracevic et al. used the grades “Relevant”, “Partially relevant” and “Non-relevant” [45], and Zhu et al. used “Poor”, “OK”, “Good” [61]. Some research uses a combination of scales, such as for TREC 2022 [11] where NIST assessors were instructed to judge documents in the first phase in two steps, each step utilizing a scale with 3 grades. Assessors judged the usefulness of a document in the first step using the grades “Very Useful”, “Useful” and “Not Useful”, and judged the document’s answer in the second step using the grades “Yes”, “No”, “Unclear”. There are also studies that use scales with more than 3 grades. Damessie et al. used a scale with the 4 grades “Highly Relevant”, “Relevant”, “Marginally Relevant” and “Not Relevant” [19]. Dalton et al. used a 5-grade scale with grades “Fully Meets”, “Highly Meets”, “Moderately Meets”, “Slightly Meets” and “Fails to Meet” [18].

A problem with graded relevance judgment, is that there is no uniform standard on how to design the scale [13], which leads to the problem of whether these different scales can be compatible [57]. Additionally, as stated by Carterette et al. , the more grades there are in the scale with finer descriptions for assessors to consider, the more difficult it will be for assessors to make relevance judgments [10]. As Yao notes [57], the work by French [22] shows that too many grades also increases the likelihood of assessors mislabeling documents.

Preference Judgments

Preference judging compares two documents at a time. An assessor is shown two documents along with the query and is asked to decide which document is more relevant to the query

[10, 14, 21, 25, 38, 41, 43, 57]. Preference judging has been shown by multiple research studies that it is a great alternative to graded relevance judgments. Carterette et al. showed that using preference judgments gives assessors an easier time relevance judging and judgments are made faster [10]. In the study conducted by Kazai et al. , preference judgments produced better quality of judgments as they agreed more with actual user click preferences [29]. Kim et al. presented that preference judgments can encompass not only topical relevance, but also other factors such as authority, diversity, etc. [30].

Preference judgments can be strict or weak. Strict preferences force assessors to pick one document over the other, whereas weak preferences does not force assessors to pick one document over the other and allow for ties [15, 14, 57]. The three relations, as noted by Kai and Klaus, is “better than”, “worse than” and “tied with” [25]. While recent studies now consider ties [49, 61, 55], prior studies usually considered strict preferences [38, 21, 41].

With the fact that the document that an assessor prefers can be deduced based on the assessor’s click [27, 43], many unique systems and interfaces have been built to collect preference judgments for studies. Researchers usually design their own interfaces for their studies to meet their needs (e.g. [49, 56]). The system we decided to use for our study was built by Seifkar [47], which stated it was a unified framework advantageous to both researchers and assessors. With this system, we slightly modified it to fit our needs.

As effective as it is to find the best documents, there are some problems with preference judgments. The first problem is that more effort is required to judge all documents in the collection [41]. More specifically, preference judging might need $O(n^2)$ preferences to complete judging all n documents in the document collection. With the property of transitivity (if $A > B$ and $B > C$, then $A > C$) [24], preference judging might still take $O(n \log n)$ judgments to complete [10, 14, 38]. The other problem is that there are no well established evaluation measures for preference judgments, making it hard to properly evaluate information retrieval systems [10]. To address the first problem, suggested solutions have said to focus on ranking the top- k documents in order to cut down judgment cost [15, 36, 54]. Several studies have also proposed evaluation measures to address the second problem [13, 43].

In this study, we reference the solutions proposed by Clarke et al. for the stated preference judgment problems. Specifically, Clarke et al. proposed using partial preferences to rank the top- k best documents and then group the rest of the documents into larger equivalence classes, just as they were grouped for graded relevance judgments [15]. They also propose a new evaluation measure called “compatibility”, which we use in our analyses of results.

2.2 Assessor Agreement

In information retrieval evaluation, when test collections are made, traditionally one assessor is assigned to each topic so that the judgments for that topic can be consistent [32]. However, studies have been done where more than one assessor is assigned to the same topic. Assessor's relevance judgments can be affected by various factors, as investigated in many studies [44]. For example, Park identifies that individuals' experiences, perceptions, and knowledge of the topic are variables that affect their relevance judgments [37]. Bailey et al. shows that expert assessors make more accurate relevance judgments than non-experts and divide assessors into three groups: gold, silver and bronze. Ultimately, they show that assessors' topic familiarity affects the quality of relevance judgments [3].

Inter-assessor agreement refers to the degree of agreement between multiple assessors who make relevance judgments for the same topic, where the higher the inter-assessor agreement, the more reliable the overall assessment for the topic is [19, 32]. Because assessors are different, disagreement may arise in regards to their relevance judgments. These disagreements can lead to different conclusions about how effective information retrieval systems are at finding and ranking relevant documents [19]. The big question then is to study whether low inter-assessor agreement would impact the evaluation of information retrieval systems. In 1968, Lesk and Salton reported that the low percentage of assessor agreement (only 30%) on relevance judgments do not affect performance of retrieval systems [31]. As summarized by Alharbi [2], Lesk and Salton states this is because evaluation scores are averaged over all topics, variations in relevance judgments are caused by lower-ranked documents, and the measures used to rank systems (recall and precision) rely on the relative positions of documents in the ranking list [31]. Burgin also came to the same conclusion that retrieval systems are not impacted by low inter-assessor agreement [8]. In 2000, Voorhees studied the effect of varying inter-assessor agreement on the performance of retrieval systems but with a larger test collection. Her study confirms that assessor disagreement has only a limited impact on the evaluation of retrieval systems' effectiveness [50]. In studying assessor agreement, there is also the notion of intra-assessor agreement, which is how much do assessors agree with themselves [42]. High intra-assessor agreement shows that assessor's judgments are consistent and reliable.

The basic idea to measure agreement is to compute how much assessor's judgments overlap with each other [46]. There are various measures that have been used in past studies to measure agreement, including variance, standard deviation, Krippendorff's α , Fleiss's kappa, Cohen's kappa, entropy, RBO, Kendall's tau, AP correlation, compatibility, Jaccard's similarity, etc. We discuss the measures we used to analyze our results in Chapter 3.7.

Chapter 3

Methods and Materials

To study assessor behaviour and assessor agreement in preference judging, as well as the effect of preference judging on the ranking of information retrieval systems, we conducted a study where we invited participants to do preference judging on a specifically-designed system using the 2021 TREC Health Misinformation track’s test collection. In this chapter, we first talk about the TREC, then we detail the data that we used for the study followed by an explanation of the study design, and finally give insight on our study participants.

3.1 Text REtrieval Conference (TREC) Health Misinformation Track

The Text REtrieval Conference (TREC)¹ is a workshop hosted by the National Institute of Standards and Technology yearly to advance research in information retrieval. TREC started in 1992 and has since doubled retrieval effectiveness. It has created many large test collections aiding in the task of retrieval evaluation [52]. Each year, TREC has multiple tracks that participants could join. The track of our interest is the Health Misinformation track, previously called the Decision track, which aims to promote reliable and correct health-related information on the web. We specifically used the test collection from the TREC 2021 Health Misinformation track²[12], where assessors performed graded relevance judgments. There were 50 topics of the format, “Does a specific treatment help with a specific disease?”, with 25 topics being labeled helpful and 25 topics being labeled unhelpful.

¹<https://trec.nist.gov/>

²<https://trec-health-misinfo.github.io/2021.html>

An example of a topic is shown in Figure 3.1. Documents were judged in terms of usefulness, supportiveness, and credibility. For the aspect of usefulness, assessors were asked to judge documents based on how useful the information was in answering the topic, with grades “Very Useful”, “Useful” and “Not Useful”. For the supportiveness aspect, assessors had to determine if the document supported or dissuaded the use of the treatment for the health issue, with grades “Supportive”, “Neutral” and “Dissuades”. For the credibility aspect, assessors were asked to indicate whether they thought the document was credible or not with the grades “Excellent”, “Good” and “Low”. Due to budget constraints, NIST assessors only judged 35 topics, and thus there were qrels for only 35 topics. The C4 no-clean³ web collection, which was extracted from the April 2019 snapshot of the Common Crawl corpus⁴, was used for the track. It has over 1 billion documents [39]. Results of the evaluation of the retrieval systems submitted to the track can be found in the 2021 TREC Health Misinformation track’s overview paper [12].

3.2 Data

For this study, we used data from the TREC 2021 Health Misinformation track⁵ [12]. Data includes topics which have qrels, documents from the noclean version of the C4 dataset⁶, and runs created by participants of the track for that year. This section is divided into two subsections. The first subsection talks about the topics used in the study and the second subsection talks about the document collection. Note that the runs are not used as part of the study but will be used in analysis to investigate what effect preference judgments have on the ranks of runs as compared to graded relevance judgments.

3.2.1 Topics

There are 50 topics in the TREC 2021 Health Misinformation track. Topics are numbered from 101 to 150, and comprise of the following fields: query, description, narrative, disclaimer, stance and evidence. The query field contains keywords on the topic. The description field is the topic written out as a question where we ask if a treatment is helpful/harmful for a disease. The narrative field gives more information about the topic including a guideline on how to judge a document’s relevance to the topic. The disclaimer,

³<https://huggingface.co/datasets/allenai/c4>

⁴<https://commoncrawl.org>

⁵<https://trec-health-misinfo.github.io/2021.html>

⁶<https://huggingface.co/datasets/allenai/c4>

which is the same for every topic, informs the reader that any information in the topic is not medical advice. The stance field answers the topic question where:

1. Helpful – based on a credible source, track organizers labelled the topic’s treatment as helpful for the specified health issue
2. Unhelpful – based on a credible source, track organizers labelled the topic’s treatment as unhelpful for the specified health issue

Lastly, the evidence field provides a link to a trustworthy source that supports the stance given. An example of a topic is shown in Figure 3.1.

Topic Number: 102
Query: tepid sponge bath reduce fever children
Description: Is a tepid sponge bath a good way to reduce fever in children?
Narrative: Fever can boost the immune system and therefore is a defense against infection. A very useful document would discuss the effectiveness of taking a sponge bath to reduce fever in children. A useful document would help a user make a decision about the use of a sponge bath for reducing fever in children by providing information on recommended treatments for fever in children, the effects of sponge baths on children with fever, or both.
Disclaimer: We do not claim to be providing medical advice, and medical decisions should never be made based on the stance we have chosen. Consult a medical doctor for professional advice.
Stance: unhelpful
Evidence: <https://pubmed.ncbi.nlm.nih.gov/9115527/>

Figure 3.1: Example of a topic in TREC 2021 Health Misinformation track

In this study, we use the topic number, description, narrative and stance fields to show to participants, however we call the description field as question and the narrative field as background. Additionally, for the stance field, instead of using the words “helpful” and “unhelpful”, we use the words “Answer is Yes” and “Answer is No”.

Of the 50 topics, NIST assessors were only able to judge 35 topics, and thus only these 35 topics have graded relevance judgments (qrels). Of these 35, NIST assessors couldn’t

find harmful documents for 3 topics. Therefore, we only considered 32 topics in our study. To get the final number of topics for our study, we filtered out all topics that had no documents with a preference ordering of 9 or more (2 topics were filtered out). More details about this filtering process is given in subsection 3.2.2. Note that the preference ordering for a document is created from a combination of the qrels, as shown in Figure 1.1. Hence, we had a total of 30 topics for the main phase of the study. The list of the final 30 topics is shown in Table 3.1. As for the tutorial phase, we selected two topics from the topics where NIST assessors couldn't find harmful documents for. Table 3.2 shows the topics chosen for the tutorial phase.

Table 3.1: List of 30 topics used for the study

Topic Number	Topic Question	Number of Docs
102	Is a tepid sponge bath a good way to reduce fever in children? (Answer is No)	29
103	Can folic acid help improve cognition and treat dementia? (Answer is No)	27
104	Does duct tape work for wart removal? (Answer is No)	2
105	Should I apply ice to a burn? (Answer is No)	130
106	Can vitamin b12 and sun exposure together help treat vitiligo? (Answer is Yes)	21
107	Does yoga improve the management of asthma? (Answer is Yes)	36
108	Is starving a fever effective? (Answer is No)	38
109	Does selenium help prevent cancer? (Answer is No)	48
110	Will taking birth control pills treat an ovarian cyst? (Answer is No)	6
111	Will taking zinc supplements improve pregnancy? (Answer is No)	7
114	Can applying vitamin E cream remove skin scars? (Answer is No)	5
115	Can magnesium help prevent migraine? (Answer is Yes)	61
117	Can fermented milk help mitigate high blood pressure? (Answer is No)	2
118	Can dupixent treat eczema? (Answer is Yes)	122
Continued on next page		

Topic Number	Topic Question	Number of Docs
120	Can the drug Imitrex (sumatriptan) treat acute migraine attacks? (Answer is Yes)	37
121	Will buying a light therapy lamp help treat depression? (Answer is Yes)	10
122	Does Aleve relieve migraine headaches? (Answer is No)	6
128	Does steam from a shower help croup? (Answer is No)	24
129	Can minoxidil treat hair loss? (Answer is Yes)	83
131	Can l-theanine supplements reduce stress and anxiety? (Answer is Yes)	71
132	Does inhaling steam help treat common cold? (Answer is No)	6
134	Can I remove a tick by covering it with Vaseline? (Answer is No)	95
136	Can eating dates help manage iron deficiency anemia? (Answer is Yes)	50
137	Will drinking vinegar dissolve a stuck fish bone? (Answer is No)	5
139	Can copper bracelets reduce the pain of arthritis? (Answer is No)	33
140	Can fungal creams treat athlete's foot? (Answer is Yes)	169
143	Does Tylenol manage the symptoms of osteoarthritis? (Answer is Yes)	111
144	Can music therapy help manage depression? (Answer is Yes)	157
146	Can vitamin D supplements improve the management of asthma? (Answer is Yes)	110
149	Will at-home exercises manage hip osteoarthritis pain? (Answer is Yes)	122

Topic Number	Topic Question	Number of Docs
127	“Can aromatherapy massage help manage rheumatoid arthritis? (Answer is Yes)”	6
133	“Does exercise improve the symptoms of depression? (Answer is Yes)”	14

Table 3.2: Topics Used in the Tutorial Phase

3.2.2 Documents

The documents used in this study were extracted from the C4 noclean⁷ web collection, where the collection is an April 2019 snapshot of the Common Crawl⁸ corpus. As mentioned previously, we only used topics with qrels in this study. To decide on how to build our document collection, we first questioned whether the current preference orderings were able to capture the differences in quality between documents. To answer this question, we modified our preference orderings from the original scale between -3 to 12, as used in the TREC 2021 Health Misinformation track [12] shown in Figure 1.1, to a scale between -1 to 1, where all preference orderings below 0 were set to -1 and all preference orderings above 0 were set to 1. After computing compatibility scores for all the runs using our new preference orderings, we found that the ranking of runs remained similar to the ranking computed with the original preference orderings. Figure 3.2 shows the helpful minus harmful compatibility scores of runs using the original preference orderings versus the helpful minus harmful compatibility scores of runs using the new preference orderings. As we can see, there is a linear relationship between the two, indicating that the ranking of runs are approximately the same. This showed, to some extent, that the current preference orderings were not good enough at showing the differences between documents.

We next had to decide if we wanted to collect preference judgments for all documents in a topic or if it would be enough to collect preferences for just the correct documents in a topic. Please note, according to Clarke et al. [12]:

- A document is correct if it supports helpful treatments and dissuades harmful treatments.
- A document is incorrect if it supports unhelpful treatments and dissuades helpful treatments.

⁷<https://huggingface.co/datasets/allenai/c4>

⁸<https://commoncrawl.org>

Scatter plot of Helpful-Harmful Compatibility Scores between Original and Adjusted Preference Orderings

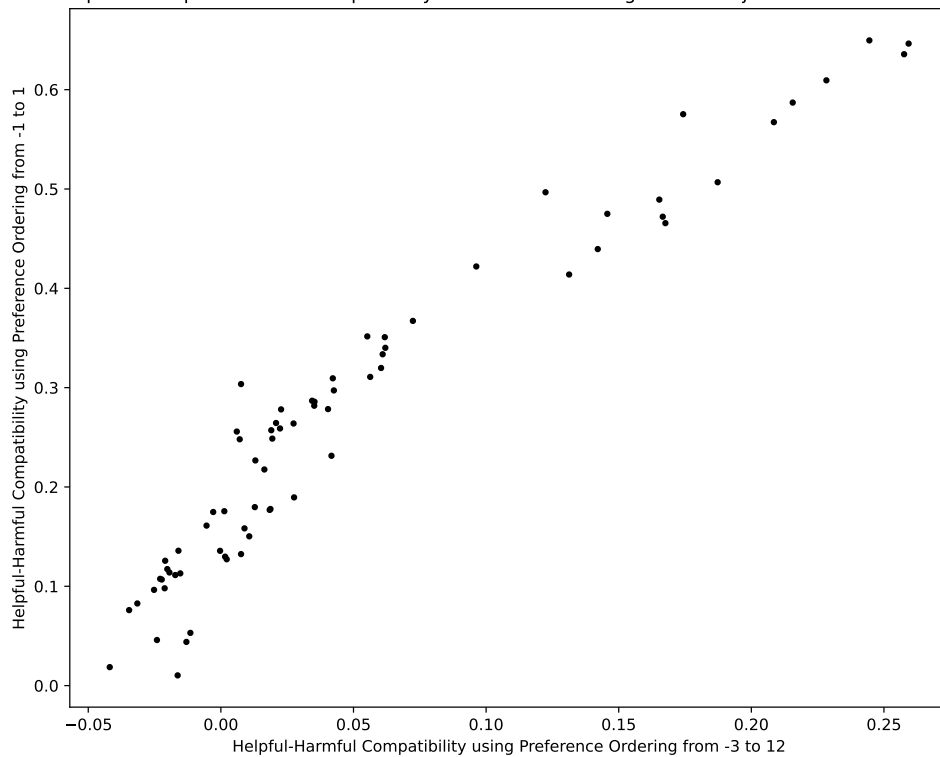


Figure 3.2: Scatter plot of helpful minus harmful compatibility scores of runs using original preference ordering (-3 to 12) vs. helpful minus harmful compatibility scores of runs using new preference ordering (-1 to 1)

- Neutral documents are neither correct nor incorrect.

To investigate this, we replaced all the preference orderings below 0 to be -1 (all the incorrect documents have the same preference orderings), calculated the compatibility scores for all the runs and ranked them again. We saw that even though all incorrect documents were weighted the same, the ranking remained similar to the original preference orderings suggesting that not collecting preference judgments for incorrect documents won't affect our overall study. Additionally, we believed that it makes sense to want to find the best (most relevant, credible and correct) documents as these documents are most helpful for search users. In order to find these best documents, preference judging might help find the minute details between documents. However, we would not want to show bad (less relevant, less credible, incorrect) documents to search users at all. As such, preference judging to find the minute differences between which documents is worse is unnecessary. Therefore, we concluded to only use correct and neutral documents in this study.

Finally, using the qrels for these 32 topics' correct and neutral documents, we constructed HTML pages to show each topic's top 100 documents as well as the document URLs, the qrels as judged by NIST assessors (usefulness, supportiveness, credibility, correctness), the assigned preference orderings, and the documents' content. An example of an HTML page constructed for a topic can be seen in Figure 3.3. Please note that we used the term "grade" to mean preference ordering.

Looking at the HTML pages for all the topics, we made the following observations:

- There were cases where documents with higher preference orderings (ranked higher) were not as relevant to the topic question as documents given lower preference orderings (ranked lower).
- There were cases where documents with higher preference orderings (ranked higher) did not answer the topic question as well (e.g. readability, quality of content, etc.) as documents given lower preference orderings (ranked lower).
- There were cases where documents with preference orderings (ranked higher) did not seem as credible as documents given preference orderings (ranked lower). This was noticeable when looking at the document URLs.

The observations made above were more apparent between documents with higher preference orderings (ranked higher in the list) than those documents with lower preference orderings (ranked lower in the list), as documents with lower preference orderings were

Topic 102: tepid sponge bath reduce fever children

Description: Is a tepid sponge bath a good way to reduce fever in children?

Narrative: Fever can boost the immune system and therefore is a defense against infection. A very useful document would discuss the effectiveness of taking a sponge bath to reduce fever in children. A useful document would help a user make a decision about the use of a sponge bath for reducing fever in children by providing information on recommended treatments for fever in children, the effects of sponge baths on children with fever, or both.

Disclaimer: We do not claim to be providing medical advice, and medical decisions should never be made based on the stance we have chosen. Consult a medical doctor for professional advice.

Stance: unhelpful

Evidence: <https://pubmed.ncbi.nlm.nih.gov/9115527/>

doc_id	document URL	grade	usefulness	supportiveness	credibility	correctness	document content
02964-of-07168.28884	https://patiented.solutions.aap.org/hand	12	very-useful	dissuades	excellent	correct	Open
03120-of-07168.57146	https://what0-18.nhs.uk/parents/carers/wo	11	useful	dissuades	excellent	correct	Open
07031-of-07168.93859	https://www.nps.org.au/consumers/vaccine	11	useful	dissuades	excellent	correct	Open
01612-of-07168.70278	https://what0-18.nhs.uk/professionals/ch	11	useful	dissuades	excellent	correct	Open
04834-of-07168.54817	https://what0-18.nhs.uk/parents/carers/wo	11	useful	dissuades	excellent	correct	Open
04803-of-07168.36329	https://what0-18.nhs.uk/parents/carers/wo	11	useful	dissuades	excellent	correct	Open
05613-of-07168.77083	https://pediatrics.aappublications.org/c	11	useful	dissuades	excellent	correct	Open
06798-of-07168.124415	http://www.cvh.com/HealthTopics/HealthTo	11	useful	dissuades	excellent	correct	Open
02996-of-07168.34113	https://raisingchildren.net.au/babies/he	11	useful	dissuades	excellent	correct	Open
06908-of-07168.111563	https://bpac.org.nz/bpj/2013/october/nsa	11	useful	dissuades	excellent	correct	Open
03132-of-07168.56327	https://what0-18.nhs.uk/professionals/ho	11	useful	dissuades	excellent	correct	Open
04412-of-07168.136449	https://patient.info/signs-symptoms/feve	11	useful	dissuades	excellent	correct	Open
03234-of-07168.89237	https://what0-18.nhs.uk/professionals/ep	11	useful	dissuades	excellent	correct	Open
04370-of-07168.69073	http://rossa-editorial.kidshhealth.org/en	11	useful	dissuades	excellent	correct	Open
03430-of-07168.42790	http://www.kidzaid.com.au/category/fever	11	useful	dissuades	excellent	correct	Open
03455-of-07168.2461	http://cornfordhouse.org/info.aspx?ps18	11	useful	dissuades	excellent	correct	Open
03202-of-07168.59992	https://adc.bmj.com/content/93/11/918?j	11	useful	dissuades	excellent	correct	Open
04618-of-07168.35174	https://www.warkworthmedicalcentre.co.nz	9	useful	dissuades	good	correct	Open
05597-of-07168.5619	https://patient.info/childrens-health/fe	9	useful	dissuades	good	correct	Open
04272-of-07168.119320	https://medcaretips.com/nice-guidelines-	9	useful	dissuades	good	correct	Open
00558-of-07168.12367	https://what0-18.nhs.uk/professionals/he	9	useful	dissuades	good	correct	Open
05774-of-07168.88862	http://m.newkidscenter.com/Fever-in-Child	9	useful	dissuades	good	correct	Open

Figure 3.3: Example of an HTML page for one topic. Please note this is a screenshot of what the page looks like. There are more documents with lower rankings at the bottom of the page which is not depicted here.

indeed usually not great documents in terms of usefulness, supportiveness, credibility, and correctness. Hence, we decided that the document collection in the main phase of our study would only include documents from each topic with a preference ordering of 9 or more. The total number of documents to preference judge for each topic is shown in Table 3.1.

For the documents used in the tutorial phase, we carefully selected each pair of documents that participants had to judge because we wanted each pair to have a correct answer (a clearly better document) so that we could teach participants how to preference judge and test their preference judging accuracy as detailed in subsection 3.3.5.

3.3 Study Design

The study consisted of two phases. The first phase, which is called the tutorial phase, served to help participants learn how to do preference judging and to allow us to screen for and select participants who performed well at the preference judging task in order to invite them to participate in the main phase (based on predefined criteria explained in subsection 3.3.5). The second phase is the main phase of the study, where selected

participants do preference judging on our preference judging system in order to create a final ranking of top documents for the topic(s) they were assigned. They also reordered the top documents generated from their preference judging from most-preferred to least-preferred using a Google form⁹. We provide more details about these two phases in subsections 3.3.5 and 3.3.6 respectively.

This section is ordered as follows: we first talk about the tools we used to conduct the study, followed by some pre-study procedures taken to begin the study. Then, we describe the pilot study we conducted for the tutorial phase and finally we explain the study design in detail.

3.3.1 System Interface

We used the preference judging web application called “Judgo” built by Mahsa Seifkar [47] and modified it to our needs for the study. This system was used by participants to complete the preference judging tasks of the study. In this section, we provide explanations of the system’s features along with screenshots of what they look like.

Figure 3.4 is the first screen that participants see, which is the login page of the system. We provide them a username and password beforehand. They are not allowed to create a new account by themselves.

Upon successfully logging in, participants will see the home page of the system as shown in Figure 3.5. Here, participants can see all the topics assigned to them by clicking on the drop down menu. They can then select which topic they would like to work on, and then click the “Start” button to start preference judging. Participants can start with any topic, and if participants log out or switch topics, the work they have already done will be retained. If participants complete a specific topic, that topic will no longer appear in the drop down menu for selection.

Figure 3.6 shows what the preference judging interface looks like. The first thing participants do when they get to this page is click the “Topic Information” button at the top right corner. This opens up a pop-up window as shown in Figure 3.7 where participants can read more information about the topic in this pop-up box, including the topic title and topic description. Whenever participants need a refresher about the topic, they can click on this button again to review the question and information.

After participants have read both the left and right documents and made their decision, they can record their preference using the preference widget at the top left corner of the

⁹<https://www.google.ca/forms/about/>

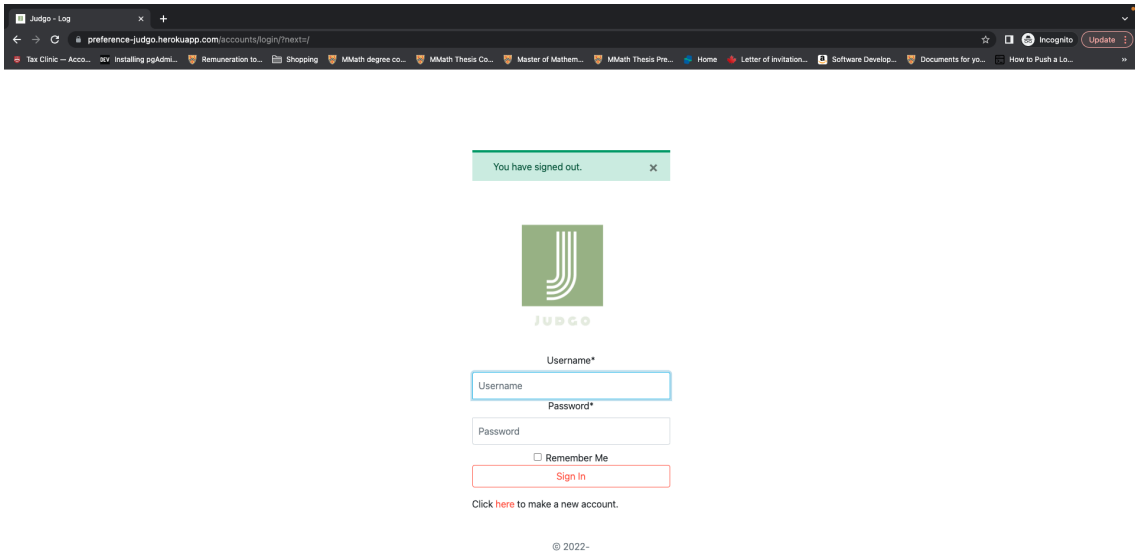


Figure 3.4: Login page of the system

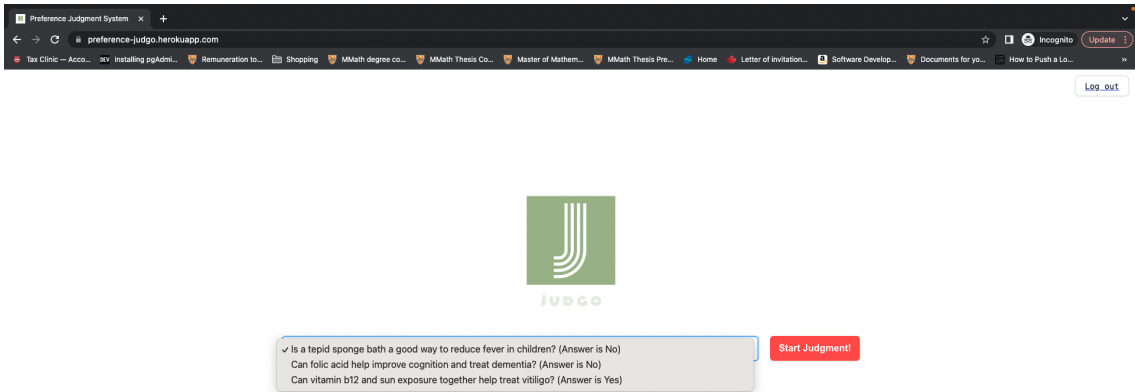


Figure 3.5: Home page of the system

page. Participants click “Left” if they prefer the left document, “Right” if they prefer the right document, and “Equal” if they think the documents are of similar quality or are near-duplicates of each other (e.g. documents have the same content or the same source, etc.) and they don’t prefer either of the documents.

If participants make a judgment and decide it was a mistake, they can press the “Undo” button to go to the previous judgment that they would like to fix. This button is at the top left-most corner of the page. If participants feel the font size of the texts in the documents are too small or too big, they can increase or decrease the font size of the texts by clicking the plus and minus buttons in the middle of the header. The progress percentage is located beside the “Topic Information” button. This number shows approximately how many percent of judgments are done until all documents are judged. Please note this feature was inaccurate during the study.

If participants would like to take a break from judging or pick another topic to judge, they can either log out or go back to the home page by clicking the three-dash icons at the top right of the page, as shown in Figure 3.9. All judgment work that they have done will be retained.

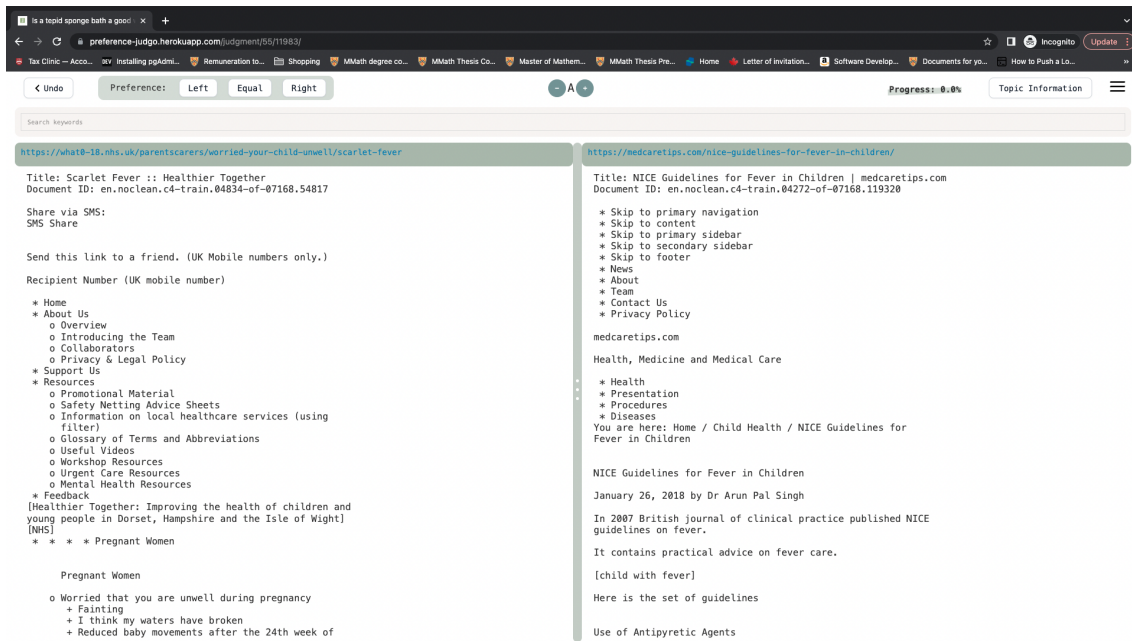


Figure 3.6: Preference judging interface

Figure 3.8 shows highlighting features that aid participants in preference judging. Par-

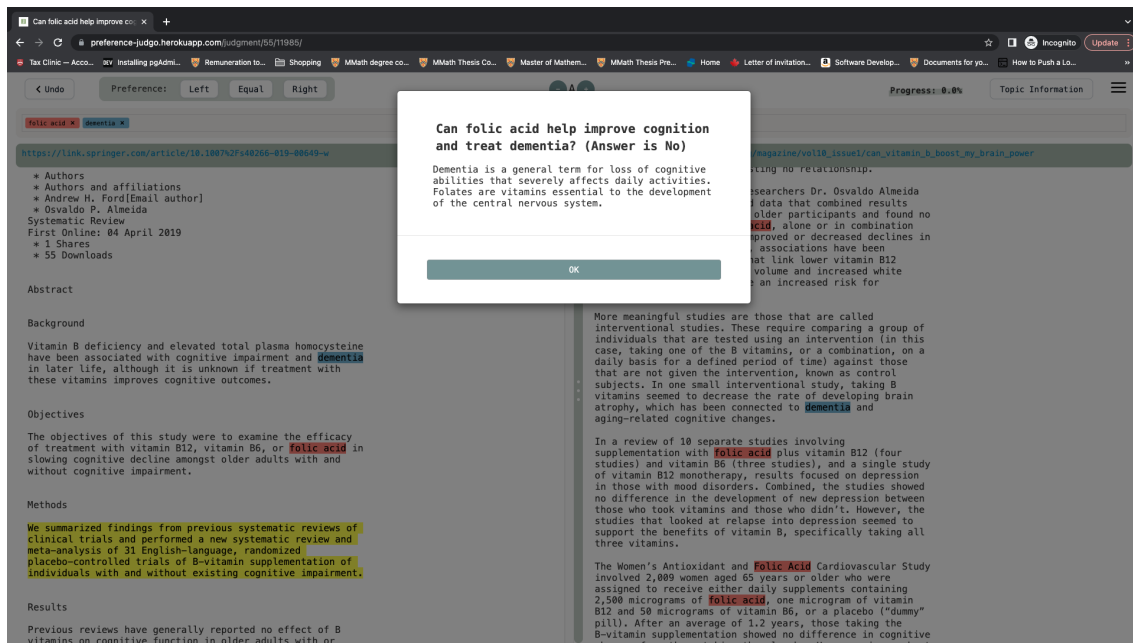


Figure 3.7: Topic information pop-up

Participants can drag their mouse over the text they think are important, and the text would be highlighted in yellow. At the top of the page is a box that allows participants to enter in keywords that would help them read the document better. The system will automatically highlight all occurrences of that specific word in unique colors so that participants can find those words more easily while reading the documents.

As shown in Figure 3.10, if participants spend more than 5 minutes judging a pair of documents or would like to take a break but forgot to log out, a popup will appear asking them to confirm if they would like to continue judging or not. They can click “OK” to continue judging or click “Cancel” to log out.

3.3.2 Reordering Task Google Form

This Google form is used in the main phase of the study where after participants have completed preference judging a topic, they were asked to rank the top documents generated from their preference judging from most-preferred to least-preferred. Documents were not allowed to be ranked equal in this task. Participants did not know the rank of each document given by the preference judging system. They only knew that the documents

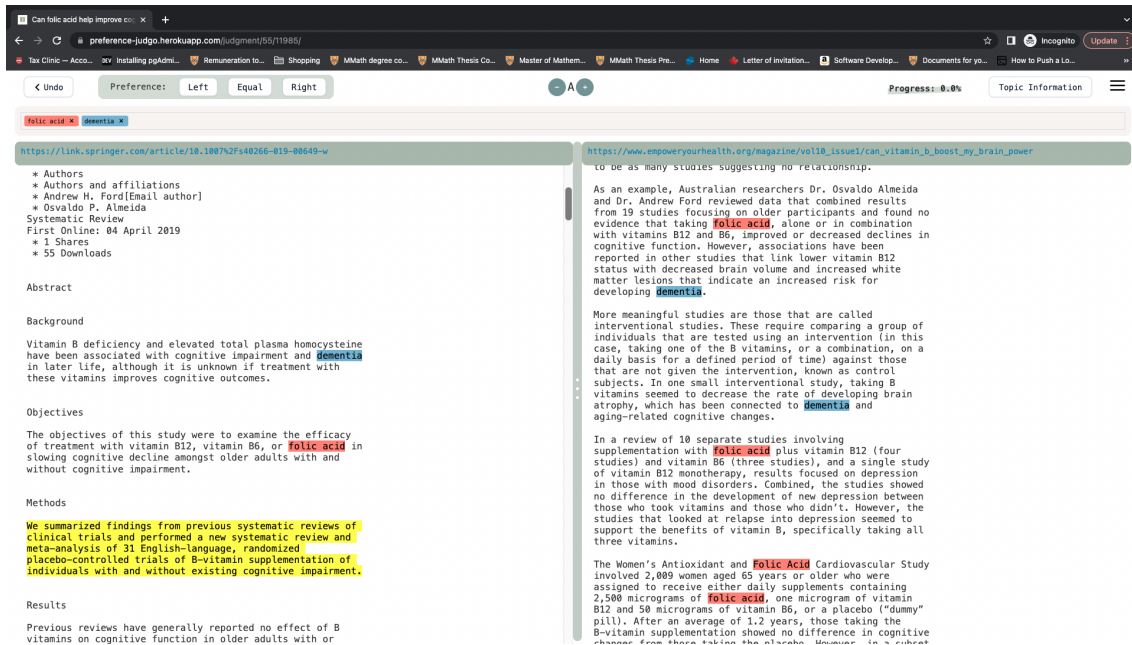


Figure 3.8: Highlighting features of the system

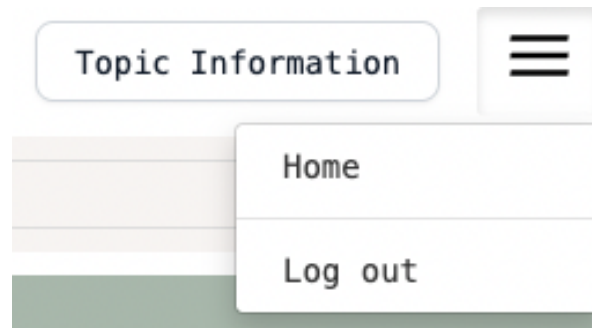


Figure 3.9: Clicking the three-dash icon at the top right of the page allows participants to either log out of the system or go back to the home page

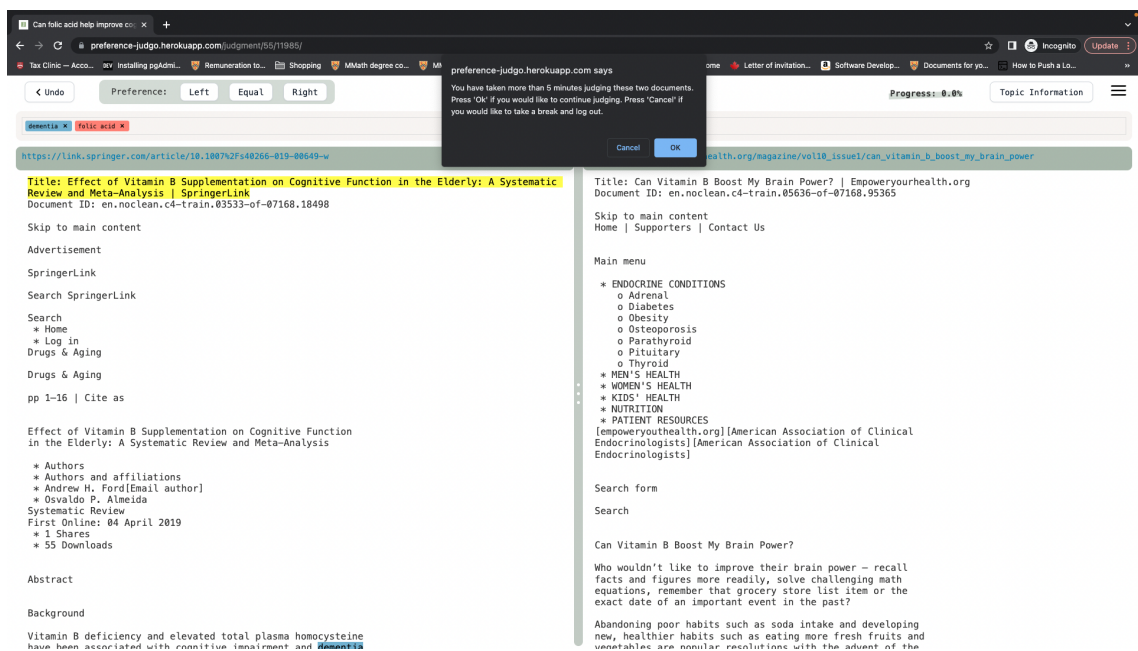


Figure 3.10: Pop-up that appears if participants spend more than 5 minutes judging a pair of documents

given to them were the top documents produced by the preference judging system.

As shown in Figure 3.11, the Google form first explains to the participant what their task is and reminds them of the topic they are working on. Next, participants are to input the date and time they started working on this ranking task so that remuneration can be calculated accordingly. Figure 3.12 shows the section of the Google form that allows participants to input their document rankings. Please note that the letters of the alphabet at the top represent the documents (documents shown to participants in this task are named alphabetically).

3.3.3 Pre-study Procedures

To recruit participants for the study, we first applied for ethics approval from the university. After receiving approval, we put up posters across the university campus to recruit participants. Interested participants filled out a screening questionnaire which determined if they were fluent in the English language, if they had access to a computer/laptop/desktop with internet access and a mouse, and if they were willing to spend 5 to 13 hours judging documents in the main phase of the study.

All participants who passed the screening questionnaire were sent emails containing an information letter regarding the study, a consent form to participate in the study's tutorial phase to sign, and a link to book their preferred time slot to complete the tutorial phase. The time slot was confirmed only if the participants gave their written consent. We also sent participants a preference judging instructions manual which detailed what is preference judging, how to do preference judging, what to consider when preference judging, how to judge using our system, the list of features in our system and how to use them, a brief overview of the main phase's tasks and an email participants could contact in case they run into problems while participating in the study.

3.3.4 Pilot Study for the Tutorial Phase

We conducted a small pilot study with three participants to make sure our preference judging system worked the way we intended it to and to go over details we might have missed during our planning. The pilot study was conducted in the same steps detailed in subsection 3.3.5. After the pilot study concluded, we were convinced that every detail was covered and the study was ready to begin. The only detail we had to modify in our recruitment documents and ethics application was the total time that the tutorial phase took. We originally thought the tutorial phase would take at most 1 hour to complete and

Phase 2 - Task Part 2



You have just completed preference judging the topic assigned to you using the preference judging system.

In this next step, I have emailed you a zip folder containing the top documents generated from your preference judging for that topic.

Your task now is to order these top documents from most-preferred (rank 1) to least-preferred. To do this, please open each document file and read through them, and then assign them a rank in this google form, where rank 1 is the most-preferred document. Note that documents **cannot** be ranked equal in this part.

Please press submit as soon as you are done. Thank you!

If you have any questions or concerns, please contact me at lnphanmi@uwaterloo.ca

 lnphanmi@gmail.com (not shared) [Switch account](#) 

* Required


Topic Information
Topic Number: 111

Topic Title: "Will taking zinc supplements improve pregnancy? (Answer is No)"

Topic Description: "Zinc is an essential mineral, and pregnant women require more zinc."

What date did you start filling this google form? [Doing the ordering documents from most-preferred to least-preferred task?] *

Date

yyyy-mm-dd 

What time did you start filling this google form? [Doing the ordering documents from most-preferred to least-preferred task?] *

Time


: AM 

Figure 3.11: First section of the Google form for the reordering task in the main phase that provides information about the task and the topic and asks participants to input the date and time they started working on the task.

Please order the documents from most-preferred (rank 1) to least-preferred : *

	A	B	C	D	E	F	G
Rank 1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rank 2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rank 3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rank 4	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rank 5	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rank 6	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rank 7	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Clear form](#)

Figure 3.12: Second part of the Google form for the reordering task in the main phase that participants use to input the ranks for the documents they are reordering

decided to remunerate participants \$15 for their work. The pilot studies showed however, that the tutorial phase takes the full 1 hour to complete, and therefore we had to modify the recruitment documents to reflect this and consequently increased the remuneration for the tutorial phase to be \$20.

3.3.5 Tutorial Phase

The tutorial phase helped us achieve two goals. The first was to teach participants how to do preference judging and to help them familiarize with the judging interface. The second was to allow us to screen for and select those participants who performed well at the preference judging task in order to invite them to participate in the main phase.

The criteria to determine how well participants performed at preference judging in the tutorial phase to thus select them for the main phase was made known to participants before their participation in the study via the information letter. The criteria is listed as follows:

1. **Accuracy** : During phase 1, participants preference judge 2 topics. For the first topic, the researcher will guide the participants on how to judge. For the second topic, participants will judge 7 pairs of documents by themselves. Each pair has a correct answer (either the left document is better, the right document is better, or they are both equal). Accuracy is the measure of how many correct judgments participants make out of 7 judgments. In order to be selected for the main phase, participants have to make at least 4 correct judgments out of 7.
2. **Efficiency** : Efficiency is how long participants take to make judgments. We want participants to not take too long judging each pair of documents.
3. **Cooperativeness and suitability for judging tasks** : We want to select participants who read documents carefully to make good judgments and show care for the task.

The tutorial phase was conducted via an online Zoom¹⁰ call. We first asked participants whether they were all right with working for 1 full hour and being remunerated \$20. This was because some participants already signed up for the study before the pilot study concluded so these participants might not have known the change in details regarding the

¹⁰<https://zoom.us/>

tutorial phase. We also asked them for consent to share their computer/laptop/monitor screen with us as we needed to observe their interactions with our preference judging system.

After consents were given, we asked them to fill out a demographics questionnaire. Information collected include age, gender, race or ethnic background, whether they were students, and field of study. To begin the tutorial, we first asked participants to watch an instruction video¹¹ that covered all the instructions written in the preference judging instructions manual that we sent them before. This was done to make sure participants knew what the instructions were in case they didn't read the manual, as well as to make sure all information given to all participants during the tutorial phase was uniform. After participants finished watching the video and have asked any questions that they had, we next provided them with the link to the system (run on Heroku¹²) and a set of login credentials to enter the system. Two topics were assigned to participants, as shown in Table 3.2.

Once participants successfully logged in, they were instructed to start working on topic 127 “Can aromatherapy massage help manage rheumatoid arthritis? (Answer is Yes)”. For this topic, we guided them on what to do. Specifically, participants were directed on what buttons to press, what kinds of keywords to input into the system, how to highlight important texts, and how to do preference judging. Details about the system interface is given in subsection 3.3.1. In this topic, for each pair of documents shown to the participants, there was a correct answer. After participants made a judgment, we told them whether their decision was correct or wrong. In the case that their decision was wrong, they were given an explanation as to why their answer was wrong so that they could learn how to preference judge better. Working on the first topic gave participants a better understanding of how to do preference judging and how to use the system to complete the tasks.

After participants completed the first topic and have asked any questions they had, they were instructed to start working on the second topic, topic 133 “Does exercise improve the symptoms of depression? (Answer is Yes)”. For this topic, participants worked by themselves with no help from us. They were instructed to preference judge 7 pairs of documents. As with the first topic, each pair of documents in this topic had a correct answer. Participants' preference judgments for each pair were recorded to be compared to the correct answer later on to determine their judgment accuracy. After participants completed preference judging the two topics, they were given an opportunity to ask any last

¹¹https://youtu.be/QXGn_E0y1Kc

¹²<https://www.heroku.com/>

questions that they had after which the tutorial phase of the study concluded. Participants were then remunerated for their work and were instructed to wait for 2 business days for a decision about whether they were selected to participate in the main phase of the study or not. Decisions were made using the criteria detailed above.

It is to be noted that midway through conducting the tutorial phase, for one of the seven judgments in the second topic (topic 133), we found that all participants had judged differently compared to our correct answer, which led us to eliminate this specific pair of documents. To elaborate, participants still had to judge this specific pair of documents, but we did not count this judgment towards the accuracy score. Hence, for the remainder of the participants participating in the tutorial phase, instead of making 4 out of 7 correct judgments, they now had to make 4 out of 6 correct judgments in the second topic to be selected for the main phase.

3.3.6 Main Phase

All participants selected to participate in the main phase of the study were sent an invitation email with a consent form to fill, the preference judging instruction manual for reference and a guidelines manual reiterating what participants needed to do in the main phase. After participants signed the consent form, they were assigned topics to complete and received their login credentials via email to access the preference judging system. Participants performed preference judging at their own time and place and were allowed to take breaks as needed.

When they completed the assigned topic, they were asked to reorder the top documents (usually top-10 documents) that were generated from their preference judging for that topic from rank 1 being their most-preferred document to rank 10 being their least-preferred document. Documents were not allowed to be ranked equal. The reordering task was done using Google forms. To be specific, the top documents for the topic were zipped into a folder and a Google form was created where participants could input their ranking for those top documents. These were then sent to the participants via email. Details and screenshots of the Google form is given in subsection [3.3.2](#).

Participants were allowed to work for a maximum of 13 hours in the study. Hence, if participants completed preference judging and reordering for a topic and did not work for more than 13 hours yet, they were allowed to request for more topics to work on. In this study, there are 30 topics to collect preference judging data for. Each topic was preference judged by 3 participants independently. Assignment of topics to participants was on a random, first-come-first-serve basis. To be specific, participants who gave consent for the

main phase first were assigned topics first. Participants who completed work on a topic and requested more work first were assigned more topics first, as long as topics were available. The topics were assigned as in Figure 3.13, with columns being filled one-by-one.

All actions participants made while interacting with the preference judging system was logged along with the time that those actions were made. To be specific, the system created a log when:

1. the participants logged in to the system and landed on the homepage
2. the participants started judging a specific topic
3. the participants started working on a pair of documents
4. the participants made a preference judgment for a pair of documents
5. the participants decided to change their preference judgment and hit the “Undo” button
6. the participants completed judging the topic
7. the participants logged out of the system
8. anytime the participants landed on the homepage

The study ended once all topics had preference judgments done by 3 participants. All participants were remunerated for the work they completed. Although all participants were allowed to work for a maximum of 13 hours, only 5 participants worked for more than 10 hours. This was because there were more participants willing to work than there were topics available. Lastly, all preference judgment data from the main phase was stored and used for analysis, which is discussed in chapter 4.

3.4 Participants

For the pilot study for the tutorial phase, the participants included my supervisor and two fellow graduate students.

To recruit participants for the actual study, we put up posters across the university campus. Interested participants filled out a screening questionnaire linked on the poster, as detailed in subsection 3.3.3. We received a total of 121 responses of which 3 responses were

Topic Number	Assessor 1	Assessor 2	Assessor 3
102	participant58	participant3	participant33
103	participant3	participant25	participant55
104	participant4	participant53	participant5
105	participant33	participant5	participant21
106	participant32	participant33	participant40
107	participant42	participant26	participant39
108	participant17	participant47	participant9
109	participant23	participant54	participant34
110	participant32	participant45	participant61
111	participant27	participant4	participant9
114	participant18	participant2	participant39
115	participant17	participant13	participant8
117	participant8	participant20	participant42
118	participant21	participant29	participant9
120	participant7	participant20	participant22
121	participant29	participant15	participant39
122	participant13	participant32	participant5
128	participant9	participant4	participant13
129	participant39	participant21	participant34
131	participant12	participant41	participant53
132	participant41	participant9	participant8
134	participant43	participant40	participant39
136	participant19	participant20	participant33
137	participant39	participant27	participant2
139	participant22	participant29	participant9
140	participant18	participant57	participant54
143	participant42	participant29	participant53
144	participant8	participant45	participant27
146	participant5	participant56	participant8
149	participant6	participant49	participant54

Figure 3.13: Topic Assignment for the Main Phase

removed due to being a duplicate response to another and another 3 responses were removed as they didn't satisfy one of the three necessary requirements stated in the screening questionnaire.

Participants who passed the screening questionnaire were asked to read an information letter detailing the study and fill out a consent form if they agreed to participate in the tutorial phase. A total of 83 participants gave their consent and booked a time slot to participate in the tutorial phase of the study, however only 67 participants actually showed up and completed the tutorial tasks.

Participants' demographic information were collected during the tutorial phase. Of the 67 total participants, 36 participants identified as Woman, 30 participants identified as Man, and 1 participant identified as Non-binary/Non-conforming. Participants were aged between 17 to 31 years old, with 23 participants less than 20 years old, 35 participants between 20 to 25 years old, and 9 participants older than 25 years old (average age of 21). All participants except one were students. Participants were from different majors, specifically 20 from mathematics, 15 from engineering, 14 from science, 10 from arts, 4 from environment, 3 from health, and 1 unanswered. In terms of race or ethnic background, 27 participants were South Asian descent, 19 were Chinese, Korean, Japanese, Taiwanese descent, 7 were European descent, 5 were Southeast Asian descent, 4 were African, Afro-Caribbean, African-Canadian descent, 2 were Arab, Persian, West Asian descent, 1 was Latin American or Hispanic descent, and 2 preferred to self-identify as Turkish/European and Indian/European.

Of the 67 participants who completed the tutorial phase of the study, 2 participants stated they did not want to continue the study, hence we consequently removed their study data. Fifty-one participants passed the tutorial phase and were invited to participate in the main phase of the study, of which 49 participants gave consent to participate. However, not all 49 participants completed the main phase. Three participants were unable to contribute to the study due to time constraints, and 5 participants discontinued their involvement mid-way through the study, thereby compromising the usability of their collected data. In the latter case, we had to reassign those topics to other available participants. Additionally, one participant who only worked on one topic had their data affected by the system bug mentioned in section 3.5. Therefore, for the main phase of the study, we collected preference judgments from 40 participants in total. Participants were remunerated \$20 for the tutorial phase and \$20/hour for the main phase of the study rounded up to the nearest half-hour.

3.5 System Bugs in the Preference Judging System: Observations and Remedial Actions

During the main phase of the study, we observed certain system bugs due to the utilization of a newly developed preference judging system that had not been extensively tested. One of the identified system bugs resulted in the inadvertent loss of previously identified best documents by the participants, thereby increasing the total judging time for a topic beyond what was necessary to retrieve the top-10 documents. This error was due to the system's inability to manage scenarios where participants rapidly clicked on an action button in the preference widget. Upon discovering this issue, prompt action was taken to fix the bug. We also had to remove all data of topics that were affected by this bug as they were deemed unusable. The affected topics were subsequently re-assigned to other participants for preference judging.

3.6 Data Cleaning

After the study's conclusion, we noticed issues with the recorded times for when judgments were created that required correction before data analysis could begin. To elaborate, the system records the timing of judgment creation and completion for each pair of documents presented to the participants. Notably, the created time of a new judgment equals the completed time of the previous judgment, as participants are shown fresh document pairs after rendering their decision by clicking on the preference widget buttons. However, when a participant logs out of the system before clicking the buttons to record their preference judgment decision and subsequently logs back in, the created time for the specific pair of judgments remains fixed at the time the participant initially viewed that document pair, instead of being updated to the log-in time. Consequently, we had to correct the created time of these specific judgments to reflect the respective participant's log-in time.

3.7 Measures Used for Data Analysis

In this thesis, we used the following measures to analyze our results:

Kendall's tau is a measure commonly used in information retrieval to measure the correlation between two ordered lists. It is a measure of rank correlation with values ranging from -1 to 1, where -1 indicates a perfectly negative correlation, 0 means no correlation,

and 1 means perfect correlation [48]. Vorhees used the Kendall’s tau measure in her study to analyze inter-assessor consistency using qrels [50]. Scholer et al. also used Kendall’s tau in their study of consistencies of relevance judgments [46].

AP correlation is a measure of correlation between two ranked lists, based on the idea of precision and recall. It is based on the notion that documents at the top of the ranking carries more significance. Thus, it is computed by determining the likelihood of each item being ranked correctly in regards to the items above it and then taking the average of all items in the list. The measure ranges from -1 to 1, where -1 means perfect negative correlation, 0 means no correlation, and 1 means perfect correlation [58].

Rank Biased Overlap (RBO) is a measure of similarity between two ranked lists, where the idea is that items at the top of the list are more likely to be seen by users than those at the bottom. RBO gives different weights to items in the list based on their position in the list. The range for RBO is from 0 to 1, where 0 indicates no overlap between the two lists and 1 indicates perfect overlap [53]. In our study, we compute RBO using Webber et al. ’s formulation for handling ties, except that we stop at rank 10. The code that we used to compute RBO can be found in Appendix B.

Compatibility calculates the maximum similarity between an information retrieval system’s ranking and an ideal ranking. Rank biased overlap (RBO) [53] is used to calculate compatibility [14, 15, 13].

Cohen’s kappa is a measure of inter-assessor agreement, taking into account the possibility of agreement by chance. Cohen’s kappa ranges from -1 to 1, where Cohen suggests that values ≤ 0 indicates no agreement, 0.01–0.20 indicates none to slight agreement, 0.21–0.40 indicates fair agreement, 0.41– 0.60 indicates moderate agreement, 0.61–0.80 indicates substantial agreement, and 0.81–1.00 indicates almost perfect agreement [35].

Fleiss’ kappa is a measure that extends Cohen’s kappa to calculate inter-assessor agreement for more than two assessors. The kappa score ranges from 0 to 1. A score of 0 indicates no agreement beyond chance and a score of 1 indicates perfect agreement [20].

Chapter 4

Results and Discussion

In this chapter, we initially report results from the tutorial phase (section 4.1), and then make three analyses on the collected data from the main phase (sections 4.2, 4.3, 4.4). From the data from the main phase, we first look at assessor behavior when preference judging. Second, we measure assessor agreement to see how consistent they are at making preference judgments. Third, we observe how much preference judging to identify the top documents can affect the ranking of runs compared to the ranking produced by NIST assessors' graded relevance judgments as reported in the TREC 2021 Health Misinformation track.

4.1 Analysis of Results from Tutorial Phase

In this section, we report results about participants' performance measured using the criteria listed in chapter 3.3.5. As mentioned previously in chapter 3.4, of the 67 participants who participated in the tutorial phase, 51 participants passed and were invited to participate in the main phase of the study. Of the 16 participants who were not invited to participate in the main phase, 2 participants stated they did not want to participate in the main phase of the study and had their data consequently removed from the study. Additionally, there was one participant who completed the tutorial phase in our preference judging system that was, at the time, undergoing system maintenance. Hence we cannot include the participant's tutorial data in this analysis. Thus, we report statistics for only the 64 participants who participated in the tutorial phase.

In terms of accuracy, we measured participants' preference judging performance by looking at how many correct judgments they made out of 6 total document pairs in the

second topic (topic 133) they worked on (note that we disregarded one pair as mentioned in chapter 3.3.5). Of the 64 participants, 13 participants judged less than 4 pairs correctly and 51 participants judged 4 or more pairs correctly. Table 4.1 shows the frequency of participants’ preference judgment accuracy for the second topic of the tutorial phase. The table shows that only 6 participants were able to judge all pairs of documents correctly, where the majority of participants judged 4 out of 6 pairs correctly. None of the participants judged less than 2 pairs of documents correctly. In general, we can see that participants did well in the tutorial phase.

Number of Correct Pairs	Number of Participants
0	0
1	0
2	5
3	8
4	28
5	17
6	6

Table 4.1: Frequency of Participants’ Preference Judgment Accuracy for the Second Topic of the Tutorial Phase

In terms of efficiency, as expected, participants in general took approximately 1 hour to complete the tutorial phase. There were 3 participants who took considerably longer (more than 1 hour 30 minutes). Consequently, these participants did not pass the tutorial phase. Lastly, in terms cooperativeness and suitability for judging tasks, all participants were enthusiastic and performed preference judging seriously.

4.2 Analysis of Assessor Behaviour

We used the data that we collected on participants’ judgment time and interactions with the system to study assessor behavior when preference judging. Each judgment made by a participant is calculated by taking the difference between the time they completed a judgment and the time they started a judgment. We analyze judgments made without an “Undo” action performed and judgments made because of an “Undo” action separately. To reiterate, “Undo” actions happen when participants already made a judgment for a pair of documents but clicked the “Undo” button to go back to that specific pair of documents to change their judgment.

4.2.1 Analysis of Judgments Made without “Undo” Actions

For all results and analysis in this subsection, we only consider judgments that participants made without clicking on the “Undo” button. Additionally, we removed all judgments that took more than 600 seconds. From this data, we first looked at how much time participants took to make each judgment for all the judgments collected in the study. Figure 4.1 shows the judgment time distribution for 60 seconds, figure 4.2 shows the judgment time distribution for 180 seconds and figure 4.3 shows the judgment time distribution for 600 seconds. We can see here that all graphs are right skewed, showing that the majority of judgments were made relatively quickly, with fewer cases taking significantly longer. In general, most judgments took less than 20 seconds to make. There are a few judgments that took close to 600 seconds to make, which can be considered outliers. Over all the judgments made in the study, the average time to make a judgment is 57.75 seconds.

Next, we calculated each participant’s micro-average judgment time across all the topics that they worked on by taking the total time they worked on all the topics they judged divided by the total number of judgments they made. The histogram in Figure 4.4 shows the trend where most participants took a shorter amount of time to make a judgment and fewer participants took longer to make a judgment. In this study, we see most participants making a judgment within 15 to 70 seconds. There are two participants who took relatively longer than the rest of the participants to make judgments (between 200 to 230 seconds per judgment), which could either be because the topics they worked on had difficult documents to read, the topics were harder to judge, or they were just slow assessors. More research and analysis on the data needs to be made here in order to answer this question. The average of the average of participant’s judgment time is 80 seconds.

We also calculated each topic’s micro-average judgment time. A topic is judged by 3 participants. We take the sum of all 3 participants’ judgment time divided by the sum of all 3 participants’ number of judgments. Figure 4.5 reflects this calculation, where we see that the shape of the graph is skewed right, similar to the graph in Figure 4.4. For most topics, it takes less than 50 seconds to make a judgment and fewer topics take longer. The average of the average of topic’s judgment time is 83.27 seconds, which is around the same as the average of the average of participant’s judgment time.

To have a better idea of how much time it takes to find the top-10 preferred documents using preference judging, we plotted the number of judgments it took for each participant to judge a topic versus the number of documents in that topic in the scatter plot shown in Figure 4.6. It is evident that as the number of documents in a topic increases, so does the number of judgments required to identify the top-10 documents within that topic. Specifically, the number of judgments needed to find the top-10 documents in a topic

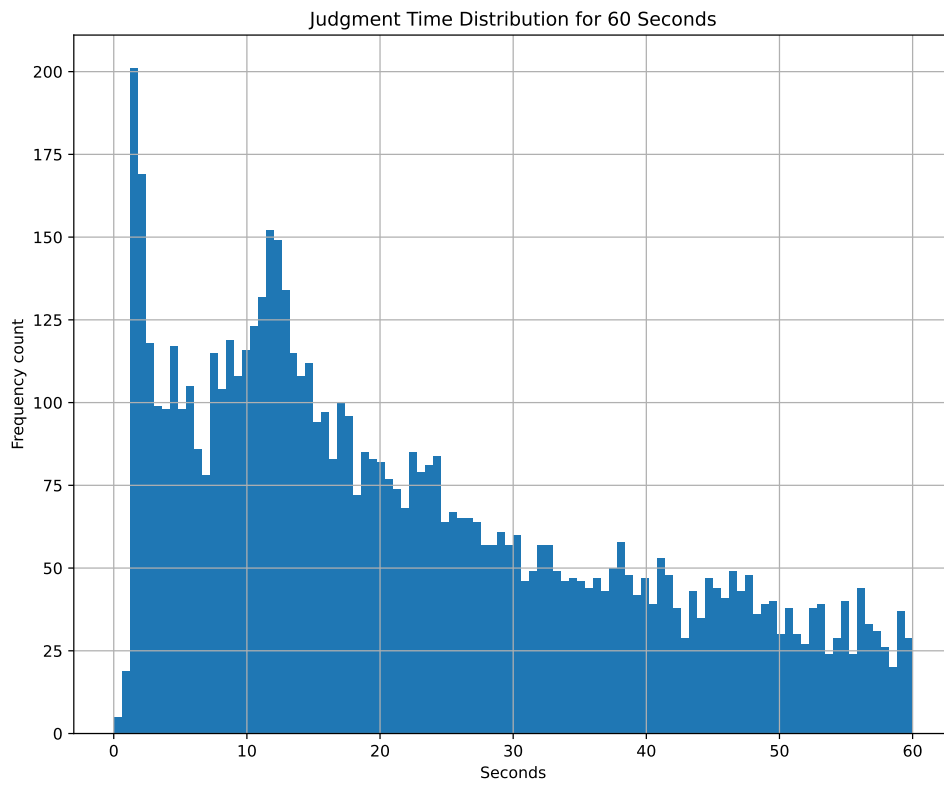


Figure 4.1: Judgment Time Distribution for 60 Seconds

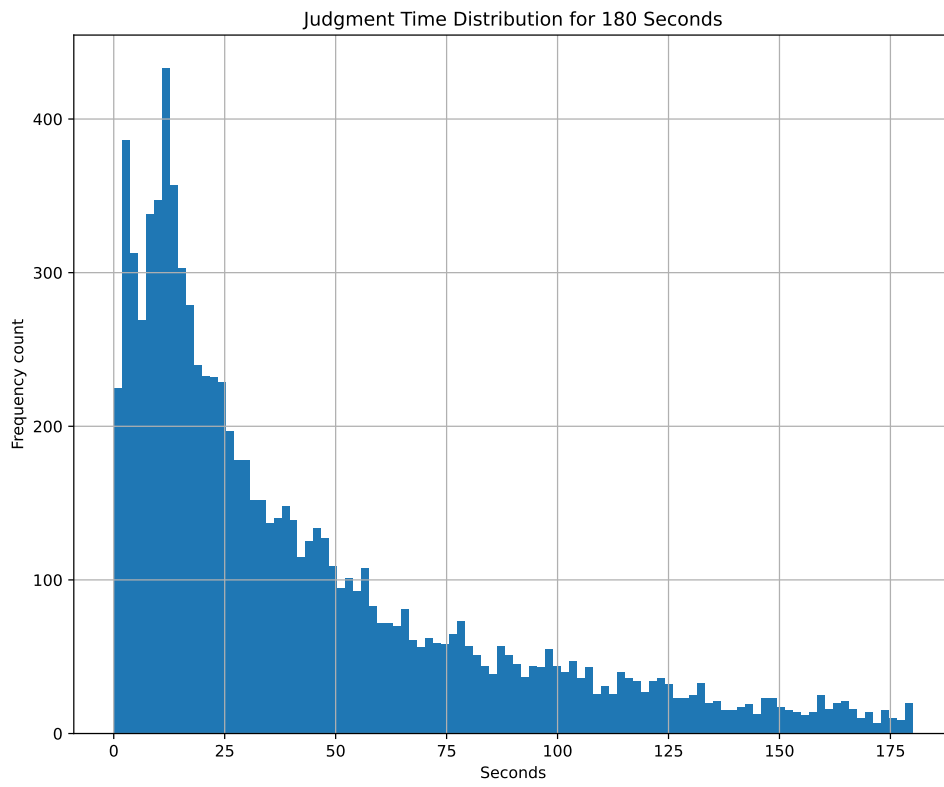


Figure 4.2: Judgment Time Distribution for 180 Seconds

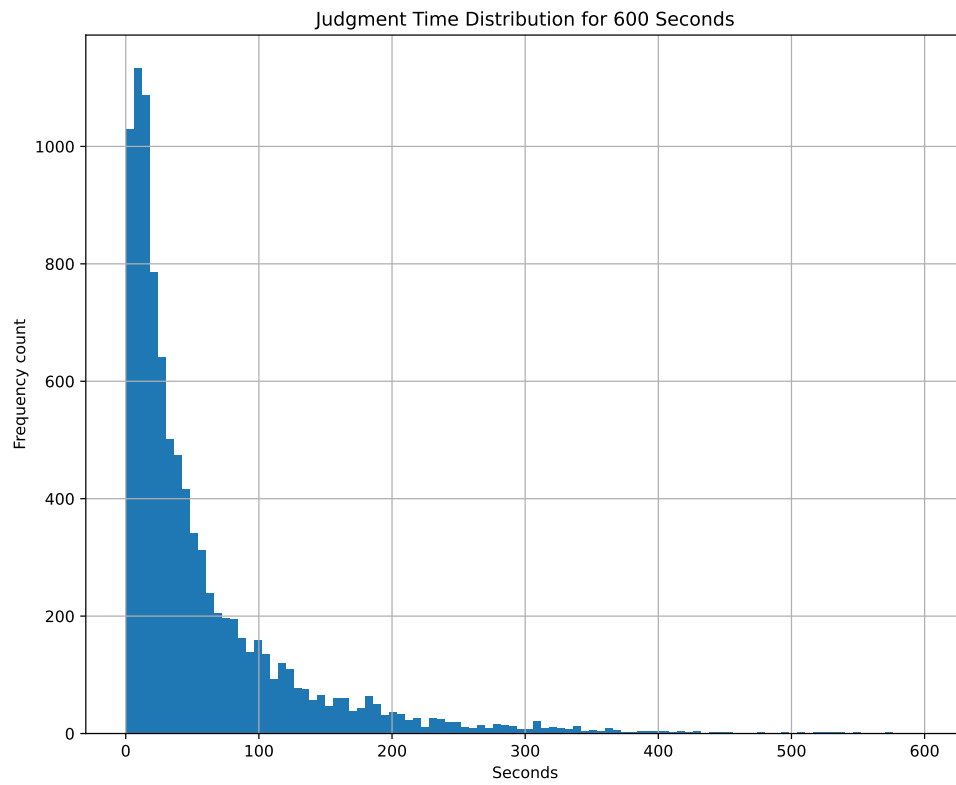


Figure 4.3: Judgment Time Distribution for 600 Seconds

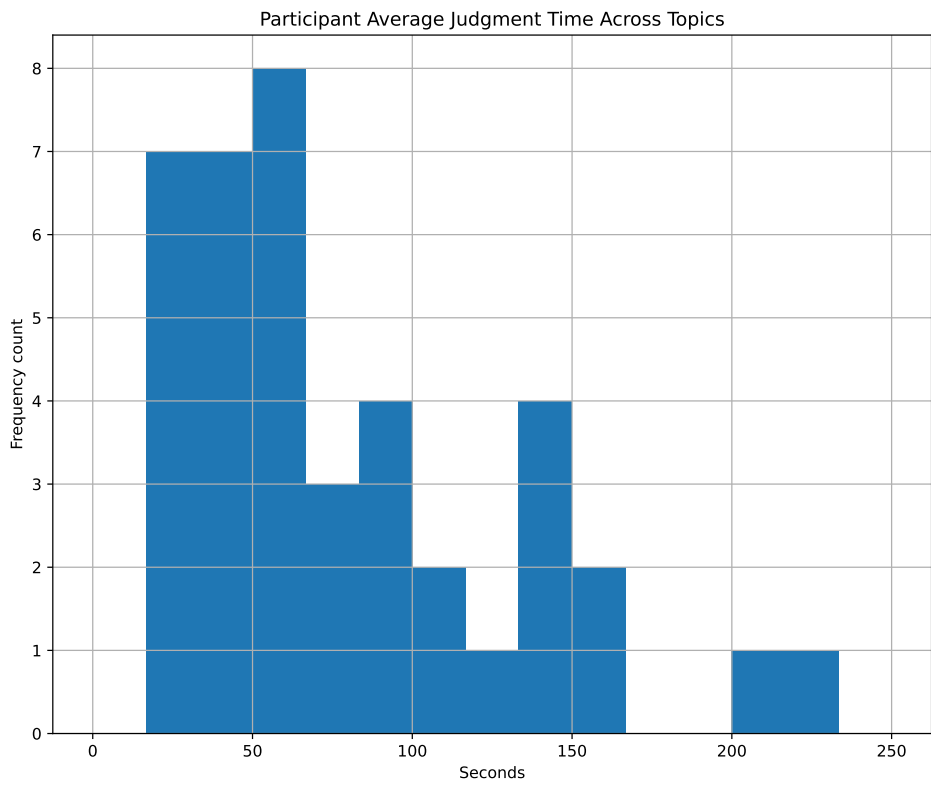


Figure 4.4: Participant Average Judgment Time Across Topics

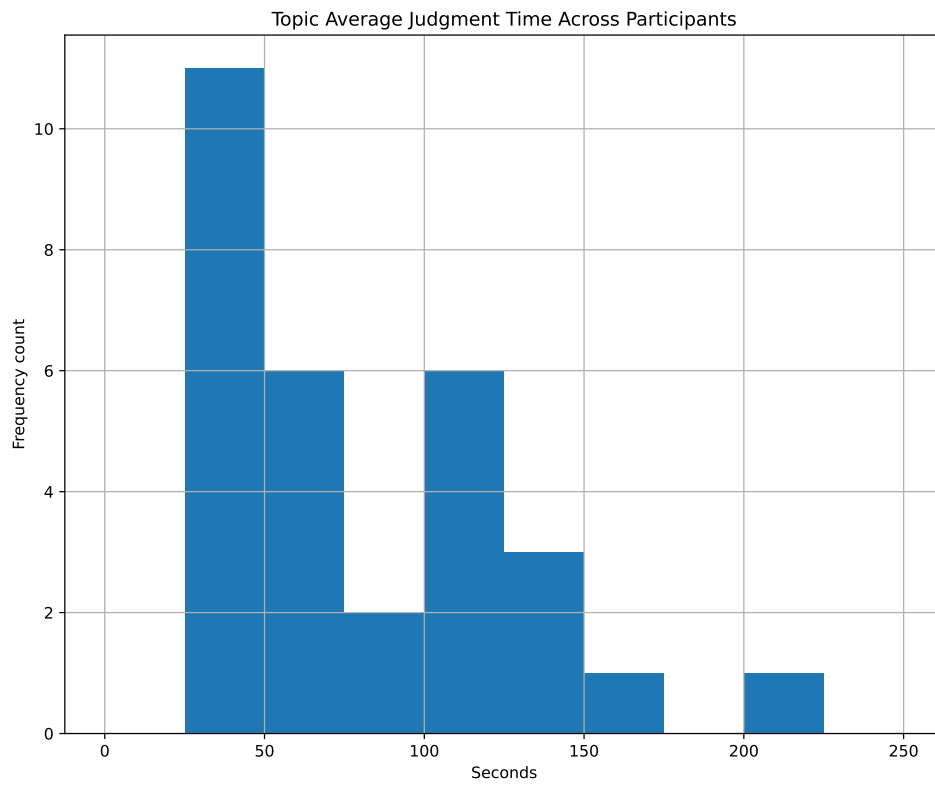


Figure 4.5: Topic Average Judgment Time Across Participants

is approximately twice the number of documents in the topic. Similar conclusions can be made by looking at Figure 4.7, where we average the number of judgments made for a topic first and then plot it against the number of documents in that topic. This observation may be useful for future studies, as it provides insight into determining the optimal number of documents to judge and estimating the associated costs.

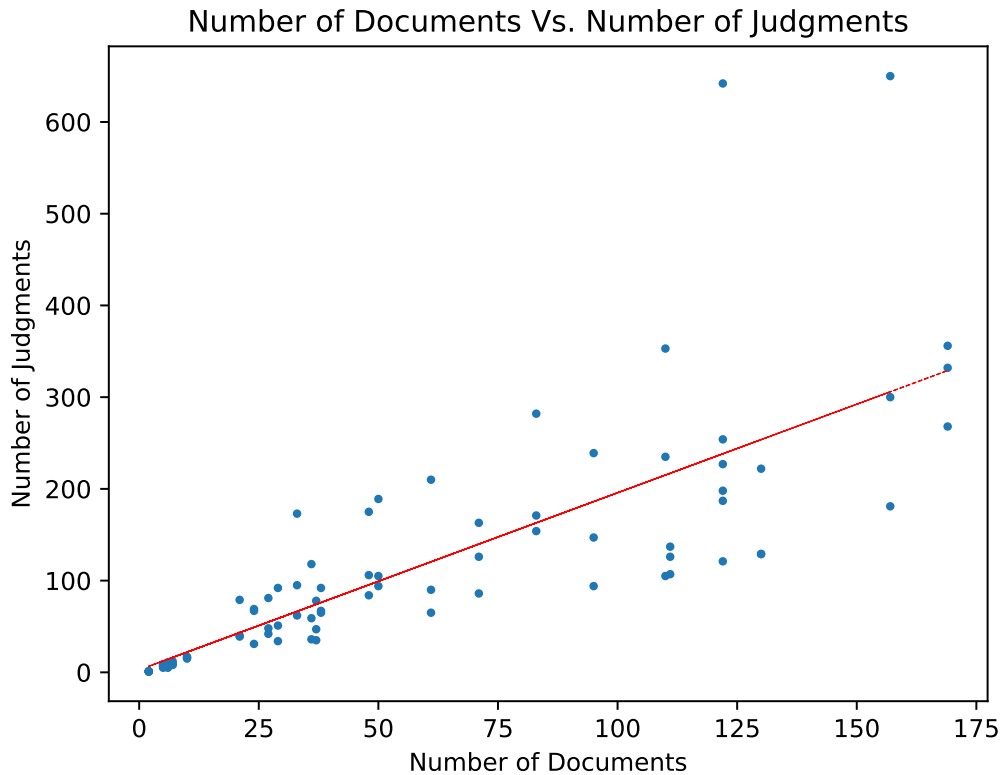


Figure 4.6: Number of Documents Vs. Number of Judgments

4.2.2 Analysis of Judgments Made because of an “Undo” Action

Undo actions occur when participants change their initial preference judgment for a document pair by clicking the “Undo” button. In this study, 9940 judgments were made across all topics, with only 126 of them involving undo actions. Of the 40 participants who participated in the main phase of the study, 25 participants performed at least one

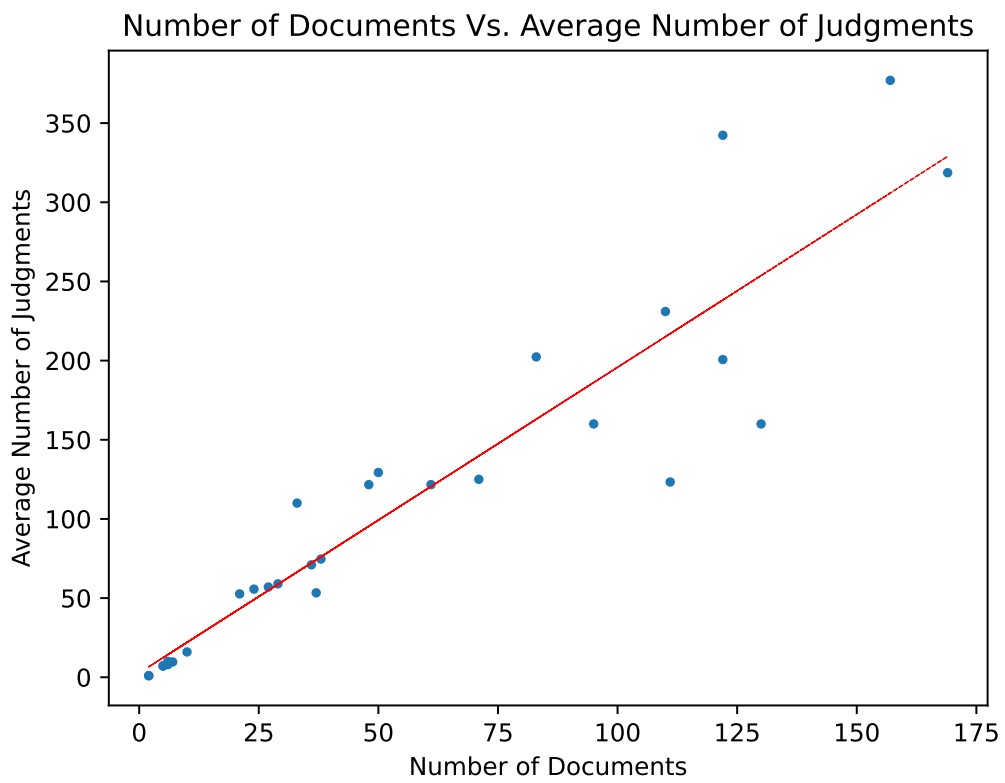


Figure 4.7: Number of Documents Vs. Average Number of Judgments per Topic

undo action. Figure 4.8 shows the distribution of undo actions. The graph shows that the majority of participants executed 1-2 undo actions, while a few outliers made up to 20 undo actions. It should be noted that the number of undo actions may have been higher for some participants if a topic had more documents to judge. Figure 4.9 displays the time distribution of judgments that involved undo actions. Note that the time of each judgment made because of an “Undo” action is from the time the “Undo” button was clicked to the time the new judgment was completed. The graph reveals that the majority of undo actions took less than 10 seconds to complete, with a few outliers taking nearly 600 seconds. To calculate the average undo time across all participants and topics, we first calculated the average undo time for each participant across all the topics they made undo judgments on. We then calculated the average of these averages to get the overall average undo time for participants, which was 39.68 seconds. Finally, we also observed that out of the 126 judgments involving undo actions, in 94 cases, participants retained their initial decision before and after the undo action, while only in 32 cases did participants alter their initial decision.

4.3 Analysis of Agreement between Assessors

Each topic in our study was preference judged by 3 participants. In this section, we measure inter-assessor and intra-assessor agreement to study assessors’ judgment consistency in preference judging.

4.3.1 Inter-assessor Agreement

To assess the degree of agreement between assessors in our study, we computed Kendall’s tau, Rank Biased Overlap (RBO), Cohen’s kappa and Fleiss’ kappa. For each topic, we measured similarity of the rankings between:

- participant 1 vs. participant 2
- participant 1 vs. participant 3
- participant 2 vs. participant 3

To figure out how similar assessors’ document rankings are to each other, we calculated Kendall’s tau scores. For the participants for each topic, we appended the participants’ top

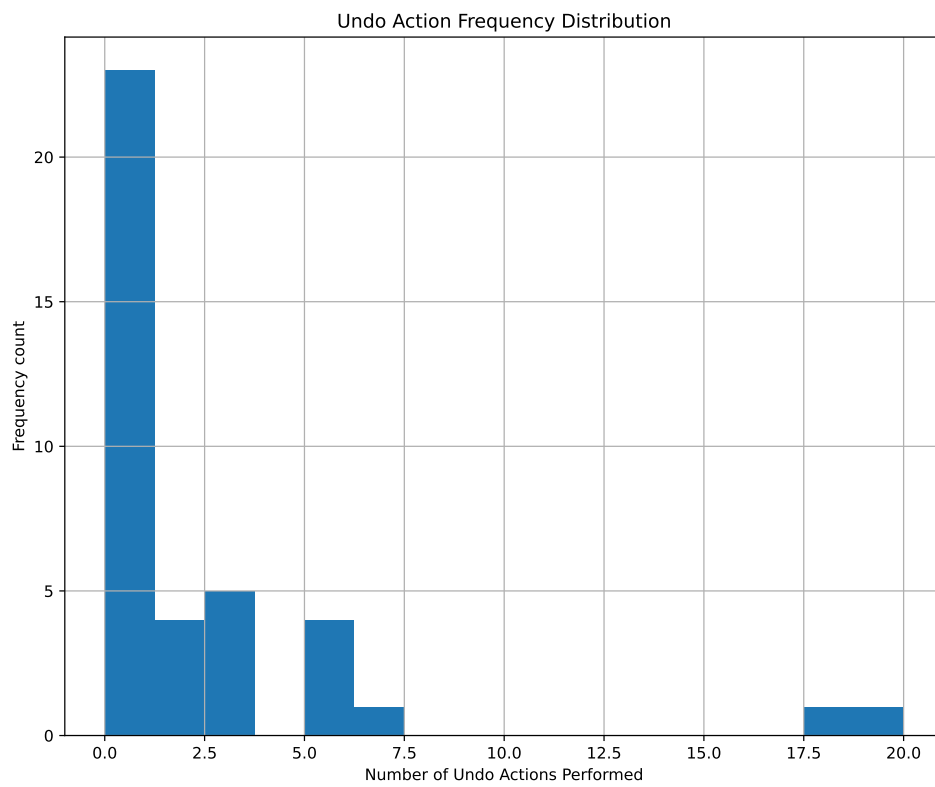


Figure 4.8: Undo Action Frequency Distribution

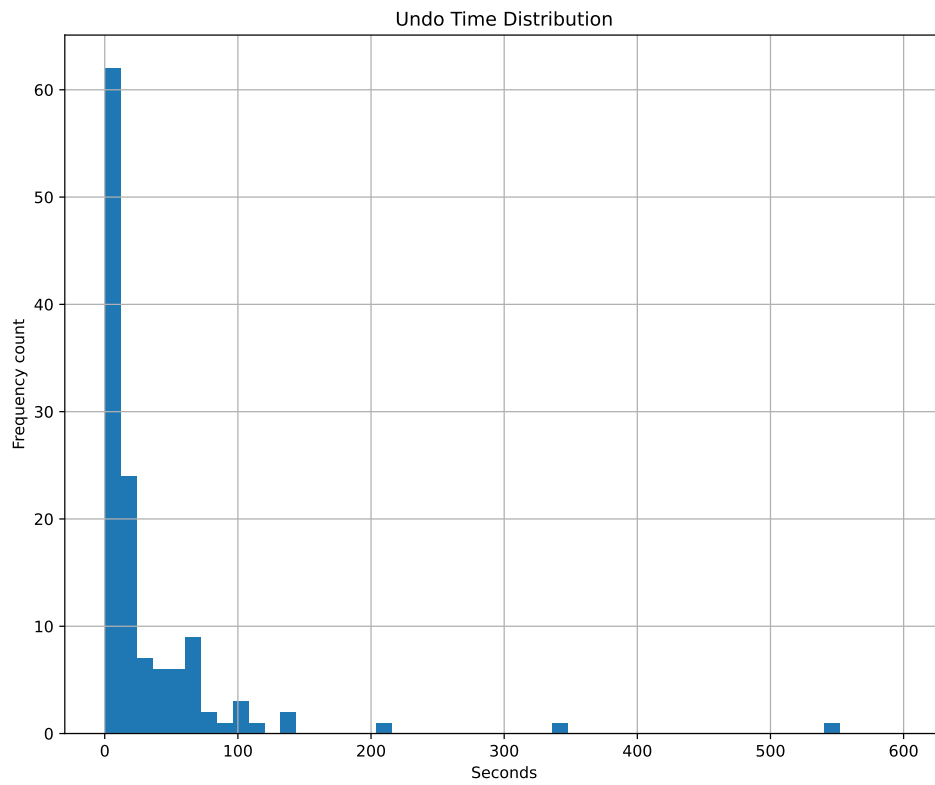


Figure 4.9: Time Distribution of Judgments that Involved Undo Actions

documents from their preferences to the top of the rest of the documents in that topic that was not ranked to be in the top documents. Figure 4.10 shows the Kendall’s tau scores for each comparison of assessors by the number of documents in the topic. We see that the majority of the Kendall’s tau scores are greater than 0, indicating a positive correlation between the rankings provided by the assessors. Nonetheless, the variability of these scores implies that the level of agreement between assessors is not absolute. Additionally, there are some ranking comparisons that have negative scores, which shows that some assessors don’t agree with each other for certain topics. Another notable observation is that in cases where the number of documents for a given topic is less than or equal to 10, the Kendall’s tau scores tend to indicate an almost perfect positive correlation (1) or negative correlation (-1). This is largely due to the fact that our study seeks to identify the top 10 documents for each topic, but when there are only 10 documents to rank, the assessment focuses more on how the assessors reorder these documents, rather than on their ability to identify the same documents as the top documents.

Next, in order to investigate to what extent participants’ preferences overlap, we computed RBO scores by comparing participants’ top 10 documents for each topic. Figure 4.11 shows the RBO scores with patience parameter set to 0.7 for each comparison of assessors by the number of documents in the topic. Note the parameter is set to 0.7 to put more emphasis on the top ranked documents. We can see from the plot that for topics with less than or equal to 10 documents, as our study sought to identify the top-10 documents for a topic, the RBO scores are near perfect overlap as all assessors’ preferences consist of the same documents. However, for topics with more than 10 documents, the RBO scores drop. A majority of the RBO scores are less than 25%, indicating low overlap between assessors’ preferences on a topic. This observation led to the question of whether assessors working on a given topic were even able to identify the same documents as their top-10 documents.

To determine whether assessors working on a given topic were able to identify the same top documents beyond chance or not, we computed Cohen’s kappa and Fleiss’ kappa. For these calculations, we decided to exclude any topics that had less than or equal to 10 documents as we were more interested in measuring agreement among participants based on their ability to identify and rank the same documents among their top 10, rather than observing how these documents were re-ordered. For each topic, we took all the documents for that topic and asked to what extent participants agreed which documents are considered a “top-10” document and which documents are not. Figure 4.12 shows Cohen’s Kappa scores for each comparison of assessors versus the number of documents in the topic. Note that the dashed line is meant to help readers read the plot better. From the interpretation of Cohen’s Kappa values given in the paper by McHugh [35], we see that most values are above 0, which shows that there is, in general, agreement among assessors

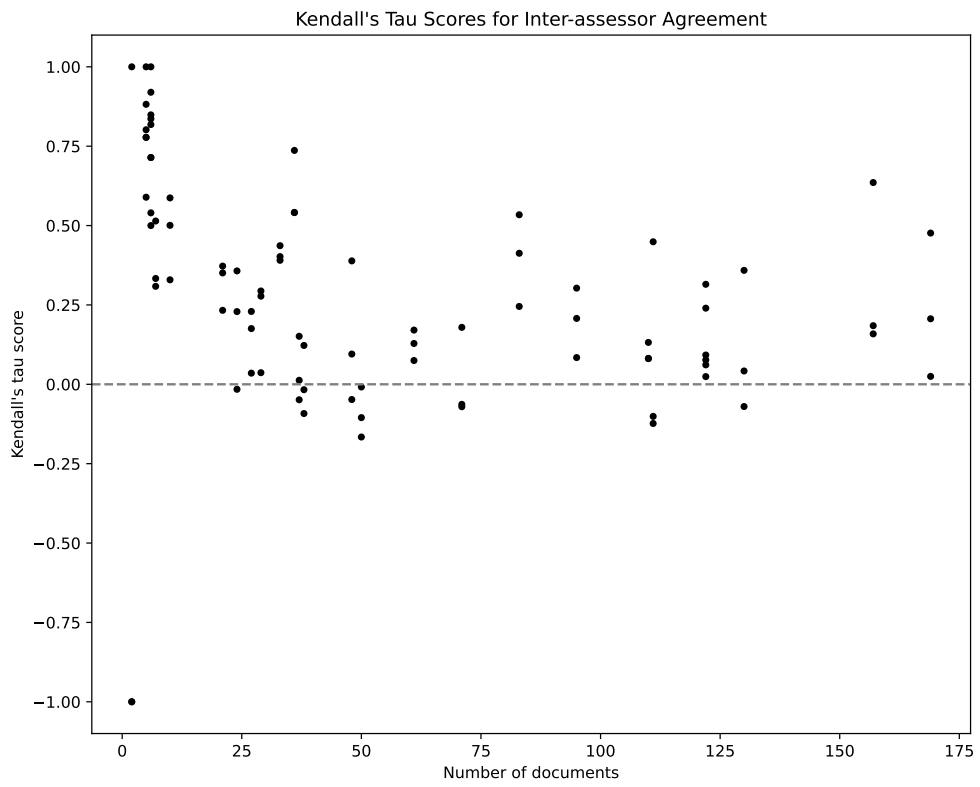


Figure 4.10: Kendall's Tau Scores for Inter-assessor Agreement

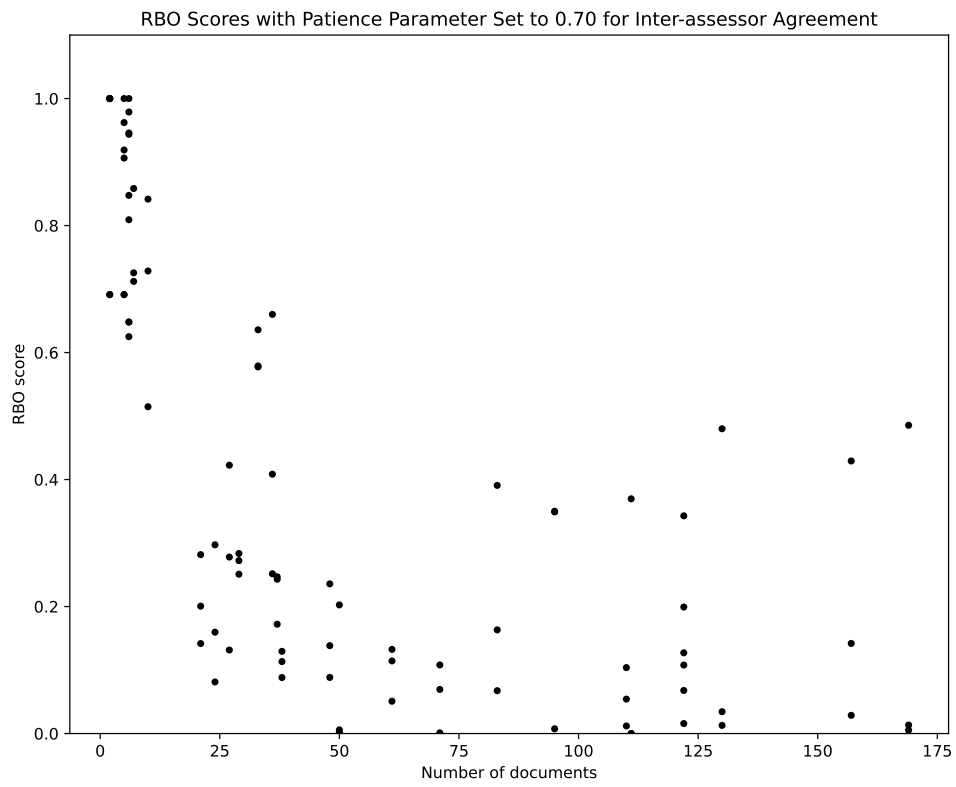


Figure 4.11: RBO Scores with Patience Parameter Set to 0.70 for Inter-assessor Agreement

beyond that expected by mere chance. Taking the mean of all the Cohen’s kappa scores, we get a value of 0.171 which means there is only slight agreement among assessors. To figure out what is the maximum possible agreement among our assessors, we assumed that there is a participant who performed preference judging badly among the three participants that worked on a topic. Hence, we removed this participant for each topic and calculated the mean to get a Cohen’s Kappa value of 0.328. This value indicates that in this study, the maximum possible agreement among our three assessors is fair, which is only a slight increase in agreement.

Figure 4.13 shows the Fleiss’ Kappa scores for each topic taking into account all three participants. Note that the dashed line is meant to help readers read the plot better. We see that all the scores except two were above 0, which shows that there is agreement among assessors although the degree of agreement varies. Assessors agreed more in some topics than others. For two topics, however, there is negative agreement among assessors. The two topics with negative Fleiss’ kappa scores are shown in Table 4.2. Upon initial inspection of the documents in these two topics, we noticed that in general, the documents we saw were of poor quality. While the documents did address the topic, they either did not give a clear answer or were not credible enough. We suspect poor document quality might have affected participants and caused them to have a hard time finding the same top documents for the topics. More investigation is needed. The average Fleiss’ kappa score for all topics with more than 10 documents is 0.165.

Topic Number	Topic Question	Number of Docs
108	“Is starving a fever effective? (Answer is No)”	38
136	“Can eating dates help manage iron deficiency anemia? (Answer is Yes)”	50

Table 4.2: Topics with Negative Fleiss’ Kappa Scores

4.3.2 Intra-assessor Agreement

In addition to measuring the inter-assessor agreement, we also assess the intra-assessor agreement. For this analysis, in order to analyze whether assessors are agreeing with themselves or not, we calculate compatibility scores with patience parameter set to 0.70 between participants’ preferences for a topic produced by the preference judging system and their re-ordered preferences for the same topic. For brevity, we call this a participant-topic pair. To reiterate, we have 30 topics in the study and each topic was preference judged by 3 participants. Hence, we have 90 participant-topic pairs in total. Also note the parameter

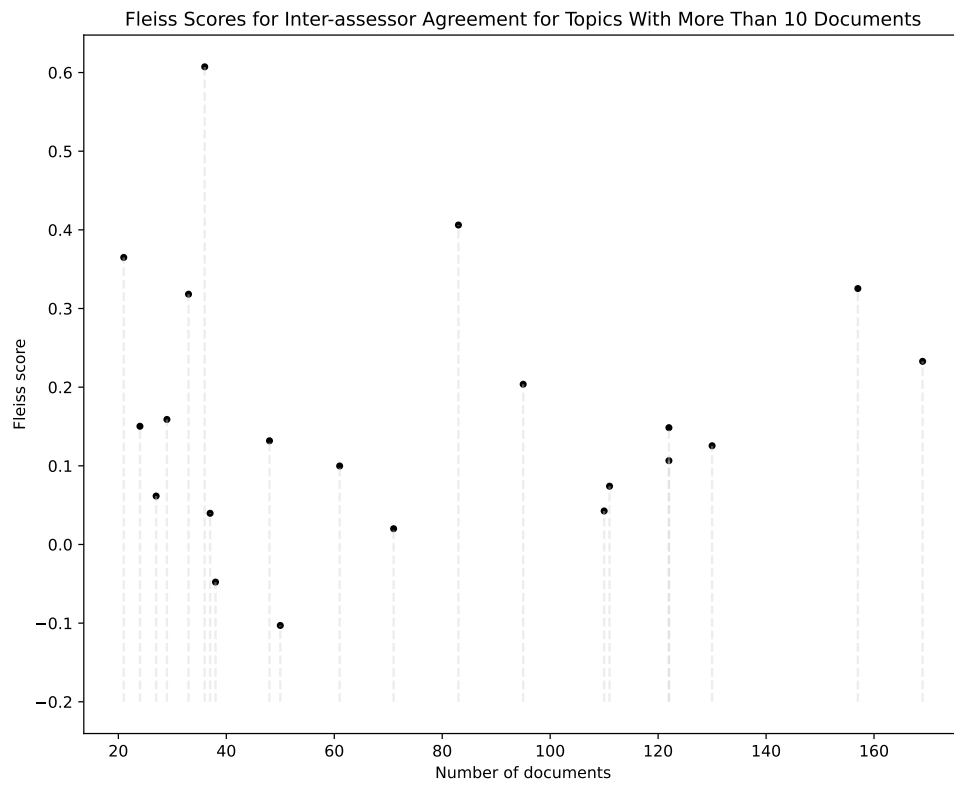


Figure 4.13: Fleiss' Kappa Scores for Inter-assessor Agreement

is set to 0.70 here to put more emphasis on the top ranked documents. Figure 4.14 shows the distribution of compatibility scores for all participant-topic pairs in the study. The plot shows that one-third of the participant-topic pairs have a high compatibility score of 1.0, of which 10 are participant-topic pairs with preferences where all documents had the same grade. The compatibility scores for the rest of the participant-topic pairs varies, with some having higher compatibility score than others. The average compatibility score for all the participant-topic pairs is 0.781. Figure 4.15 shows the distribution of compatibility scores for all topics with less than or equal to 10 documents. The average compatibility score for participant-topic pairs for topics with less than or equal to 10 documents is 0.834. Figure 4.16 shows the distribution of compatibility scores for all topics with more than 10 documents. The average compatibility score for participant-topic pairs for topics with more than 10 documents is 0.759. We can see that for topics with less than or equal 10 documents, there is high intra-assessor agreement. This is still apparent for topics with more than 10 documents, but not as frequent. Overall, it can be inferred that the assessors agree with themselves, but the level of agreement is varied and not particularly strong.

4.3.3 Summary of Assessor Agreement

To summarize, we find that based on the four measures used to assess inter-assessor agreement, while assessors do agree with each other at a rate greater than chance, this degree of agreement is not significant. We also see from our analysis of intra-assessor agreement that assessors do tend to agree with their own preference orderings, but this level of self-agreement varies.

4.4 Analysis of the Effect of Preference Judgments on the Ranking of Runs

To assess how much preference judging affected the ranking of runs compared to that of NIST’s, we used the preferences generated from our participants’ preference judging to create new rankings for the runs submitted in TREC 2021, and then compared these new rankings to the ranking produced by NIST assessors’ graded relevance judgments. In the next few paragraphs, we first detail the process of creating these new rankings and then provide results and analysis.

In our study, each topic was preference judged by 3 different participants, as shown in Figure 3.13. The first participant who judged a specific topic is called “participant

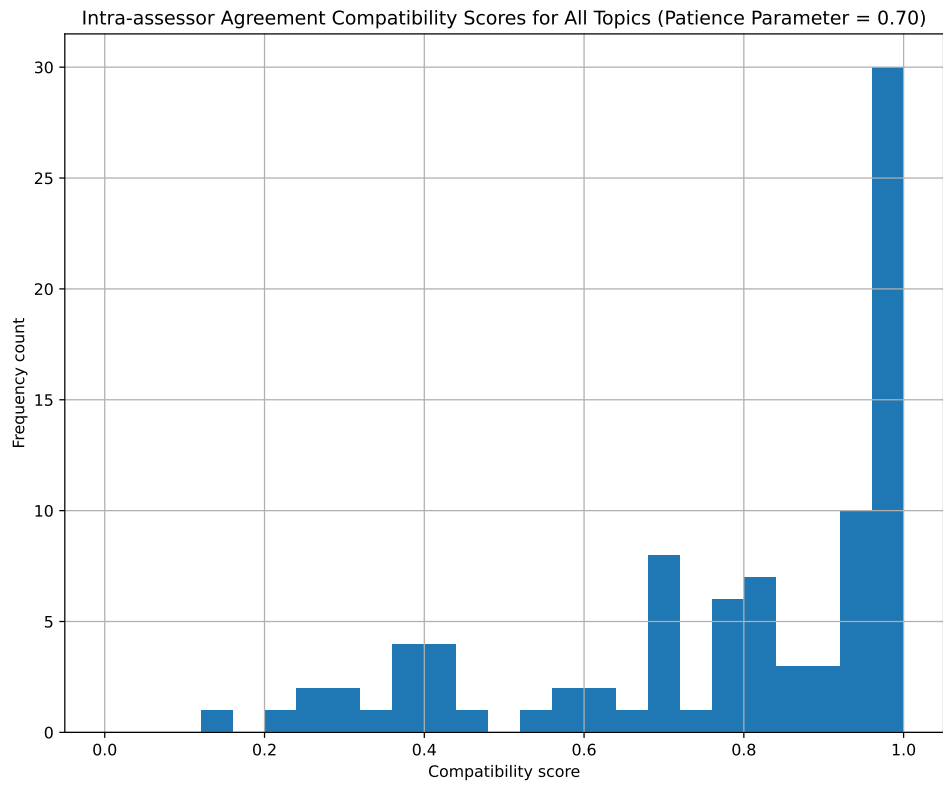


Figure 4.14: Intra-assessor Agreement Compatibility Scores for All Topics with Patience Parameter Set to 0.70

Intra-assessor Agreement Compatibility Scores for Topics ≤ 10 Documents (Patience Parameter = 0.70)

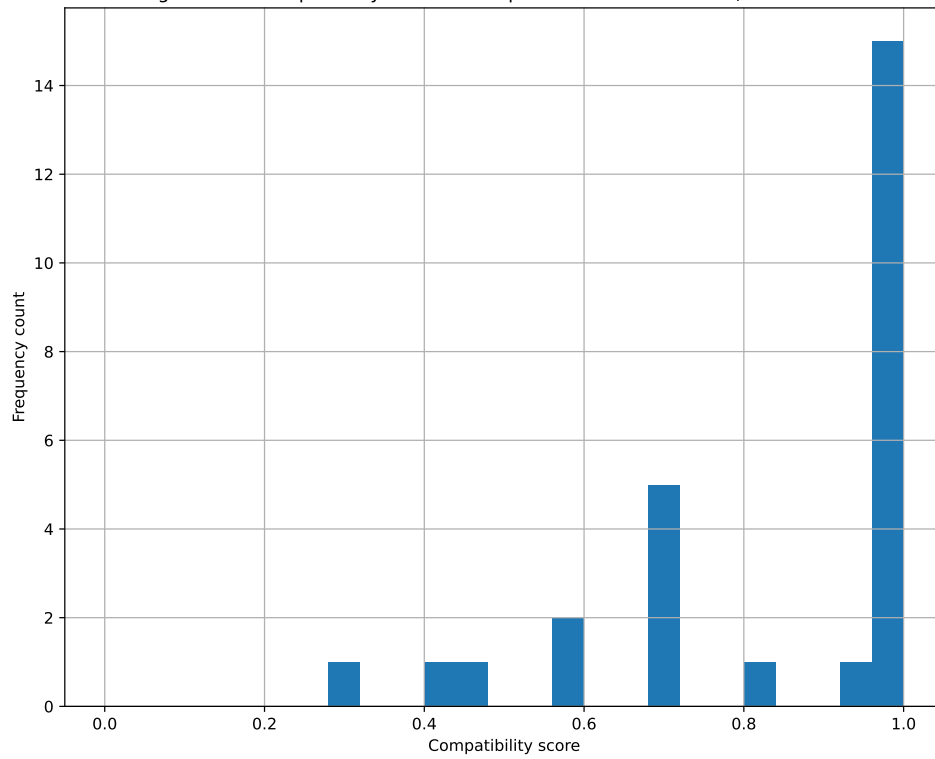


Figure 4.15: Intra-assessor Agreement Compatibility Scores for Topics with ≤ 10 Documents with Patience Parameter Set to 0.70

Intra-assessor Agreement Compatibility Scores for Topics > 10 Documents (Patience Parameter = 0.70)

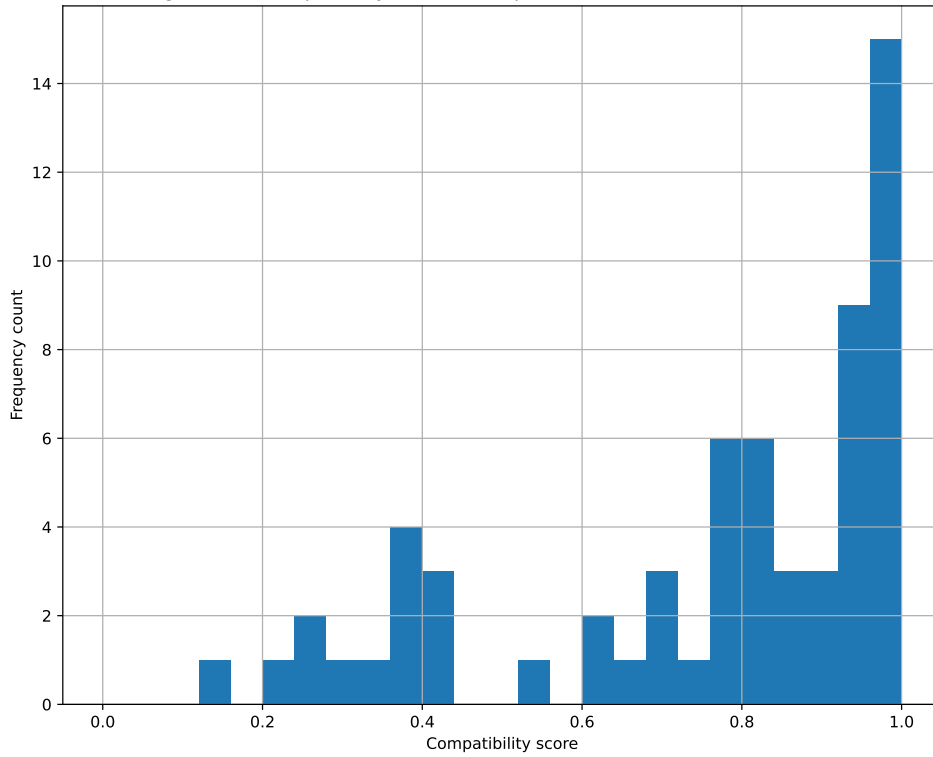


Figure 4.16: Intra-assessor Agreement Compatibility Scores for Topics with > 10 Documents with Patience Parameter Set to 0.70

1”, the second participant who judged that same topic is called “participant 2”, and the third participant who judged that same topic is called “participant 3”. Thus, 3 sets of top documents were produced for a topic. Using these sets of top documents, we created 3 new qrels corresponding to each participants’ preference judgments. Note that in this paper, we refer to the qrels created from NIST assessors’ graded relevance judging in TREC 2021 as “NIST qrels” for the sake of brevity. For each participant, we follow the steps described below to create new qrels:

1. First, for each topic, we reverse the preference values of the top documents found by preference judging so that they align with the preference values used in TREC 2021. In TREC 2021, the best documents have higher preference values, whereas worse documents have lower preference values. However, the opposite is true for the preference values assigned by the preference judging system, where better documents have lower preference values and worse documents have higher preference values. To be able to append the documents found using preference judging to the documents in the NIST qrels, we need to invert the preferences values of the top documents found by preference judging. We achieved this by first finding the highest preference value amongst the top documents and then subtracting each document’s preference value from this highest preference value plus 1. This way, the best documents will now have the highest preference values and the worst documents will have the lowest preference values.
2. The documents found by preference judging should be considered the best documents in the new qrels. To do this, we determined the highest preference value in the NIST qrels and then added this value to all the preference values of the top documents found by preference judging. Thus, the documents found by preference judging should now have their original preference values in the NIST qrels removed and updated such that their values are higher than the preference values of the rest of the documents in the NIST qrels.
3. We then append the top documents found by preference judging to the top of the NIST qrels.
4. We repeat these 3 steps for all 30 topics using the participant’s preference judging data. This results in the new qrels. In this study, we refer to the new qrels created using participant 1’s data as “participant1 qrels”, participant 2’s data as “participant2 qrels”, and participant 3’s data as “participant3 qrels”.

Next, the new qrels were used to calculate compatibility scores for the runs submitted in TREC 2021, with patience parameter set to default at 0.95. Note that since we only included the correct and neutral documents in the study, the new qrels only changed runs' helpful compatibility scores and not the harmful compatibility scores. Then, we plotted the new qrels' compatibility scores against NIST qrel's compatibility scores. We created the following plots:

- NIST helpful compatibility scores vs. Participant 1 helpful compatibility scores
- NIST helpful compatibility scores vs. Participant 2 helpful compatibility scores
- NIST helpful compatibility scores vs. Participant 3 helpful compatibility scores
- NIST helpful minus harmful compatibility scores vs. Participant 1 helpful minus harmful compatibility scores
- NIST helpful minus harmful compatibility scores vs. Participant 2 helpful minus harmful compatibility scores
- NIST helpful minus harmful compatibility scores vs. Participant 3 helpful minus harmful compatibility scores

From the 6 plots created, we observed that the ranking of runs produced using the participants' qrels were similar to the ranking of runs produced using NIST qrels where the ranking of runs did not change much. The runs that did well with NIST qrels were also the best runs using the participants' qrels, and the lower-ranked runs remained lower in rank. There were some slight changes in ranks between some runs using the participants' qrels, but not enough to significantly change the overall rankings. As all plots showed similar results, we only include two of the plots in this paper. Figure 4.17 shows the helpful compatibility scores of runs calculated using NIST qrels and participant 1 qrels. Figure 4.18 shows the helpful minus harmful compatibility scores for runs. We can see from the figures that there is a linear relationship between NIST compatibility scores and participant 1 compatibility scores for the ranking of runs, demonstrating that the ranking of runs are approximately the same. We also show the top-10 runs using NIST qrels and how their ranks change under each participant's qrels based on the average compatibility scores in Table 4.3 to further elaborate how the ranking of the best runs didn't change much.

Even though the compatibility plots showed that participant1, participant2, and participant3's qrels did not change the ranking of runs much compared to NIST's qrels ranking, there were still slight ranking changes that occurred. These changes mostly happened

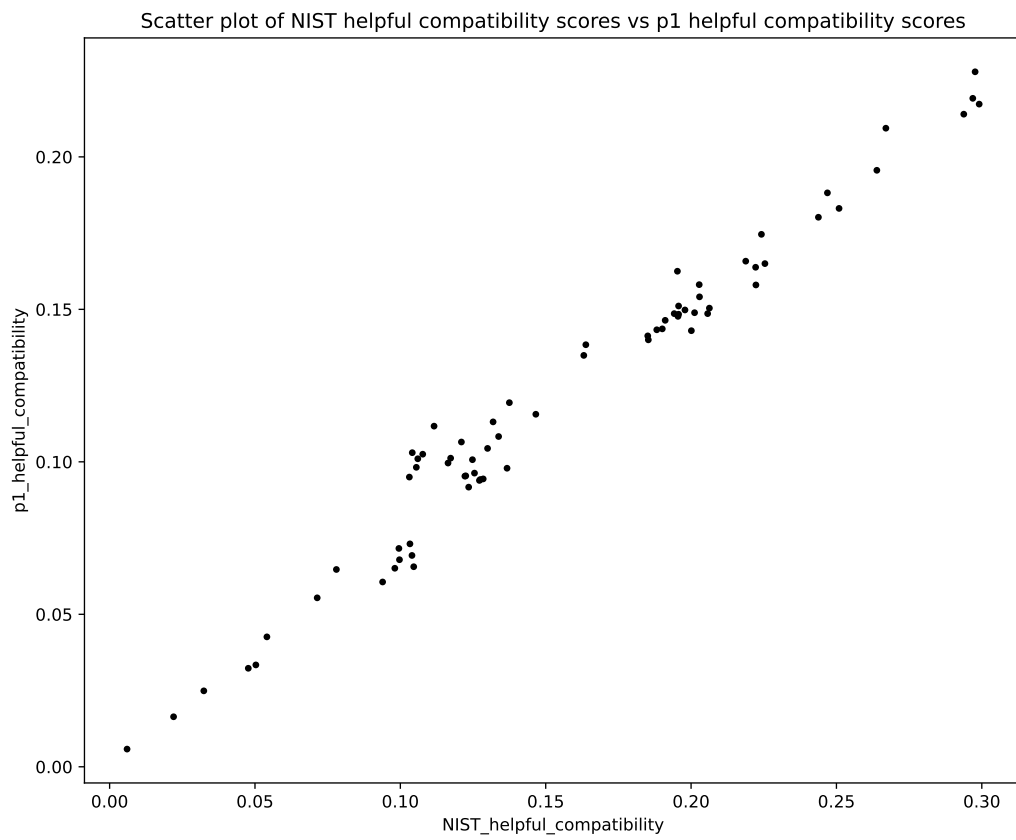


Figure 4.17: NIST Helpful Compatibility Scores vs. Participant 1 Helpful Compatibility Scores for Runs

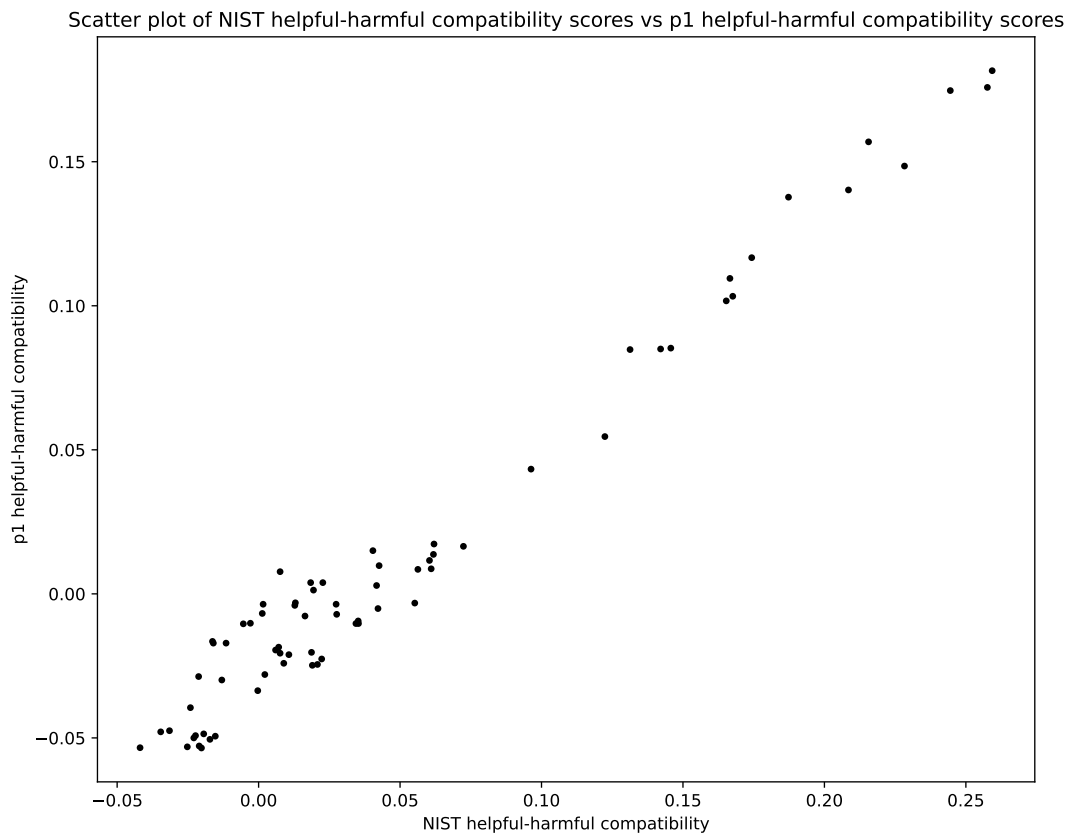


Figure 4.18: NIST Helpful Minus Harmful Compatibility Scores vs. Participant 1 Helpful Minus Harmful Compatibility Scores for Runs

Run	NIST Rank	P1 Rank	P2 Rank	P3 Rank
vera_mt5_0.5	1	1	1	1
vera_mdt5_0.5	2	2	2	2
vera_mt5_0.95	3	3	3	3
vera_mdt5_0.95	4	5	5	5
vera0	5	4	4	4
WatSMC-Correct	6	6	7	7
WatSMT-SD-S1	7	7	6	6
mt5_r	8	8	10	8
WatSMC-CALQAHC2	9	10	8	11
WatSMM-CALQAAll	10	9	9	10

Table 4.3: NIST Top 10 Runs and How Their Ranks Change Under the 3 Participants' Qrels Based on Average Compatibility Scores

between the lower-ranked runs. To determine the effectiveness of preference judging to produce the same ranking of retrieval systems with different sets of assessors, we performed 1000 simulations between our participants' rankings and the NIST ranking to obtain a distribution of Kendall's tau scores. We also performed 1000 simulations to obtain a distribution of AP correlation scores, as Kendall's tau scores did not reflect the fact that we care about the top systems' rankings more than the lower systems' rankings. Specifically, we performed the following steps:

1. Each topic in our study was preference judged by 3 participants. So for every topic, we randomly picked a participant.
2. Next, for that specific topic for all the runs, we selected that specific participant's helpful minus harmful compatibility score.
3. After all the topics for every run had a helpful minus harmful compatibility score, we calculated the average score for the run by taking the sum of all the topics' helpful minus harmful compatibility scores and dividing it by the number of topics.
4. Once all runs had an average helpful minus harmful compatibility score, we sorted the scores in descending order (largest to smallest) to create a ranking for the runs.
5. We then calculated Kendall's tau and AP correlation between this ranking and NIST ranking and store the values in two separate lists.
6. We repeated the above steps 1000 times to obtain 1000 Kendall's tau scores and 1000 AP correlation scores.
7. Finally, we plotted these scores to see the distribution of Kendall's tau scores and the distribution of AP correlation scores for our rankings.

Figure 4.19 shows the distribution of Kendall's tau scores between our participants' rankings and the NIST ranking that we ran 1000 times. The mean of the distribution is 0.784. The range of Kendall tau's values varies from 0.647 and 0.863, which shows that there is variability between different assessors. Figure 4.20 shows the distribution of AP correlation scores between participants' rankings and the NIST ranking. The mean of the distribution is 0.811. The range of AP correlation scores varies from 0.702 to 0.876. Even though the mean of the distribution is higher, the variability remains the same.

This led to the next question, "What impact does this difference in assessors have on evaluation?" To answer this question, we performed 1000 simulations to obtain distributions of Kendall's tau scores and AP correlation scores again, but this time between

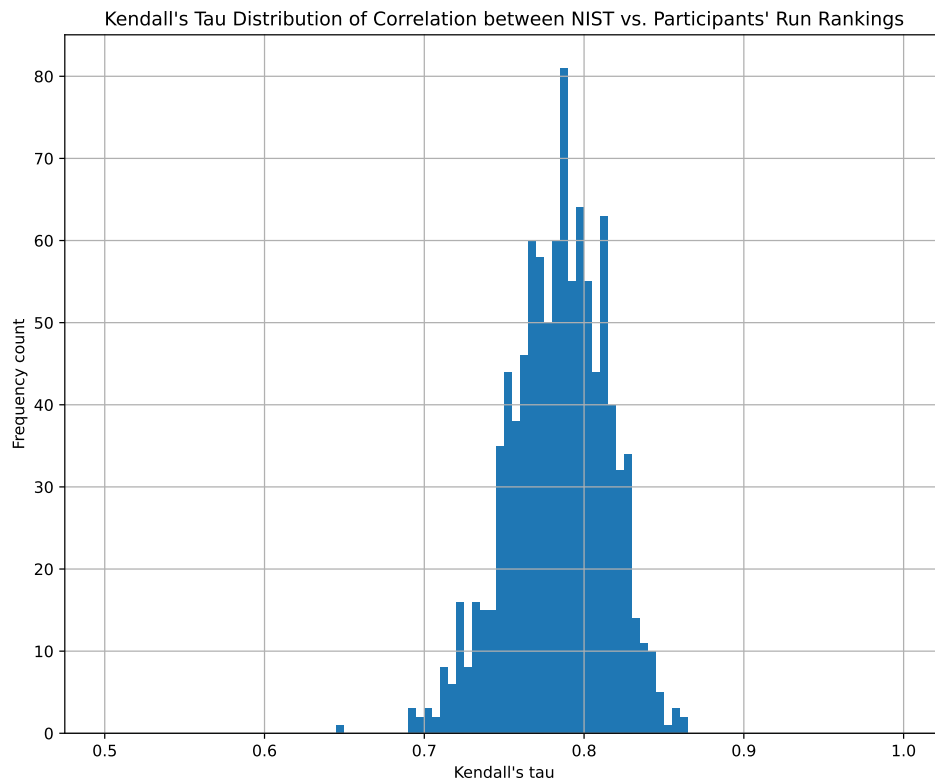


Figure 4.19: Kendall's Tau Distribution of Correlation between NIST vs. Participants' Run Rankings

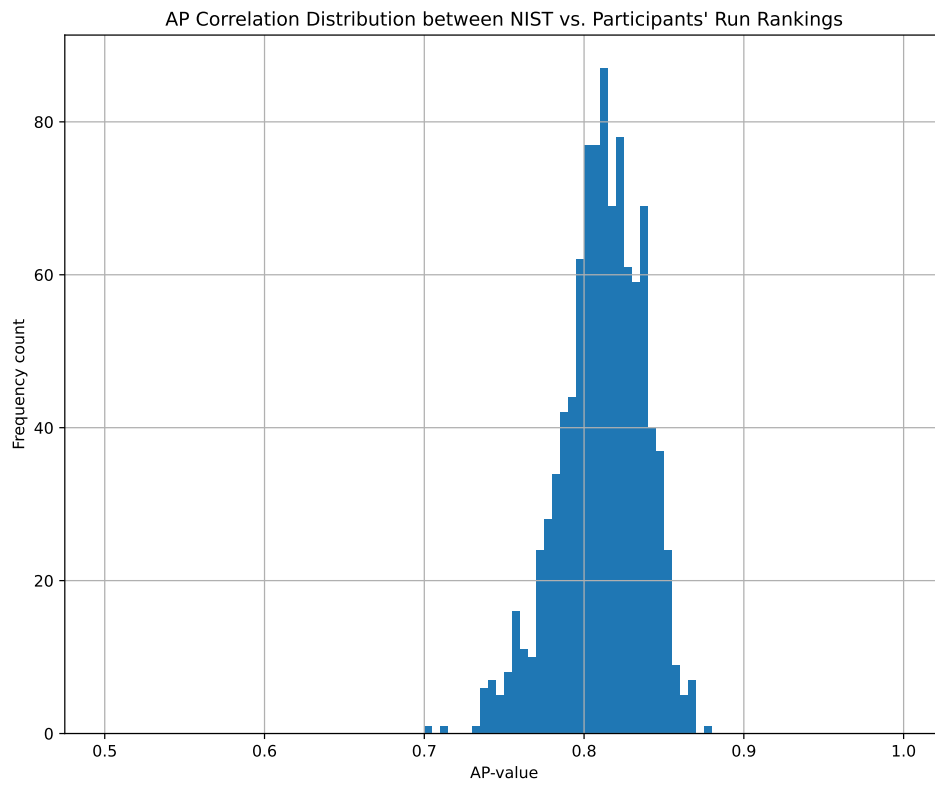


Figure 4.20: AP Correlation Distribution between NIST vs. Participants' Run Rankings

assessor versus assessor. We thus wanted to generate two assessors' rankings from the participants that we have. To do this, we randomly selected 2 out of the 3 participants that we have to be the two assessors for each topic, making sure that the same participant cannot represent both assessors for a given topic. The steps we performed are similar to that of the previous simulation, with slight modifications as follows:

1. Each topic in our study was preference judged by 3 participants. So for every topic, we randomly picked two participants to be the two assessors, making sure that the same participant cannot represent both assessors.
2. Next, for an assessor, for that specific topic for all the runs, we selected the specific participant's helpful minus harmful compatibility score.
3. After all the topics for every run for each assessor had a helpful minus harmful compatibility score, we calculated the average score for the run by taking the sum of all the topics' helpful minus harmful compatibility scores and dividing it by the number of topics.
4. Once all runs had an average helpful minus harmful compatibility score, we sorted the scores in descending order (largest to smallest) to create a ranking for the runs for each assessor.
5. We then calculated Kendall's tau and AP correlation between the ranking of one assessor and the ranking of the other assessor and stored the values in two separate lists.
6. We repeated the above steps 1000 times to obtain 1000 Kendall's tau scores and 1000 AP correlation scores.
7. Finally, we plotted these scores to see the distribution of Kendall's tau scores and the distribution of AP correlation scores for our rankings.

Figure 4.21 shows the distribution of Kendall's tau scores between one assessor and another assessor's rankings. The mean of the distribution is 0.849. The range of Kendall's tau values varies from 0.746 to 0.928. According to Voorhees [51], Kendall's tau score must be greater than or equal to 0.9 to declare that the rankings of information retrieval systems are similar. We see however, that the mean value is less than 0.9 and only the tail of the distribution is over 0.9. Thus, the rankings produced by assessors are different. Figure 4.22 shows the distribution of AP correlation scores between one assessor and another assessor's

rankings. The mean of the distribution is 0.850. The range of Kendall's tau values varies from 0.761 to 0.925. From this, we observe that even if we care more about the top systems, the mean value is still less than 0.90 and only the tail of the distribution is over 0.9, giving us the same conclusion that the rankings produced by assessors are different.

Hence, from our simulations, we reach a conclusion that having one non-expert assessor is not enough to do preference judging to evaluate information retrieval systems.

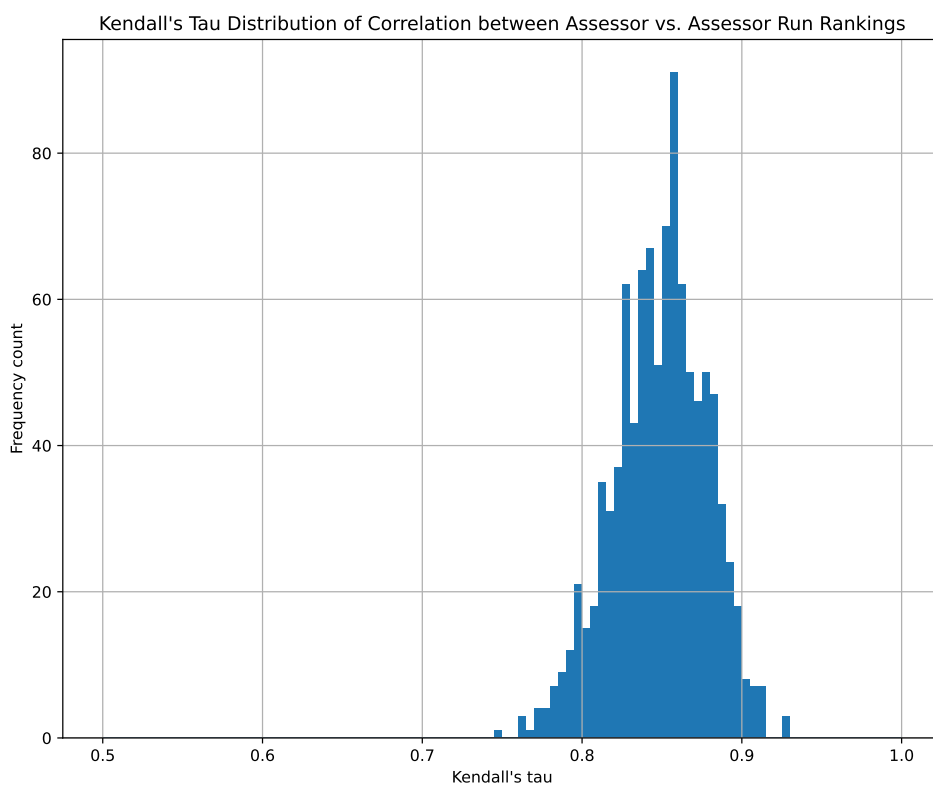


Figure 4.21: Kendall's Tau Distribution of Correlation between Assessor vs. Assessor Run Rankings

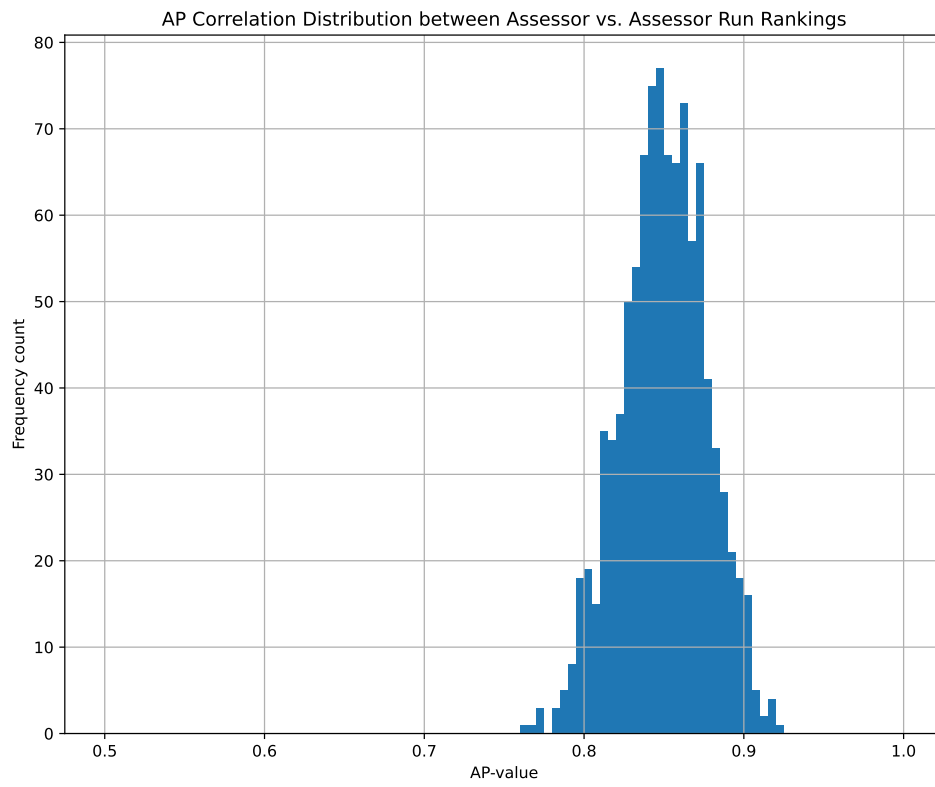


Figure 4.22: AP Correlation Distribution between Assessor vs. Assessor Run Rankings

Chapter 5

Conclusion

In this thesis, we investigated how assessors behave when preference judging, the level of agreement among multiple assessors, and whether the use of preference judging to find the top-10 documents for topics would affect the rankings of retrieval systems. A user study was conducted with 40 participants who performed preference judging tasks on 30 topics taken from the 2021 TREC Health Misinformation track. Besides the main task of preference judging, participants were also asked to re-order the top documents produced from their preference judging from most-preferred to least-preferred.

Our key findings are:

- The number of judgments needed to find the top-10 preferred documents using preference judging is about twice the number of documents in that topic.
- Participants rarely change their mind about their initial preferences, as the study shows that of the total 9940 judgments collected in the study, only 126 judgments involved an “Undo” action. Additionally, of these 126, participants maintained their original preference 94 times.
- Insufficient agreement among assessors, variable assessor self-agreement, and the variability in ranking systems as reported by our study simulations suggest that preference judging to evaluate information retrieval systems should not be done with just 1 non-professional assessor.
- According to our analysis using Kendall’s tau and AP correlation, preference judging to find the top-10 documents does significantly change the rankings of runs as compared to the rankings reported in TREC 2021 Health Misinformation Track. Most

of these changes happen between the lower-ranked runs rather than the top-ranked runs.

Our current study shows low inter-assessor agreement and varying levels of intra-assessor agreement. Additionally, different assessors produce different orderings of retrieval systems as shown by the simulations of participants' run rankings. Thus, we cannot be confident with the results produced if we re-ranked runs based on the current preference judgments that we have. More work needs to be done to create a set of preference judgments that we can be confident in. There are various possible directions we can take. One way is to conduct a user study to confirm that the preference judgments already collected are indeed favored by users. Another approach is to have professional assessors preference judge the documents in hopes there will be more agreement. Once we have a set of preference judgments that we are confident in, we can then work on addressing the question of whether preference judgments are better than graded judgments at ranking documents and evaluating retrieval systems.

References

- [1] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2006.
- [2] Alharbi, Aiman. *Studying Relevance Judging Behavior of Secondary Assessors*. PhD thesis, University of Waterloo, 2016.
- [3] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: Are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 667–674, New York, NY, USA, 2008. Association for Computing Machinery.
- [4] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1):3–23, 1999.
- [5] Maryam Bashir, Jesse Anderton, Jie Wu, Peter B Golbus, Virgil Pavlu, and Javed A Aslam. A document rating system for preference judgements. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 909–912, 2013.
- [6] Michael Bendersky, Xuanhui Wang, Marc Najork, and Donald Metzler. Learning with sparse and biased feedback for personal search. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5219–5223. IJCAI, 2018.
- [7] Abraham Bookstein. Relevance. *Journal of the American Society for Information Science*, 30(5):269–273, 1979.

- [8] Robert Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing Management*, 28(5):619–627, 1992.
- [9] Stefan Büttcher, Charles L. A. Clarke, and GV Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.
- [10] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in Information Retrieval*, pages 16–27, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.
- [11] Charles L. A. Clarke, Maria Maistro, Mahsa Seifkar, and Mark D. Smucker. Overview of the TREC 2022 health misinformation track (notebook). pages 1–8, 2022.
- [12] Charles L. A. Clarke, Maria Maistro, and Mark D. Smucker. Overview of the TREC 2021 health misinformation track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC-30)*, 2021.
- [13] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. Offline evaluation by maximum similarity to an ideal ranking. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26. ACM, 2009.
- [14] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. Assessing top-k preferences. *ACM Trans. Inf. Syst.*, 39(3), May 2021.
- [15] Charles L.A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, ICTIR '20*, page 185–192, New York, NY, USA, 2020. Association for Computing Machinery.
- [16] William S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971.
- [17] W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*, chapter 1.2. Addison-Wesley, 2010.
- [18] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. TREC CAsT 2019: The conversational assistance track overview. 2019.

- [19] Tadele T. Damessie, Falk Scholer, Kalvero Järvelin, and J. Shane Culpepper. The effect of document order and topic difficulty on assessor agreement. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, ICTIR '16*, page 73–76, New York, NY, USA, 2016. Association for Computing Machinery.
- [20] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382, 1971.
- [21] H.P. Frei and P. Schäuble. Determining the effectiveness of retrieval algorithms. *Information Processing Management*, 27(2):153–164, 1991.
- [22] Simon French. *Decision Theory—An Introduction to the Mathematics of Rationality*. Ellis Horwood Limited, Chichester, 1986.
- [23] Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1003–1004. ACM, 2013.
- [24] Kai Hui and Klaus Berberich. Transitivity, time consumption, and quality of preference judgments in crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 105–106. ACM, 2014.
- [25] Kai Hui and Klaus Berberich. Low-cost preference judgment via ties. In Joemon M. Jose and et al., editors, *Advances in Information Retrieval*, volume 10193 of *Lecture Notes in Computer Science*, pages 626–632. Springer International Publishing AG, 2017.
- [26] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. *SIGIR forum*, 36(2):41–48, 2002.
- [27] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7–es, apr 2007.
- [28] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pages 781–789. ACM, 2017.
- [29] Gabriella Kazai, Emine Yilmaz, Nick Craswell, and S.M.M. Tahaghoghi. User intent and assessor disagreement in web search evaluation. In *Proceedings of the 22nd ACM International Conference on Information amp; Knowledge Management, CIKM '13*, page 699–708, New York, NY, USA, 2013. Association for Computing Machinery.

- [30] Jinyoung Kim, Gabriella Kazai, and Imed Zitouni. Relevance dimensions in preference-based ir evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, page 913–916, New York, NY, USA, 2013. Association for Computing Machinery.
- [31] M.E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4(4):343–359, 1968.
- [32] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. Considering assessor agreement in IR evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, page 75–82, New York, NY, USA, 2017. Association for Computing Machinery.
- [33] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.
- [34] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. In Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, editors, *Introduction to Information Retrieval*, chapter 3.3.4, pages 73–75. Cambridge University Press, Cambridge, UK, 2008.
- [35] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.
- [36] Shuzi Niu, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. Top-k learning to rank: Labeling, ranking and evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 751–760, New York, NY, USA, 2012. Association for Computing Machinery.
- [37] Taemin Kim Park. The nature of relevance in information retrieval: An empirical study. *The Library Quarterly: Information, Community, Policy*, 63(3):318–351, 1993.
- [38] Kira Radinsky and Nir Ailon. Ranking from pairs and triplets: Information quality, evaluation methods and query complexity. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 105–114, New York, NY, USA, 2011. Association for Computing Machinery.
- [39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

- [40] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
- [41] Mark E. Rorvig. The simple scalability of documents. *Journal of the American Society for Information Science*, 41(8):590–598, 1990.
- [42] Ian Ruthven, Leif Azzopardi Glasgow, Mark Baillie, Ralf Bierig, Emma Nicol, Simon Sweeney, and Murat Yakici. Intra-assessor consistency in question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 727–728, New York, NY, USA, 2007. Association for Computing Machinery.
- [43] Tetsuya Sakai and Zhaohao Zeng. Good evaluation measures based on document preferences. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 359–368, New York, NY, USA, 2020. Association for Computing Machinery.
- [44] Tefko Saracevic. 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. *Journal of the American Society for Information Science and Technology*, 58:2126, 01 2007.
- [45] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison. A study of information seeking and retrieving. *Journal of the American Society for Information Science*, 39(3):161–176, 1988.
- [46] Falk Scholer, Andrew Turpin, and Mark Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1063–1072, New York, NY, USA, 2011. Association for Computing Machinery.
- [47] Mahsa Seifikar. A preference judgment interface for authoritative assessment. Master's thesis, University of Waterloo, 2023.
- [48] David Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 4th edition, 2007.

- [49] Ruihua Song, Qingwei Guo, Ruochi Zhang, Guomao Xin, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. Select-the-best-ones: A new way to judge relative relevance. *Information Processing Management*, 47(1):37–52, 2011.
- [50] Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.
- [51] Ellen M. Voorhees. Evaluating the evaluation: A case study using the TREC 2002 question answering track. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 260–267, 2003.
- [52] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, 2005.
- [53] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4), nov 2010.
- [54] Fen Xia, Tie-yan Liu, and Hang Li. Statistical consistency of top-k ranking. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.
- [55] Xinyi Yan, Chengxi Luo, Charles L. A. Clarke, Nick Craswell, Ellen M. Voorhees, and Pablo Castells. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 567–577, New York, NY, USA, 2022. Association for Computing Machinery.
- [56] Ziyang Yang, Alistair Moffat, and Andrew Turpin. Pairwise crowd judgments: Preference, absolute, and ratio. In *Proceedings of the 23rd Australasian Document Computing Symposium*, ADCS '18, New York, NY, USA, 2018. Association for Computing Machinery.
- [57] Y.Y. Yao. Measuring retrieval effectiveness based on user preference. *Journal of the American Society for Information Science*, 46(2):133–146, 1995.
- [58] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 587–594, New York, NY, USA, 2008. Association for Computing Machinery.

- [59] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 587–594, New York, NY, USA, 2008. Association for Computing Machinery.
- [60] Zhaohui Zheng, Keke Chen, Gordon Sun, and Hongyuan Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 287–294, New York, NY, USA, 2007. Association for Computing Machinery.
- [61] Dongqing Zhu and Ben Carterette. An analysis of assessor behavior in crowdsourced preference judgments. 2010.

APPENDICES

Appendix A

Preference Judging Instructions Manual

Relevance Judging for Evaluation of Health Search Guidelines

Version: 12 Sept 2022

Overview

The purpose of this study is to see if comparing two documents next to each other and selecting the preferred one will create a better final ranking of documents than scoring documents one-by-one separately from each other. In this study, participants will use a specially designed document judging platform. For a specific query given to them, they will be shown two documents at a time and will be asked to keep selecting the document that they prefer until a final ranking of documents is created.

Detailed Instructions

We need your help to determine which documents would help a searcher (someone who is using a search engine) determine the correct answer to a health-related question. For different tasks, each task will have a unique question, for example, “Does yoga improve the management of asthma?”. We will tell you what the correct answer is for each question and present you with documents that should contain the correct answer, but the documents will be of different quality. You will use a preference judging system to compare pairs of documents. The preference judging system will show two documents side by side, their URLs, and provide you tools for finding and marking relevant material in each document. Preference judging systems work by asking you to compare pairs of documents and select the preferred document.

When comparing two documents, you are to **prefer the document that would best help the searcher reach a correct decision.**

When deciding which document to prefer, consider which document you would want a search engine to show you before the other. Usually, we want to see **well written documents** that **focus on the question at hand** from **credible sources**. Both the quality of the answer and the credibility of the answer source are important, but if in doubt, prefer a more credible website over a less credible site.

- If you **clearly prefer** one document to the other, then you should record your preference.
 - Keep in mind that sometimes you may simply prefer one document to the other because its source is more trusted by you. Other times, you might prefer the content of one document to the other.
 - It might seem trivial, but trust your gut feelings about preferences. It is okay to make fast decisions if you know you prefer one to the other.
 - **If you have a reason to prefer one document to the other, then record your preference.**
- If you find yourself **struggling to say one is better than the other**, you should say they are **equal**.

Sometimes, you will see two documents that look identical or very similar. This is not a mistake

and you should mark them as equal. Always highlight the title of the document first to help you know that you've seen it previously. Then highlight the portions of the document which you feel are useful for you to make your decision. You must highlight portions of the document that helped you reach your decision. Later in the instructions we detail how to highlight portions of a document.

You are encouraged to take into consideration all factors that you think would matter to a searcher and **influence the searcher to make a correct decision**. In addition to containing a correct answer, factors may include, but are not limited to the following:

- Quality of explanation for the answer, i.e. searchers may make better decisions when a document has a correct answer with an explanation and reasoning as opposed to simply having the correct answer.
- Presentation quality. Is the answer and document written in a manner that is easy to read and comprehend?
- Some documents will have more expertise, authoritativeness, and trustworthiness¹. For example, www.cdc.gov has high amounts of expertise, authoritativeness, and trustworthiness. If two documents seem to contain the same information, but one has more credibility, you would assume that the more credible document would influence the searcher more and be preferred.
- An informative document from a credible source would be preferred to a document that is for advertising or marketing purposes.
- Documents written by experts would be preferred to those by non-experts.
- The whole document context should be considered. For example, a single correct sentence embedded in a document filled with scam health treatments is less likely to influence a searcher to make a correct answer than a document filled with credible information.

If you come across a document that contains an incorrect answer, please prefer the other document with its correct answer.

Please note that we will provide you with the URL (web address such as <http://www.uwaterloo.ca/>) of the original web page to help you understand the source of the document. You are encouraged to judge pages based on the document's text and URL without clicking on the URL. In many cases, it is important to consider the URL to help you judge the quality and credibility of the source. URLs from sites you know to be of high quality are better than URLs from unknown or suspicious sites.

The preference judging system will look as follows:

¹ The idea of understanding the purpose of a website before judging its quality, determining the amount of expertise, authoritativeness, and trustworthiness (E-A-T), and the cdc.gov example of high E-A-T are ideas based on Google's General Guidelines for search evaluators: <http://static.googleusercontent.com/media/www.google.com/en//insidesearch/howsearchworks/assets/searchqualityevaluatorguidelines.pdf> . Last Accessed: 17/12/2018)

[Undo](#) Preference: [Left](#) [Equal](#) [Right](#) Progress: 0.0% [Topic Information](#)

<https://www.everydayhealth.com/esophageal-cancer/does-selenium-prevent-esophageal-cancer.aspx> <https://p.lein.no/pages/research>

Title: Does Selenium Prevent Esophageal Cancer? | Everyday Health
 Document ID: en.noclean.c4-train.01158-of-07168.106659

[\[print-Logo\]](#)
 Search
 Log in My Profile
 Your Profile

- * Following Topics
- * Saved Items
- * Newsletters
- * Tools
- * My Daily Crohn's
- * My Daily RA
- * My Daily Diabetes
- * Settings
- * Logout

 Subscribe Menu

Main Menu

Conditions

- * Atrial Fibrillation
- * Cold and Flu
- * Depression
- * Heart Failure
- * High Cholesterol
- * Multiple Sclerosis
- * Psoriasis
- * Psoriatic Arthritis
- * Rheumatoid Arthritis
- * Type 2 Diabetes
- * Ulcerative Colitis
- * View All
- * Drugs A-Z
- * Symptom Checker

Healthy Living

Title: Research - Plein
 Document ID: en.noclean.c4-train.01563-of-07168.66897

Free shipping on all orders.

(0)
 [Plein]
 * Shop
 * Sign in
 * Cart
 *

By skeptics, for skeptics.

Filter by ingredient

- * All
- * Bone Support
- * Caffeine
- * Immune System
- * Caffeine
- More Info

Caffeine

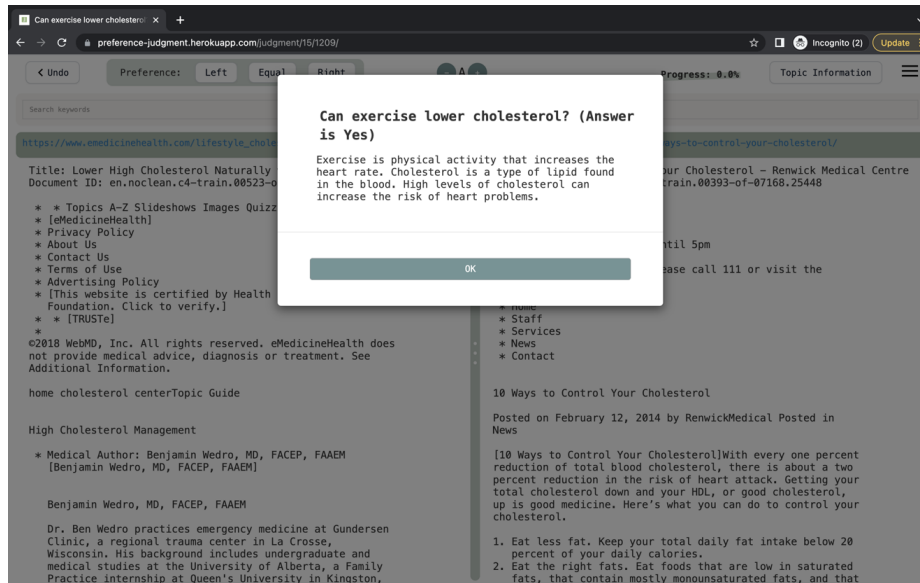
- o Alertness
- o Caffeine
- o Cognitive functions

Found in:
 Leaves and fruits of certain plants.

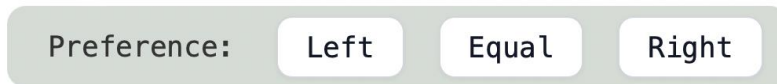
Do you need it?
 It sure feels very good those early mornings, when the 2-o'clock dip hits or when working late. Actually, it feels good to be alert and awake when you need to.

Consequences of not getting enough
 Zombie state has been observed.

You should first click on the "Topic Information" button that will bring up the task's "topic question" and background information. Whenever you need a refresher about the topic, click on this button to review the question and information.



You record your preference judgment using the preference widget:

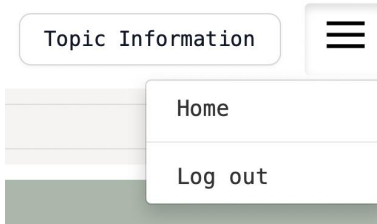


Thus, if the left document is more likely to influence a searcher to make a correct decision, you would click on "Left", and similarly if the right document is better you would click on "Right". If the documents are the same or near-duplicates with the same source, etc. then you should judge them "Equal".

If after making a judgment, you decide it was a mistake, you can go back to the previous judgment pair using the "Undo" button:



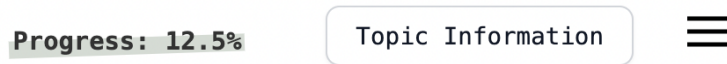
To go back to the task/topic selection page, you can click on the three horizontal lines in the upper right corner and select "Home". You can also log out whenever you want.



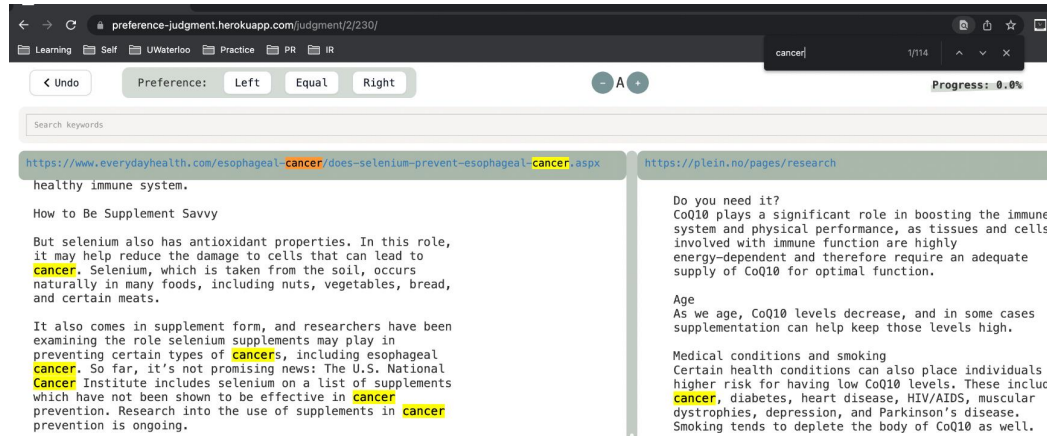
You can increase the size of text in the documents by the following feature in the middle of the header. It will be kept during the judgment process.



The progress number is beside the topic information button in the right corner, indicating approximately how many percent of judgments are done until all documents are entirely judged.



A fast way to find occurrences of a single keyword in the documents is to use the web browser's "find in page" search feature, which is brought up by typing CTRL-F. For example in Google Chrome's browser it will pop up a widget that allows you to enter a keyword and then use up and down arrow buttons to find the next or previous occurrence. The search will first go through the left document, and then move on to the right document. For example:



The judging system also offers the means to enter search keywords and phrases to highlight all occurrences automatically in documents. Keywords can only contain numbers and letters, and

they are not case sensitive. In the location that says "Search keywords", you can type a keyword or phrase and then press the "Enter" key to add that as a word to highlight:

The screenshot shows a search interface with a search bar containing the keywords: cancer x health x calcium x cLock x plants x. The progress indicator shows 0.0%. Two document snippets are displayed side-by-side.

Left Snippet (everydayhealth.com):

Researchers in China studied **health** data from people who had esophageal **cancer** and compared the information with people who were **cancer**-free. They found a correlation between those who ate foods higher in selenium and lower rates of esophageal **cancer**. The researchers concluded that selenium supplementation could help protect against esophageal **cancer**, particularly for people who do not have enough selenium in their diet.

A recent analysis out of the Netherlands found similar results. It reviewed extensive data on 120,852 middle-aged adults and found that those with higher selenium levels – as demonstrated by tests of their toenail clippings – were less likely to develop esophageal **cancer**. Study results showed that women, people who had never smoked, and those who generally had low intakes of antioxidants were most likely to benefit from selenium supplementation.

Scientists have observed that **cancer** rates are lower where there are higher levels of selenium in the soil, suggesting a tenuous link between the presence of selenium in the food supply and **cancer** risk.

Common Questions About Diet and **Cancer**

How Much Selenium to Take

Food sources can provide the daily recommended amount of selenium, which is 70 micrograms (mcg) per day. **Cancer** prevention has been linked to supplemental levels as high as 200 mcg.

You may be at risk for low levels of selenium if you smoke, use birth control pills, drink alcohol, or have **health** problems such as Crohn's disease that make it hard to absorb the nutrition you need from your food.

There are some risks to taking too much selenium – over 400 mcg for adults – per day, including:

Right Snippet (plein.no):

Title: Research - Plein
Document ID: en.noclean.c4-train.01563-of-07168.66097

Free shipping on all orders.

(0)
[Plein]
* Shop
* Sign in
* Cart
*
By skeptics, for skeptics.

Filter by ingredient

- * All
- * Bone Support
- * Caffeine
- * Immune System
- * Caffeine
- More Info

Caffeine

- o Alertness
- o Caffeine
- o Cognitive functions

Found in:
Leaves and fruits of certain **plants**.

Do you need it?
It sure feels very good those early mornings, when 2-o'cLock dip hits or when working late. Actually, feels good to be alert and awake when you need to.

Consequences of not getting enough
Zombie state has been observed.

Besides highlighting keywords through the search box, in the judging system, you can highlight sentences and paragraphs in the documents by mouse down, drag, and mouse up. The system keeps highlighted part of the documents until the end of the judgment session. You also can remove the highlighted part with the mouse by clicking and dragging over a highlighted section.

Sometimes when you select text, it stays selected rather than being converted to a highlight. If this happens, you can simply click the selected text and then it will change to highlighted text.

You should always mark in documents the material that you think a searcher will find useful to make a decision about the search question. By marking the relevant material in a document, you will be able to compare that document faster the next time you see it. You should expect to see some documents many times as you are asked to compare it to many other documents. You do not need to read documents in their entirety. You should read and search the document for what you think is relevant material that a normal searcher would use to help them make a decision. As you know, searchers do not waste their time reading non-relevant material, and nor should you, but it is important that your preference judgments be as accurate as possible while not taking a very long time for you to make a judgment. **Try to balance speed and**

accuracy while judging as fast as possible while still making accurate preference judgments.

To help you keep track of which documents you have already seen, first **highlight the title of the document at the top of the document**. When the document comes up again in the system for a different preference comparison, you'll know you've already seen the document and can scroll to your other highlighting.

The highlighting by the mouse has more priority than the search keywords. For example, in the following picture, "selenium" was entered in the search box, but when the user highlighted the first sentence in the left document, it turned yellow. If you delete highlighting, it will reveal any search keywords.



To log into the system, you will be given a username and password by the researcher. (You can't create a new account.)



JUDGO

Username*

Username

Password*

Password

Remember Me

Sign In

Click [here](#) to make a new account.

© 2022-

You will be assigned to one or more tasks/topics, which will appear in a dropdown when you log in. You can start with any task/topic. If you log out or switch topics, any work that you've done will be retained.



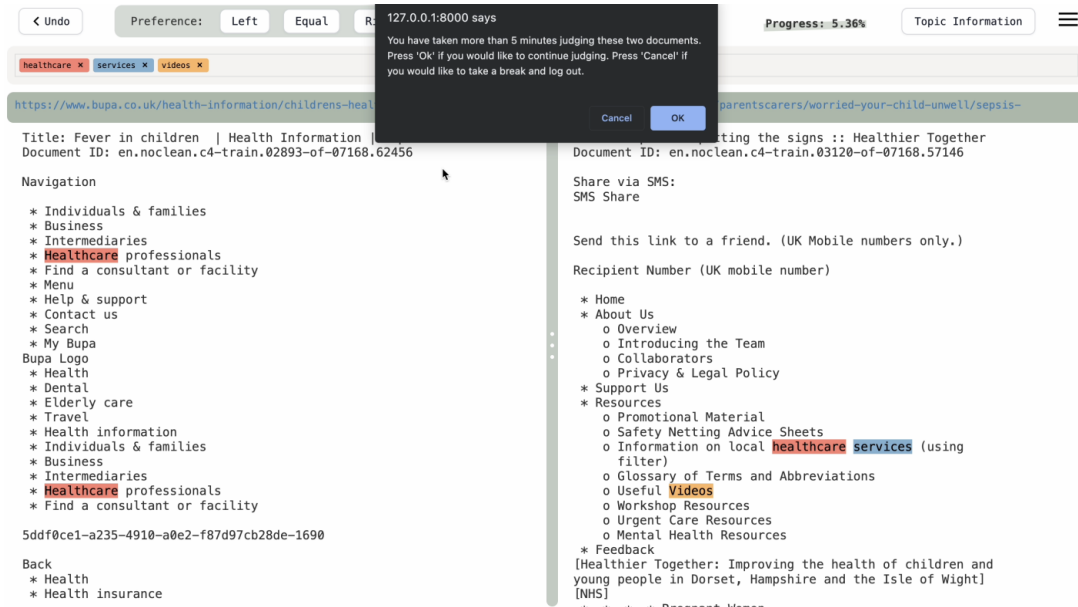
JUDGO

✓ Can exercise lower cholesterol? (Answer is Yes)

Can fish oil improve your cholesterol? (Answer is No)

Start Judgment!

If you would like to take a break, please log out of the system. If you spend more than 5 minutes judging a pair of documents or would like to take a break but forgot to log out, a popup will appear asking you to confirm if you would like to continue judging or not. Please click "Ok" if you would like to continue judging and click "Cancel" if you would like to log out.



Sometimes the system might not show a popup. Instead, it will redirect you to the “Sign Out” page. In this case, please click the browser’s “Back” button if you would like to continue judging or click “Sign Out” to log out to take a break.

Re-ranking Documents

After you complete the preference judging session for a topic, we will present you with 10 randomly ordered documents from those that you have judged and ask you to order them from most-preferred to least-preferred. You will be able to click and view each document and change the order of the documents and decide on a final ranking for these documents from rank 1 to 10, where rank 1 is the best document and rank 10 is the least. Documents cannot be ranked equal in this part of the study.

Other Notes

If you have any questions or problems during your study session, please contact the student researcher Linh Nhi Phan Minh at lnphanmi@uwaterloo.ca.

Appendix B

Code for Calculating RBO scores (rbo.py)

```
# (2*intersection of S and T to depth d)/(|S to depth d| + |T to depth d|)

# Code modified from:
# https://github.com/claclark/Compatibility/blob/master/compatibility.py

def rbo(listA , listB , p=0.75, depth=10):
    score = 0.0
    normalizer = 0.0
    weight = 1.0
    for i in range(1,depth+1):
        listA_set = set_to_depth_d( listA , i )
        listB_set = set_to_depth_d( listB , i )
        score += weight * ((2.0 * len(listB_set.intersection(listA_set))) /
                           (len(listA_set)+ len(listB_set)))
        normalizer += weight
        weight *= p
    return score/normalizer

def set_to_depth_d( theList , d):
    if d < 1:
        raise ValueError("depth_d_must_be_greater_than_0")
    return set( elements_to_depth_d( theList , d) )
```

```
def elements_to_depth_d( theList , d):
    if d < 1:
        raise ValueError("depth_d_must_be_greater_than_0")
    elts_to_depth = []
    for elt in theList:
        if isinstance(elt , list):
            elts_to_depth.extend(elt)
        else:
            elts_to_depth.append(elt)

    if len(elts_to_depth) >= d:
        return elts_to_depth

    return elts_to_depth
```

Appendix C

Code for the Rest of the Data Analysis

All code written for the analysis of collected data in this thesis can be found in the “rbo” folder (written by Professor Mark D. Smucker) and “analysis” folder (written by Linh Nhi Phan Minh) via this link: <https://github.com/UWaterlooIR/separate-vs-preference-judgments>.