Determining the Utility of Key-term Highlighting for High Recall Information Retrieval Systems

by

(Jean) Xue Jun Wang

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, 2021

Author's Declaration

I hereby declare that I am the sole author of this 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.

Abstract

High-recall information retrieval (HRIR) is an important tool used in tasks such as electronic discovery ("eDiscovery") and systematic review of medical research. Applications of HRIR often uses a human as its oracle to determine the relevance of immense numbers of documents, which is expensive in both time and money. Various methods for reducing the amount of time spent per assessment and improving the quality of assessors have been proposed to improve these systems.

For this thesis, we examine the method of presenting documents where key-terms are highlighted in place of plain-text document. This is commonly accepted as a positive feature which achieves both of the previously mentioned improvements, but there is currently a lack of empirical evidence to support its effectiveness. We describe an user study in which participants are assigned to one of two variations of a HRIR system (key-term highlighting vs plain-text) with a post task questionnaire. Our results failed to show statistically significant improvement for labelling documents with key-term highlighting over plain-text for any of the measures recall, precision, and F1, but may negatively affect retention of concepts.

Our study provides empirical evidence for how the use of key-term highlighting affects an assessor's abilities to label documents and provides insight into when including this feature may be harmful rather than helpful.

Acknowledgements

I would like to thank my supervisor Maura R. Grossman for her guidance and support throughout my research, as well as both encouraging and accommodating me to research a topic of my interest.

I would also like to thank my readers Gordon V. Cormack and Charles L. A. Clarke for reading my thesis.

I want to thank my boyfriend, Kevin Hyun, not only for his endless love and support but for also being an amazing research partner.

Finally, I would like to thank all of my colleagues for their support during my graduate studies, especially Mustafa Abualsaudall for his guidance on Gathera, and Anmol Singh for all our insightful discussions and proof reading my thesis.

Dedication

This thesis is dedicated to my parents. I would not be where I am without you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The objective of high-recall information retrieval (HRIR) is to find all or nearly all relevant documents. An important application of HRIR include electronic discovery ("eDiscovery"), which is the process where both parties in a case are required to find all or nearly all relevant material from their own document collections and provide it to the opposing party. Traditionally, an attorney would be required to review each document in a collection, taking minutes per document [15]. With the digitization of information, the sizes of document collections to search through have grown rapidly, reaching between 600,000 to 1 million documents per case [24]. Other applications include systematic review of medical research, and the construction of information retrieval (IR) test collections.

In all of these applications of HRIR, a human is used as an oracle to determine the relevance of documents. Users of HRIR systems are often required to provide immense numbers of relevance assessments, thus causing it to be both time and money consuming. In most cases, this human effort is the primary cost of high-recall tasks, which makes methods for reducing this cost highly desirable. Some methods that have been developed include reducing the total number of relevance assessments required from assessors, reducing the amount of time spent per assessment, and improving the quality of the assessors [31].

One popular method for reducing the number of labelled data needed is *active learning* (AL), which is an iterative method where the least certain unlabelled data point is assessed by an oracle and added to the set of labelled data. Different studies have shown that active learning models can effectively maintain accuracy while significantly reducing human labelling efforts [21]. *Continuous active learning* (CAL) [6, 7] is a variant of AL specifically designed for technology-assisted review (TAR), which requires retrieval and review of the substantial majority of relevant documents from a collection. The major difference between

CAL and AL is that the data points are queried by highest and lowest certainty respectively (see Section 2.2).

For this thesis, we examine the effects of presenting documents where key-terms are highlighted in place of plain-text documents for the assessment of relevance as part of a high-recall retrieval system based on continuous active learning (CAL). Key-term highlighting is commonly accepted in information retrieval (IR) systems, both academically and commercially, as a positive feature. Several examples include:

- Bulls-eye System of Zheng et. al. [33], which highlights passages to "help speed up decision about the relevance or non-relevance of a given document"
- SPIDER Retrival System of Knaus et. al. [13], who report highlighting to be "an efficient and very effective tool, especially for concentrating users' attention to the most important parts of the document".
- Google Search, which highlights search terms and relevant passages when displaying their results [16].

Our work is motivated by the lack of empirical evidence supporting the effectiveness of highlighting. Compounding to this problem, previous research in psychology and pedagogy of text highlighting show conflicting results. While some studies [9] show that highlighting can be beneficial, other studies suggest that highlight either has little effect [27, 12], or even negative effect [19, 22], particularly when highlighting is done passively (see Section 2.1). A further motivation is the result of our team's participation in the TREC 2020 COVID Track [29]. We used the same HRIR system for all five rounds, with key-term highlighting being implemented only after the first round. While all members of the team unanimously agreed that this feature was helpful, objectively, there was little change in the overall performance from the first round to the second. These results form the basis of the question we address: **since reading text with highlighted information not done by the readers themselves have been shown to negatively, if at all, impact our abilities to make inferences of the material, information retention, as well as comprehension of the text as a whole, does including key-term highlighting as a feature in the IR system yield the assumed benefits?**

To study the assumed hypothesis that, for a given amount of time, users will be able to find a greater number of relevant documents if they judge documents with key-terms highlighted in place of plain-text documents, we designed a study where we compare the performance of users with and without key-term highlighting while assessing documents using a CAL system. Our results showed that labelling documents with key-term highlighting had **no statistically significant improvement** over plain-text (see Section 4), suggesting that the benefits of highlighting are subjectively perceived through users, rather than objectively improving performance.

Chapter 2

Background and related work

In this chapter we review related works in psychology, information retrieval (IR), and high-recall information retrieval (HRIR). First, we start by reviewing early studies in applying highlighting to reading material in the field of psychology. We then summarize previous works of highlighting in the context of IR. Finally, we provide background on the high-recall methods we use - continuous active learning (CAL).

2.1 Review of Highlighting in Psychology

There has been a considerable amount of research focused on the utility of highlighting in learning and how it can effect the retention and understanding of text material. While specific implementation of highlighting differ, key-words or phrases are considered to be highlighted if a noticeable emphasis is placed on them. Popular highlighting techniques from text typesetting include background colouring, changing the font weight (bold face), and underlining words and phrases of interest. We also define highlighting as *active* if the user chooses the words and phrases to highlight, *passive* otherwise.

In a study done by Fowler and Barker [9], 78 undergraduate students were tasked with reading excerpts from scientific articles and completing a retention test based on their readings one week after. Furthermore, the participants were split into four groups; one group actively highlighted while studying, another group studied from the highlighted text created by the previous group. The third group studied from texts highlighted by the experimenter, while the last group utilized no highlighting. The results of this study showed that participants who actively highlighted scored better than the other groups, and that readers who studied from previously highlighted material scored better when the readers had "maximum faith that the highlighter could discriminate between important material and trivia."

In contrast, a similar study done by Wade and Trathen [27] showed that noting of ideas in a text (underlining, highlighting, or taking notes) made little to no difference. Furthermore, when a similar study was conducted with ten-year-old school children [19], the participants performed worse on conceptual post-reading questions with highlighting than without. The negative effects of highlighting was the most pronounced when highlighting was passive and the participant were given the post-questions before reading.

Other studies have focused solely on passive highlighting, which is more applicable to our context. In a study done by Klare et. al. [12], participants tasked with reading a 1206-word text document from an aircraft mechanics training course for 20 minutes and asked 50 multiple choice comprehension questions afterwards. Furthermore, the text was presented to participants in one of three condtions: unpatterned/plain text, highlighting all words that would later appear in the comprehension test, and highlighting important words, regardless to whether they appear in the comprehension test. While participants showed an improvement in short term retention by using highlighting, the difference was small and not practical significant.

Silvers and Kreiner [22] further considers an adversarial setting, where the experimenters highlight trivial or insignificant (inappropriate) words and phrases. The results showed that while no significant differences were observed between the groups that used appropriate highlighting and no highlighting, participants performed worse with inappropriate highlighting.

An widely accepted explanation for both the beneficial and harmful effects of highlighting is von Restorff effect [18], which predicts that when people are presented with multiple simultaneous homogeneous stimuli, the stimulus that differs from others is more likely to be remembered. Applied to our context, this effect predicts that readers are more likely to remember highlighted text than the surrounding non-highlighted text. The act of highlight can further compound this effect since the act of deciding what to and not to mark alone leads to processing textual information at a deeper level [30]. When highlighting is done appropriately and competently, the von Restorff effect predicts that the participants will remember information that is beneficial to answering the post-task questions over the insignificant details. Conversely, when highlighting is done inappropriately or incompetently, the participants may have to try harder to recall the necessary information. However, the overall effects of highlighting seems to be small as most studies did not find significant differences between highlighting and not highlighting.

2.2 Active Learning and Continuous Active Learning

Active learning (AL) is a type of that combines supervised learning with an oracle that provides labels when queried, with the aim of minimizing the number of labelled data points while maintaining accuracy (or other measures). In this setting, the data is partition into two disjoint sets: labelled data $\mathcal{L}$ and unlabelled $\mathcal{U}$. At each iteration, we train a model $\mathbf{M}$ using $\mathcal{L}$, then use $\mathbf{M}$ to label $\mathcal{U}$. A point in $\mathcal{U}$ is chosen by $\mathbf{M}$ (by some metric) to be queried and added to $\mathcal{L}$. This process stops typically when adding new labels no longer improves the accuracy.

There are many queries strategies, but some common strategies are:

- uncertainty sampling: the point which the model is least certain about are labeled (ex. closest to the decision boundary)
- query by committee: several models are trained and the point for which causes the most disagreement among the model is labeled
- variance reduction: the point which would minimize the variance of the output is labeled.

While AL, with a variety of strategies, have been shown to be successful in its goal [21], the use of AL in technology-assisted review (TAR) in electronic discovery (e-discovery) raises critical issues, particularly determining a stopping point for the learning algorithm (also known as the "stabilization issue"). In legal settings, the stopping point must be justified in order to ensure that "a reasonable review has been conducted, and that burden or expense of continuing the review would outweigh the benefit of any additional documents" [6]. In order to improve AL in the context of TAR, Cormack and Grossman [6] propose using a variant they call Continuous Active Learning (CAL). The main difference between CAL and traditional AL is that the documents with the highest certainty are chosen to be labeled, rather than the lowest. The authors show that CAL outperforms AL with uncertainty sampling, while avoiding the stabilization issue.

There have been several versions and implementations of the CAL system since it was first developed. One of the most commonly known is the Baseline Model Implementation (BMI), a version of CAL [7] which employs logistic regression implemented by Sofia ML as the underlying machine learning model. HiCAL, developed by Abualsaud et. al., builds on BMI and adds a graphical user interface [2]. Finally, Gathera, developed by one of the members of the HiCAL team, is an open source project which builds and improves on HiCAL. This is the system we chose to use and modify for our study.

2.3 Statistical Tools

2.3.1 Performance measures

As mentioned in the introduction, high-recall information retrieval has a variety of applications, two major ones being electronic discovery and systematic review. In both eDiscovery and systematic review, the goal is to find all the relevant documents to a particular topic. In eDiscovery, a missed document could lead to legal issues, while in systematic review, it could affect the conclusion of the work. Therefore, a good first measure to look at is *recall*, which is the fraction of all the relevant documents found by an assessor:

$$recall = \frac{|U_{rel} \cap R|}{|R|}$$

where U_{rel} is the set of documents judged by the assessor as relevant, and R is the true set of relevant documents. We chose to look at recall rather than simply the number of relevant documents an assessor found as different search tasks (different topics) have different total numbers of relevant documents - recall normalizes for this.

Both eDiscovery and systematic review tasks often involve two passes of the relevance judgements, first by someone qualified to broadly identify relevant material (less expensive), then again by an expert (more expensive) who examines only material deemed relevant during the first pass to make final determinations about the documents. For example, in systematic review, a graduate student may be tasked in finding all the relevant literature to a topic for their lead researcher, who then takes the result and make the final decisions on what is truly relevant. In these cases, each non-relevant document that makes it through the first pass wastes more time of the expensive expert reviewer; therefore, our second measure looks at *precision*, which is the fraction of relevant documents identified by the assessor which was actually correct:

$$precision = \frac{|U_{rel} \cap R|}{|U_{rel}|}$$

Finally, the F_1 measure combines recall and precision, and captures the trade-off between them:

$$F_1 = \frac{2 \times recall \times precision}{recall + precision}$$

2.3.2 Statistical testing

Statistical hypothesis testing is commonly employed in order to determine whether the mean performance of one group exceeds another. The two-sample t-test computes the confidence interval of the difference in means of two populations by the following test statistic

$$T = \frac{m_1 - m_2}{S_p\sqrt{n_1^{-1} + n_2^{-1}}},$$

where

- m_i: mean of population i
- n_i: size of population i
- s_i: sample standard deviation of population i
- S_p: pooled estimate of common standard deviation $S_p = \sqrt{\frac{(m_1-1)s_1^2+(m_2-1)s_2^2}{m_1+m_2-2}}$

Then, using the value t from the student-t distribution with significance level α and degree of freedom ν, we can construct an $1 - \alpha$ confidence interval

$$\left[(m_1 - m_2) \pm tS_p\sqrt{n_1^{-1} + n_2^{-1}}\right],$$

which estimates where the true difference between the means may lie. If zero exists in this interval, we cannot reject the null hypothesis and there exists no statistical significance between the means of the two populations.

Chapter 3

Study Design

In this section, we describe our experiment in detail. We describe the search topics and document collection, the study design, the high-recall system and its implementation, and other details of the experiment including how we measure performance and determine statistical significance.

3.1 Corpus and Topics

We use the TREC 2020 COVID Track [25] test collection for our search topics and documents. The track's task was ad-hoc retrieval of documents from the CORD-19 data-set [28], which contains over 500,000 scientific papers and scholarly articles on COVID-19 and related historical coronavirus research. Of the set 50 NIST assessed topics we use the first 20 topics.

These topics were written by the organizers of TREC-COVID with biomedical training [1] with inspiration drawn from consumer questions submitted to the National Library of Medicine (NLM), "medical influencers" on social media, and Twitter using the #COVID-Search tag. They range from general public questions such as "what is the origin of COVID-19" (topic number 1) to very specific scientific ones such as "are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19?" (topic number 20).

3.2 Study Participants

A recruitment e-mail for study participants was sent to the University of Waterloo's computer science graduate students list (D.1). However, participants were not limited to those on this e-mail list as recruitment had passed through word of mouth from one participant to others. When interested subjects did not violate any inclusion/exclusion criteria for the study (D.2), we chose to allow them to participate.

A total of 41 subjects participated in the study, however one subject was removed due to incomplete data collection, yielding a total of 40 participants that we use for our data analysis. The study was performed in 41 2-hour sessions with a single participant per session. Each participant was provided $30 as remuneration at the end of their session.

Of the 40 participants who completed the study, 33 were of age between 20-29, and 6 between 30-39. There were 30 male and 10 female; 37 who studies science, technology, engineering, or math, and 2 who studies arts. Furthermore, 20 had completed a masters' degree, and 18 a bachelors' degree. Not every category adds up to 40 as participants were allowed to skip questions. Figure 3.1 contains a summary of the above data.

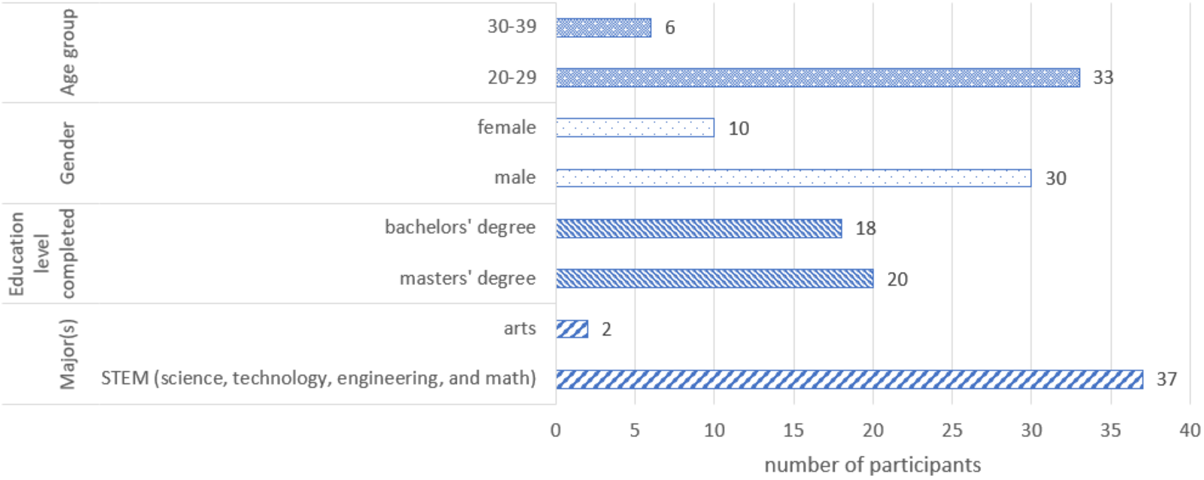


Figure 3.1: Summary of demographic questionnaire data

3.3 User Study Procedure

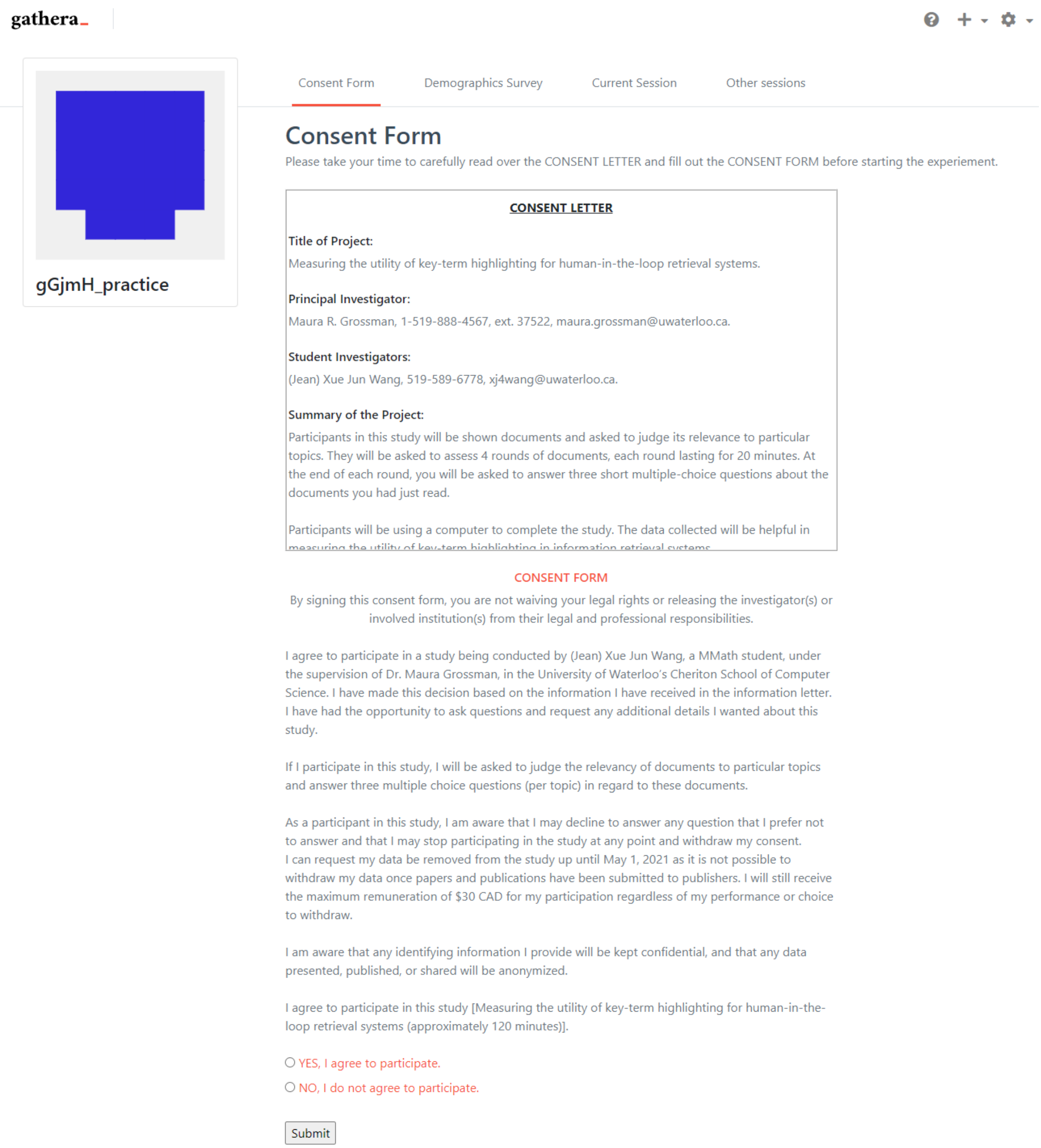
Before proceeding with their assigned tasks, participants were asked to digitally sign a consent form and complete a basic demographic questionnaire. Figure 3.2 shows the interface participants used to give consent, Figure 3.3 shows the interface they used to fill in their demographic information, and Table 3.1 contains all the questions collected in the demographics questionnaire.

Questions	
DQ1:	Age group
DQ2:	Gender
DQ3:	Education level completed
DQ4:	Major(s)

Table 3.1: Demographic questions

After concluding the administrative portion of the study, each participant underwent a virtual tutorial (via screen sharing over Microsoft Teams) walking them through how to use the system interface and what they were expected to perform at each task. One topic was used to give participants practice making graded relevance judgements and answering the post-task questions as part of the tutorial. This topic was not part of their assigned group of 4, as shown in Table 3.2, and is not counted in the final data.

Participants were given one of two different variations of a high-recall retrieval system to find as many relevant documents as possible within 20 minutes. In one variation, participants judged paragraph-length excerpts of documents shown in plain text. In the other variation, excerpts were displayed with various key terms, as determined by the machine learning model, highlighted in yellow. Throughout the rest of the paper, we will refer to each of the variations by their shorthand: **CAL-C** for the plain text variation (the control condition) and **CAL-H** where various key terms in the documents excerpts are highlighted. Figures 3.4 and 3.5 show the interfaces participants used for **CAL-C** and **CAL-H** respectively. The user interface was developed on top of the open source project Gathera (Section 2.2) and was hosted on a University of Waterloo server. Furthermore, the 20 topics were divided into 5 groups of 4 topics each at random using a random number generator (RNG). Each of the 5 groups of 4 topics were also assigned an additional unique topic, at random using a RNG, to be used during the tutorial. Participants were randomly assigned a topic group (on top of being assigned one of **CAL-C** and **CAL-H**). The topic groups are summarized in Table 3.2.



Consent Form

Please take your time to carefully read over the CONSENT LETTER and fill out the CONSENT FORM before starting the experiement.

CONSENT LETTER

Title of Project:

Measuring the utility of key-term highlighting for human-in-the-loop retrieval systems.

Principal Investigator:

Maura R. Grossman, 1-519-888-4567, ext. 37522, maura.grossman@uwaterloo.ca.

Student Investigators:

(Jean) Xue Jun Wang, 519-589-6778, xj4wang@uwaterloo.ca.

Summary of the Project:

Participants in this study will be shown documents and asked to judge its relevance to particular topics. They will be asked to assess 4 rounds of documents, each round lasting for 20 minutes. At the end of each round, you will be asked to answer three short multiple-choice questions about the documents you had just read.

Participants will be using a computer to complete the study. The data collected will be helpful in measuring the utility of key-term highlighting in information retrieval systems.

CONSENT FORM

By signing this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

I agree to participate in a study being conducted by (Jean) Xue Jun Wang, a MMath student, under the supervision of Dr. Maura Grossman, in the University of Waterloo's Cheriton School of Computer Science. I have made this decision based on the information I have received in the information letter. I have had the opportunity to ask questions and request any additional details I wanted about this study.

If I participate in this study, I will be asked to judge the relevancy of documents to particular topics and answer three multiple choice questions (per topic) in regard to these documents.

As a participant in this study, I am aware that I may decline to answer any question that I prefer not to answer and that I may stop participating in the study at any point and withdraw my consent.
I can request my data be removed from the study up until May 1, 2021 as it is not possible to withdraw my data once papers and publications have been submitted to publishers. I will still receive the maximum remuneration of $30 CAD for my participation regardless of my performance or choice to withdraw.

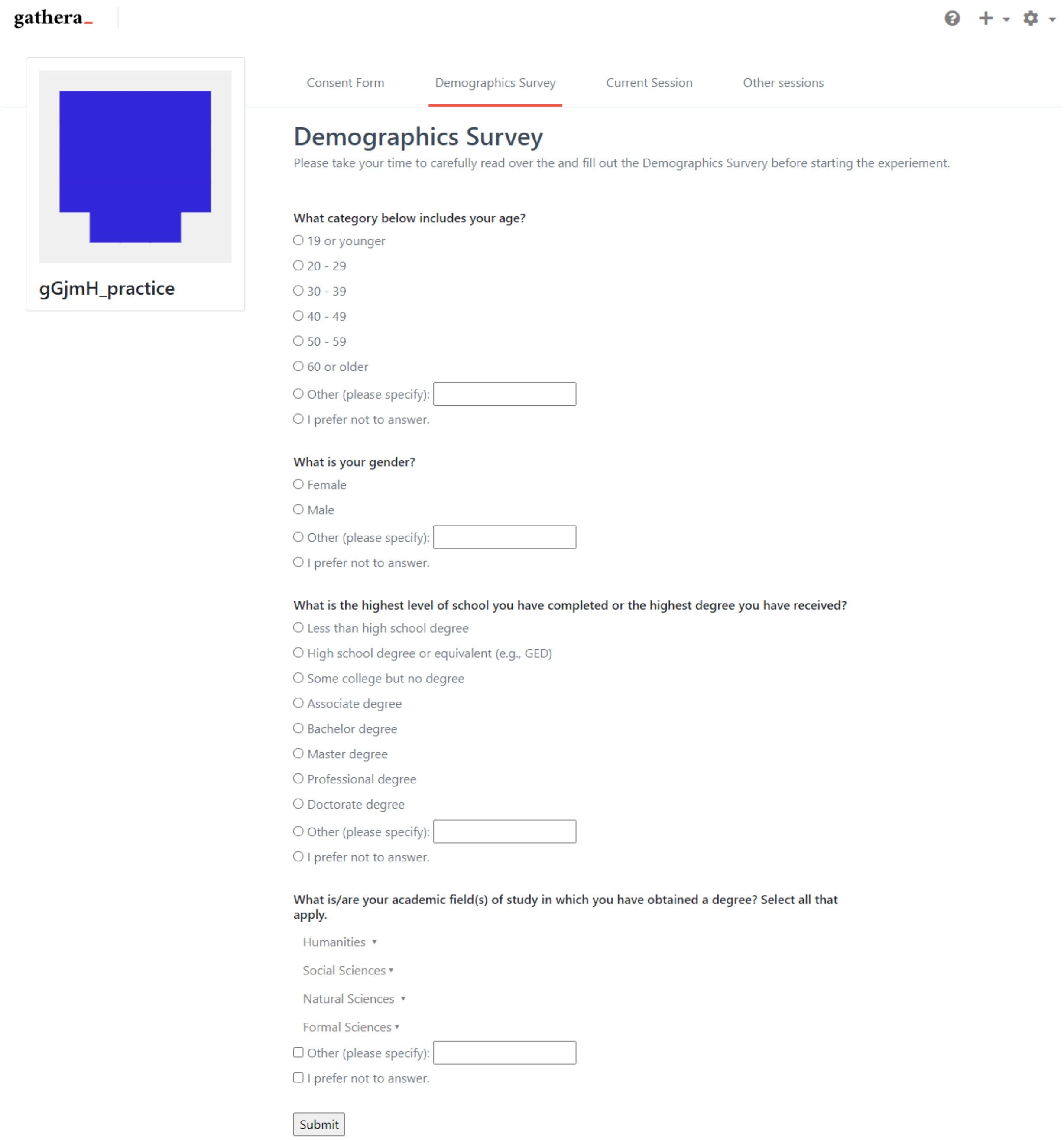
I am aware that any identifying information I provide will be kept confidential, and that any data presented, published, or shared will be anonymized.

I agree to participate in this study [Measuring the utility of key-term highlighting for human-in-the-loop retrieval systems (approximately 120 minutes)].

○ YES, I agree to participate.

○ NO, I do not agree to participate.

Figure 3.2: Consent interface



Demographics Survey

Please take your time to carefully read over the and fill out the Demographics Survery before starting the experiement.

What category below includes your age?

○ 19 or younger
○ 20 - 29
○ 30 - 39
○ 40 - 49
○ 50 - 59
○ 60 or older
○ Other (please specify): ______
○ I prefer not to answer.

What is your gender?

○ Female
○ Male
○ Other (please specify): ______
○ I prefer not to answer.

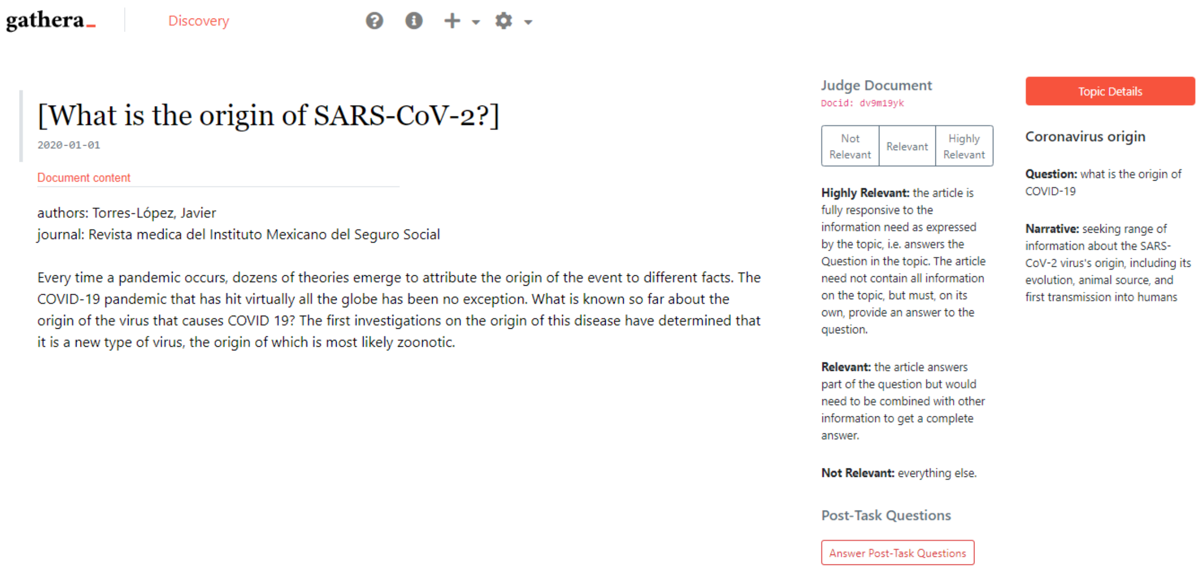
What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school degree
○ High school degree or equivalent (e.g., GED)
○ Some college but no degree
○ Associate degree
○ Bachelor degree
○ Master degree
○ Professional degree
○ Doctorate degree
○ Other (please specify): ______
○ I prefer not to answer.

What is/are your academic field(s) of study in which you have obtained a degree? Select all that apply.

Humanities ▾
Social Sciences ▾
Natural Sciences ▾
Formal Sciences ▾
□ Other (please specify): ______
□ I prefer not to answer.

Figure 3.3: Demographics interface



[What is the origin of SARS-CoV-2?]

2020-01-01

authors: Torres-López, Javier
journal: Revista medica del Instituto Mexicano del Seguro Social

Every time a pandemic occurs, dozens of theories emerge to attribute the origin of the event to different facts. The COVID-19 pandemic that has hit virtually all the globe has been no exception. What is known so far about the origin of the virus that causes COVID 19? The first investigations on the origin of this disease have determined that it is a new type of virus, the origin of which is most likely zoonotic.

Judge Document

Docid: dv9m19yk

Highly Relevant: the article is fully responsive to the information need as expressed by the topic, i.e. answers the Question in the topic. The article need not contain all information on the topic, but must, on its own, provide an answer to the question.

Relevant: the article answers part of the question but would need to be combined with other information to get a complete answer.

Not Relevant: everything else.

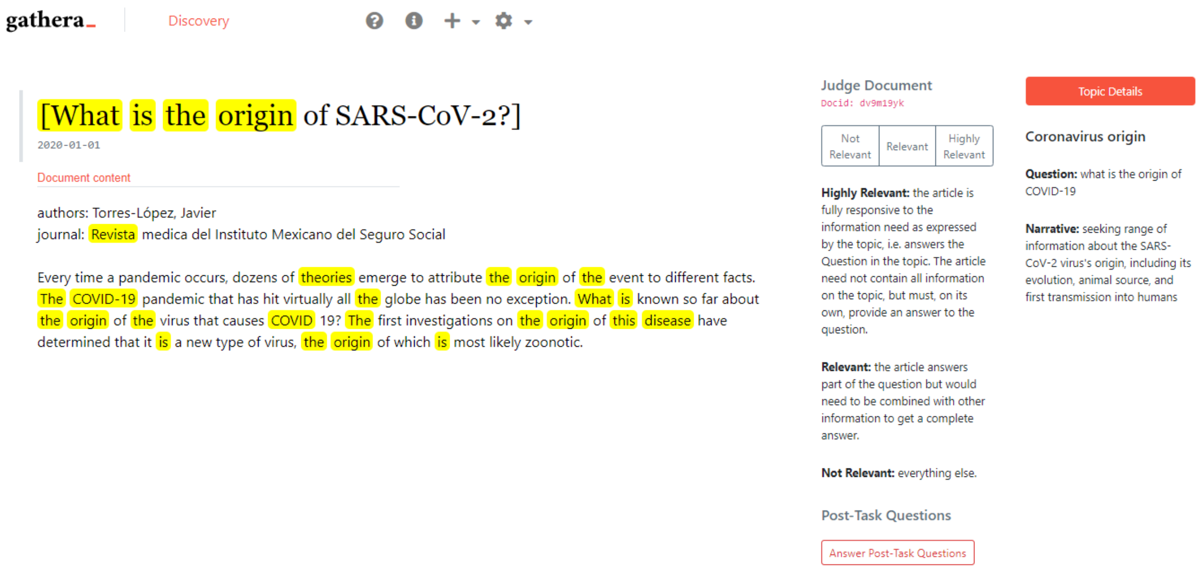
Post-Task Questions

Coronavirus origin

Question: what is the origin of COVID-19

Narrative: seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans

Figure 3.4: CAL-C interface



[What is the origin of SARS-CoV-2?]

2020-01-01

authors: Torres-López, Javier
journal: Revista medica del Instituto Mexicano del Seguro Social

Every time a pandemic occurs, dozens of theories emerge to attribute the origin of the event to different facts. The COVID-19 pandemic that has hit virtually all the globe has been no exception. What is known so far about the origin of the virus that causes COVID 19? The first investigations on the origin of this disease have determined that it is a new type of virus, the origin of which is most likely zoonotic.

Judge Document

Docid: dv9m19yk

Highly Relevant: the article is fully responsive to the information need as expressed by the topic, i.e. answers the Question in the topic. The article need not contain all information on the topic, but must, on its own, provide an answer to the question.

Relevant: the article answers part of the question but would need to be combined with other information to get a complete answer.

Not Relevant: everything else.

Post-Task Questions

Coronavirus origin

Question: what is the origin of COVID-19

Narrative: seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans

Figure 3.5: CAL-H interface

Group	**1**	**2**	**3**	**4**	**5**
Topic 1 of 4	7	2	3	1	6
Topic 2 of 4	9	4	5	10	14
Topic 3 of 4	12	11	8	13	17
Topic 4 of 4	16	15	18	20	19
Practice Topic	1	8	11	14	15

Table 3.2: Topic Groupings

As part of the tutorial, participants were instructed to follow Voorhees' definitions for graded relevance levels of highly relevant, relevant, and non-relevant [26]. Accordingly, a document was classified as:

- highly relevant if they believe that it can provide an answer to the topic question on its own
- relevant if they believe that it contains helpful information which can answer part of the topic question, but would need to be combined with other information to get a complete answer
- non-relevant otherwise.

Participants were not provided with additional information about what labels to give to specific document, outside of these official definitions, to minimize experimenter bias. Furthermore, the participants were asked to "work as quickly and accurately as they can" with no additional instructions to reduce bias.

Participants worked on their own computers in whatever location or environment they preferred, which was largely due to the COVID-19 restrictions. Any variation across the participants is random and should not bias the results. A benefit of allowing participants to work from their preferred environment is that it mimics how crowd-sourced workers might perform at this task. After completing the study, participants where remunerated $30 for their participation through electronic transfer.

After the tutorial, participants proceeded with the remainder of the study on their own with guidance from the experimenter about when to move onto the next stage. Each participants completed 4 iterations of labelling documents and answering post-task questions (see Subsection 3.3.1) using a unique search topic each time. Participants were encouraged to attempt to finished each iteration in one sitting and take breaks in between. They were not allowed to return back to a task once it is completed and were only allowed to do each task consecutively. For each of the task iterations, participants used the same system variation (either **CAL-H** or **CAL-C**). By the end of the experiment, each system variation had been applied four times to each of the 20 topics.

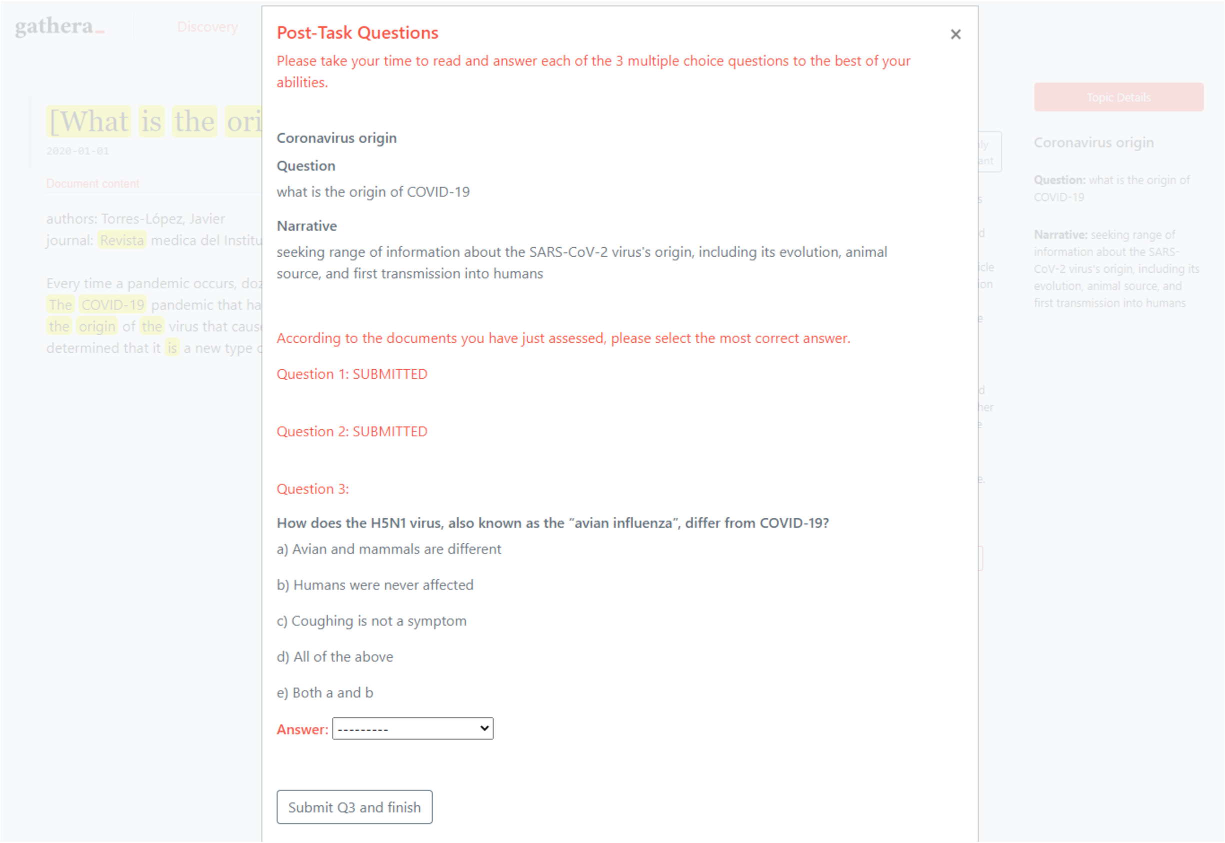
3.3.1 Post-task questions

After completing 20 minutes of document assessments, participants proceeded to the post-task questionnaires which consists of three multiple choice questions corresponding to the topic they were assigned. Figure 3.6 shows the interface participants used to complete the post-task questionnaires.

Post-task assessment questions were introduced in our study to measure knowledge acquired after assessing document excerpts during labelling. Each of the three questions in a post-task questionnaire addresses a level of Bloom's revised learning taxonomy [5]. Each question and its intended assessment is described in Table 3.3. Each of the three multiple choice questions were given one at a time as later questions sometimes contained answers for earlier ones. Table C.1 shows the repertoire of all questions with their corresponding topic, multiple choice options, and canonical answers highlighted in yellow.

Cognitive Questions	
Q1 Remembering:	*Recall specific facts* The correct answer, which exists verbatim in the corpus of relevant documents, is mixed amongst incorrect answers, which do not exist in the corpus. Participants can rely on recall and recognition to select the correct answer.
Q2 Understanding:	*Grasp meaning of instructional material* Both the correct answer and incorrect answers exist verbatim within the corpus of relevant documents. Participants will need to understand each of the multiple choice options to select the correct answer.
Q3 Applying/Analyzing:	*Use the information in a new (but similar) situation/ Take apart the known and identify relationships* The correct answer does not exist verbatim within the corpus of relevant documents. Participants will need to make inferences about the information to choose the correct answer.

Table 3.3: Post-task questionnaire's correlation to Bloom's Taxonomy knowledge domains


gathera
Discovery
[What is the ori
2020-01-01
Document content
authors: Torres-López, Javier
journal: Revista medica del Institu
Every time a pandemic occurs, doz
The COVID-19 pandemic that ha
the origin of the virus that caus
determined that it is a new type c
Topic Details
Coronavirus origin
Question: what is the origin of COVID-19
Narrative: seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans
×

Post-Task Questions

Please take your time to read and answer each of the 3 multiple choice questions to the best of your abilities.

Coronavirus origin

Question

what is the origin of COVID-19

Narrative

seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans

According to the documents you have just assessed, please select the most correct answer.

Question 1: SUBMITTED

Question 2: SUBMITTED

Question 3:

How does the H5N1 virus, also known as the "avian influenza", differ from COVID-19?

a) Avian and mammals are different

b) Humans were never affected

c) Coughing is not a symptom

d) All of the above

e) Both a and b

Answer: ---------

Submit Q3 and finish

Figure 3.6: Post-task questionnaire interface

Chapter 4

Results and Discussion

4.1 Results

The two-sample t-test (Section 2.3.2) failed to reveal any statistically significant differences between the two groups **CAL-C** and **CAL-H**. While the sample means shows that participants in **CAL-H** performed slightly better across all three performance measures, the 95% confidence intervals (which estimates the true differences between the means of the two groups) all include zero; thus, the null hypothesis that highlighting will have no difference in recall, precision, and F1 scores cannot be rejected.

Tables 4.1 and 4.2 shows the results of the T-test for each of the two groups, **CAL-C** and **CAL-H**, respectively. Table 4.3 shows the result of the two tailed two-sample T-test, with **CAL-C** as population 1 and **CAL-H** as population 2. Positive values on this table implies that the participants in **CAL-C** outperformed (on average) participants in **CAL-H** for that specific measure, conversely underperformed for negative values.

measure	mean	95%-CI
Recall	0.04283	[0.0325, 0.0531]
Precision	0.48670	[0.4203, 0.5531]
F1	0.07421	[0.0580, 0.0905]

Table 4.1: Two Tailed T-test of **CAL-C**

measure	mean	95%-CI
Recall	0.04317	[0.0334, 0.0530]
Precision	0.52349	[0.4591, 0.5879]
F1	0.07520	[0.0597, 0.0907]

Table 4.2: Two Tailed T-test of **CAL-H**

measure	mean difference	95%-CI
Recall	-0.00034	[-0.0144, 0.0137]
Precision	-0.03679	[-0.1279, 0.0543]
F1	-0.00099	[-0.0231, 0.0211]

Table 4.3: Two Tailed Two-sample T-test (**CAL-C** vs **CAL-H**)

An auxiliary measurement for participant performance was the score on the post-task questions. To measure this, each participant was awarded one mark for every answer that matched the canonical one (C.1), zero otherwise. Figure 4.1 shows the average score (in percent) obtained by participants in each level using each of the system variations **CAL-H** and **CAL-C**. Table 4.4 shows the result of the two tailed two-sample T-test for the post-task questionnaire scores. The test failed to reveal any statistically significant differences between the two groups.

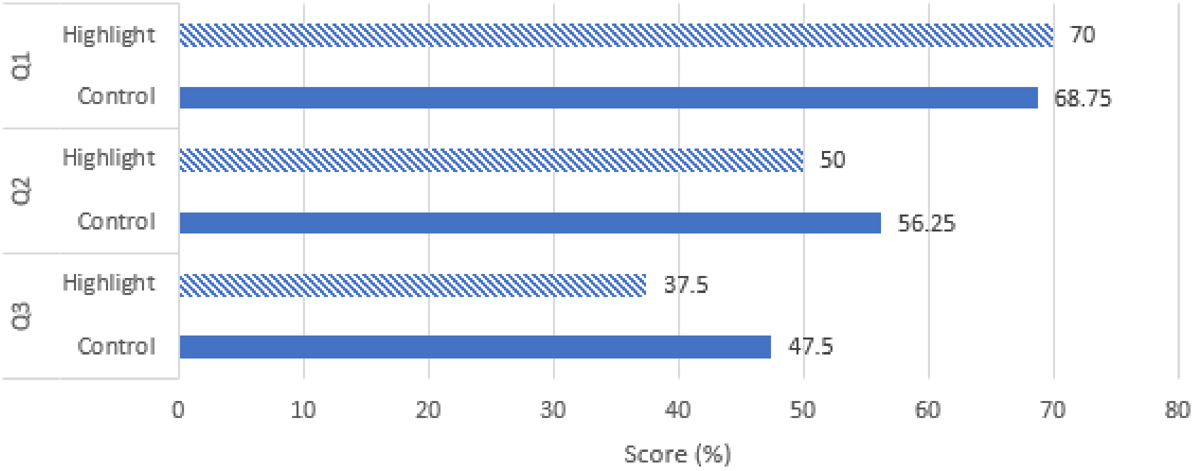


Figure 4.1: Post-task questionnaire data

Question	mean difference	95%-CI
Q1	-1.25	[-12.3, 9.8]
Q2	6.25	[-7.37, 19.87]
Q3	10	[-2.825, 22.825]

Table 4.4: Two Tailed Two-sample T-test (**CAL-C** vs **CAL-H**) for Post-task questionnaire score (%)

4.2 Discussion

The results of our study provides evidence against the hypothesis that highlighting can enhance the performance of users of human-in-the-loop information retrieval systems. In particular, our analysis shows that highlighting had no statistically significant improvement over plain-text for any of the measures recall, precision, and F1. This is inline with previous studies done in psychology on the effects of highlighting, many of which found that the effects of highlighting to be insignificant (see Section 2.1). One possible reason that highlighting had little effect on the participants of our study could be because a user's willingness to trust passively highlighted text is directly related to her faith in the highlighter [9]. When asked, participants were only informed that the key-terms that the underlying system finds important are highlighted; this may have led participants to disregard the highlighted text, thus reducing its effects.

Related to the users' confidence of the system's ability to highlight, one important aspect that we were unable to measure is the effects of highlighting during longer tasks. While our experiment setup reflects a crowd-sourced setting, it does not reflect what would happen during e-discovery, where several hours may be spent reviewing documents. As the user reviews more documents provided by the system, she will learn more about the topic as well as what the system has learned, which could increase her confidence in the system's highlighting and thus its effect. We will leave investigating this as future work.

The results of post-task questions suggest that key-term highlighting may negatively affect a user's ability to answer conceptual questions about documents she read, mirroring the results of [19, 22] in particular. As expected, participants scored better in factual level-one questions than the more conceptual (levels 2-3) questions – this is true for both groups, but is more pronounced in **CAL-H** than **CAL-C**. As previously discussed, this could be explained by von Restorff effect: since key-term highlighting only highlights important terms rather than conceptually important passages, it may hamper the users' abilities to recall concepts over facts. However, this effect does not seem to affect recall, precision, and F1 scores – perhaps implying that conceptual understanding is less important for relevancy judgements (using Voorhees' definitions).

Chapter 5

Conclusion

Overall, we found little statistical evidence that highlighting was beneficial in improving users' recall, precision, and F1 scores. Our results also indicate that key-term highlighting could have a negative effect on the users' abilities to answer conceptual questions based on the documents they reviewed. However, this does not necessarily mean that designers of high-recall systems should avoid the use of highlighting all together, but rather, designers should be mindful when to use it. For example, if high-level conceptual understanding is desirable as a part of the systemic review of documents (ex. a research assistant performing total-recall for the primary investigator), the use of highlighting may be inappropriate. However, as [2] shows, users often perceive key-term highlighting as an useful feature. If the inclusion of highlighting as a feature makes the system more likely to be used (or to be used for longer) and high recall is the most important measure, then its usage may be appropriate. While the full effects of highlighting in HRIR systems remains to be studied, our study calls into question the assumption that key-term highlighting is beneficial for all situations.

References

[1] Trec-covid, 2020.

[2] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. A system for efficient high-recall retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1317–1320, 2018.

[3] TH Anderson and BB Armbruster. Studying in pd pearson (ed), handbook on reading research (pp. 657-679), 1984.

[4] Henry F Arnold. The comparative effectiveness of certain study techniques in the field of history. *Journal of Educational Psychology*, 33(6):449, 1942.

[5] Benjamin S Bloom et al. Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, 20(24):1, 1956.

[6] Gordon V Cormack and Maura R Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 153–162, 2014.

[7] Gordon V Cormack and Maura R Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv:1504.06868*, 2015.

[8] Gordon V Cormack and Maura R Grossman. Waterloo (cormack) participation in the trec 2015 total recall track. In *TREC*, 2015.

[9] Robert L Fowler and Anne S Barker. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, 59(3):358, 1974.

[10] Stephen M Harding, W Bruce Croft, and C Weir. Probabilistic retrieval of ocr degraded text using n-grams. In *International Conference on Theory and Practice of Digital Libraries*, pages 345–359. Springer, 1997.

[11] Tereza Iofciu, Nick Craswell, and Milad Shokouhi. Evaluating the impact of snippet highlighting in search. *Understanding the User-Logging and Interpreting User Interactions in Information Search and Retrieval (UIIR-2009)*, page 44, 2009.

[12] George R Klare, James E Mabry, and Levarl M Gustafson. The relationship of patterning (underlining) to immediate retention and to acceptability of technical material. *Journal of applied psychology*, 39(1):40, 1955.

[13] Daniel Knaus, Elke Mittendorf, Peter Schauble, and Paraic Sheridan. Highlighting relevant passages for users of the interactive spider retrieval system. In *Proceedings of the fourth text retrieval conference (TREC-4)*, pages 233–244, 1998.

[14] Chester O Mathews. Comparison of methods of study for immediate and delayed recall. *Journal of Educational Psychology*, 29(2):101, 1938.

[15] Douglas W Oard and William Webber. Information retrieval for e-discovery. *Information Retrieval*, 7(2-3):99–237, 2013.

[16] Amit J Patel et al. Systems and methods for highlighting search results, January 4 2005. US Patent 6,839,702.

[17] Sarah E Peterson. The cognitive functions of underlining as a study technique. *Literacy Research and Instruction*, 31(2):49–56, 1991.

[18] H. V. Restorff. Ueber die wirkung von bereichsbildungen im spurenfeld. analyse von vorgängen im spurenfeld. i. von w. köhler und h. v. restorff. *Psychologische Forschung*, 18:299 – 342, 1933.

[19] John P Rickards and Peter R Denner. Depressive effects of underlining and adjunct questions on children's recall of text. *Instructional Science*, 8(1):81–90, 1979.

[20] Adam Roegiest and Gordon V Cormack. Impact of review-set selection on human assessment for text classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 861–864, 2016.

[21] Burr Settles. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011.

[22] Vicki L Silvers and David S Kreiner. The effects of pre-existing inappropriate highlighting on reading comprehension. *Literacy Research and Instruction*, 36(3):217–223, 1997.

[23] Kalmer E Stordahl and Clifford M Christensen. The effect of study techniques on comprehension and retention. *The Journal of Educational Research*, 49(8):561–570, 1956.

[24] John Tredennick. E-discovery, my how you've grown!', 2011.

[25] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA, 2021.

[26] Ellen M Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, 2001.

[27] Suzanne E Wade and Woodrow Trathen. Effect of self-selected study methods on learning. *Journal of educational psychology*, 81(1):40, 1989.

[28] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.

[29] Xue Jun Wang, Maura R Grossman, and Seung Gyu Hyun. Participation in trec 2020 covid track using continuous active learning. *arXiv preprint arXiv:2011.01453*, 2020.

[30] Carole L Yue, Benjamin C Storm, Nate Kornell, and Elizabeth Ligon Bjork. Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review*, 27(1):69–78, 2015.

[31] Haotian Zhang. Increasing the efficiency of high-recall information retrieval. 2019.

[32] Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. Effective user interaction for high-recall retrieval:

Less is more. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 187–196, 2018.

[33] Xi Zheng, Akanksha Bansal, and Matthew Lease. Bullseye: structured passage retrieval and document highlighting for scholarly search. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–4, 2017.

APPENDICES

Appendix A

Collected Data

A.1 CAL-C Metrics

Topic Number	**Recall**	**Precision**	**F1**
1	0.0386	0.8709	0.0738
1	0.0429	0.6666	0.0804
1	0.0185	0.65	0.0359
1	0.0815	0.6	0.1435
2	0.2119	0.9466	0.3462
2	0.1253	0.9545	0.2214
2	0.0328	1	0.0635
2	0.0567	0.95	0.1069
3	0.0644	0.5121	0.1143
3	0	0	0
3	0.0153	0.4761	0.0295
3	0.0153	0.5882	0.0296
3	0.0061	0.1025	0.011
4	0.0246	0.1346	0.0414
4	0.007	0.0655	0.0124
4	0.0105	0.1818	0.0197
4	0.007	0.2105	0.0133
5	0.017	0.1428	0.03
5	0	0	0
5	0.0154	0.2564	0.0286

Table A.1 continued from previous page

5	0.0201	0.4333	0.0383
5	0.003	0.0363	0.005
6	0.009	0.36	0.0173
6	0.002	0.019	0
6	0.01	0.4	0.0195
7	0.0171	0.6428	0.0331
7	0.0209	0.9166	0.0408
7	0.0019	0.0138	0
7	0.0286	0.5769	0.0543
8	0.0185	0.1538	0.0325
8	0.0185	0.4137	0.0354
8	0.0154	1	0.0303
8	0.0216	0.7777	0.0419
9	0.0095	0.0322	0.0143
9	0	0	0
9	0.1004	0.2079	0.1352
9	0.0239	0.2083	0.0426
10	0.0261	0.5909	0.0499
10	0.0704	0.5303	0.1241
10	0.016	0.2758	0.0301
10	0.1307	0.7471	0.2223
11	0.0452	0.2941	0.0781
11	0.0339	0.5	0.0634
11	0.018	0.5714	0.0347
11	0.0226	0.3448	0.0421
12	0.0169	0.4583	0.0324
12	0.0077	0.25	0.0147
12	0.0447	0.4603	0.0813
12	0.0138	0.3913	0.0264
13	0.0402	0.8222	0.0766
13	0.0391	0.4337	0.0717
13	0.0065	0.0689	0.0106
13	0.0619	0.3931	0.1068
14	0.0549	0.9375	0.1036
14	0.1648	0.75	0.2702
14	0.0439	0.8571	0.0834
15	0.0896	0.6153	0.1563

Table A.1 continued from previous page

15	0.0538	0.7272	0.1001
15	0.0156	0.5833	0.0302
15	0.0089	0.129	0.0159
16	0.0268	0.9166	0.052
16	0.0536	0.9166	0.1012
16	0.078	0.7272	0.1408
16	0.0243	1	0.0474
17	0.0209	0.1724	0.0372
17	0.0153	0.1447	0.0275
17	0.0083	0.0458	0.0129
17	0.0069	0.2941	0.0132
18	0.0795	0.5096	0.1374
18	0.0585	0.6842	0.1077
18	0.0345	0.7187	0.0657
18	0.0465	0.6078	0.0863
19	0.0683	0.5	0.1201
19	0.1452	0.4722	0.222
19	0.2307	0.4821	0.312
19	0.1196	0.56	0.197
20	0.0317	0.7272	0.0607
20	0.0924	0.8045	0.1656
20	0.0383	0.725	0.0727
20	0.1109	0.6942	0.1911

Table A.1: CAL-C Metrics

A.2 CAL-H Metrics

Topic Number	Recall	Precision	F1
1	0.01	0.0503	0.0165
1	0.0171	0.8571	0.0335
1	0.0257	0.6428	0.0493
1	0.03	0.8076	0.0577
2	0.0507	1	0.0965
2	0.0477	0.8888	0.0904
2	0.0268	1	0.0522

Table A.2 continued from previous page

2	0.0597	1	0.1126
3	0.0061	0.25	0.0117
3	0.023	0.4166	0.0434
3	0.0046	0.1875	0.0088
3	0.0138	0.5294	0.0268
4	0.007	0.1481	0.0128
4	0.0035	0.0625	0.006
4	0.0017	0.0384	0.0024
4	0.0017	0.0149	0
5	0.003	0.0833	0.0046
5	0.0263	0.3207	0.0484
5	0.0216	0.4117	0.0408
5	0.0061	0.1538	0.0112
6	0.0241	0.5	0.0459
6	0.001	0.0149	0
6	0.016	0.7619	0.0312
6	0.0261	0.5306	0.0495
7	0.0438	0.5897	0.0814
7	0.0591	0.7948	0.1099
7	0.0038	0.0235	0.0036
7	0.0438	0.8214	0.0831
8	0.0108	0.875	0.0213
8	0.0277	0.5454	0.0526
8	0.0154	0.8333	0.0301
8	0.0108	0.875	0.0213
9	0.0143	0.0652	0.0226
9	0.1339	0.8235	0.2303
9	0.1339	0.3835	0.1984
9	0.0669	0.4375	0.1159
10	0.1448	0.4235	0.2157
10	0.008	0.2	0.0153
10	0.0301	0.4054	0.056
10	0.0362	0.6923	0.0687
11	0.0226	0.4166	0.0428
11	0.0248	0.3437	0.0461
11	0.0203	0.6428	0.0392
11	0.0429	0.3725	0.0767

Table A.2 continued from previous page

12	0.037	0.4363	0.068
12	0.0154	0.2631	0.029
12	0.0756	0.4375	0.1288
12	0.0231	0.4545	0.0437
13	0.0597	0.4782	0.1059
13	0.0163	0.5769	0.0316
13	0.0293	0.6136	0.0558
13	0.0195	0.6	0.0377
14	0.1318	0.8372	0.2276
14	0.1208	0.9166	0.2134
14	0.0915	0.9259	0.1665
14	0.1318	0.9473	0.2313
15	0.0403	0.8571	0.0768
15	0.0022	0.0625	0.003
15	0.0112	0.5555	0.0218
15	0.0493	0.4888	0.0893
16	0.078	0.7619	0.1414
16	0.0682	0.9032	0.1267
16	0	0	0
16	0.0634	0.8666	0.118
17	0.0599	0.4942	0.1068
17	0.0027	0.0298	0.003
17	0.0223	0.2807	0.0412
17	0.0209	0.3	0.0389
18	0.018	0.48	0.0345
18	0.0585	0.709	0.108
18	0.039	0.8125	0.0743
18	0.0255	0.85	0.0494
19	0.1965	0.5227	0.2855
19	0.1709	0.5405	0.2596
19	0.0769	0.6	0.1362
19	0.1623	0.5	0.245
20	0.0964	0.4424	0.1581
20	0.0237	0.6923	0.0458
20	0.033	0.7142	0.063
20	0.0356	0.6923	0.0675

Table A.2 continued from previous page

Table A.2: CAL-H Metrics

Appendix B

Query Topics

Topic Number		
1	query	coronavirus origin
	question	what is the origin of COVID-19?
	narrative	seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans
2	query	oronavirus response to weather changes
	question	how does the coronavirus respond to changes in the weather?
	narrative	seeking range of information about the SARS-CoV-2 virus viability in different weather/climate conditions as well as information related to transmission of the virus in different climate conditions
3	query	coronavirus immunity
	question	will SARS-CoV2 infected people develop immunity? Is cross protection possible?
	narrative	seeking studies of immunity developed due to infection with SARS-CoV2 or cross protection gained due to infection with other coronavirus types
4	query	how do people die from the coronavirus
	question	what causes death from Covid-19?
	narrative	studies looking at mechanisms of death from Covid-19
5	query	animal models of COVID-19
	question	what drugs have been active against SARS-CoV or SARS-CoV-2 in animal studies?
	narrative	papers that describe the results of testing drugs that bind to spike proteins of the virus or any other drugs in any animal models. Papers about SARS-CoV-2 infection in cell culture assays are also relevant
6	query	coronavirus test rapid testing
	question	what types of rapid testing for Covid-19 have been developed?
	narrative	looking for studies identifying ways to diagnose Covid-19 more rapidly
7	query	serological tests for coronavirus
	question	are there serological tests that detect antibodies to coronavirus?
	narrative	looking for assays that measure immune response to COVID-19 that will help determine past infection and subsequent possible immunity

Table B.1 continued from previous page

8	query	coronavirus under reporting
	question	how has lack of testing availability led to under reporting of true incidence of Covid-19?
	narrative	looking for studies answering questions of impact of lack of complete testing for Covid-19 on incidence and prevalence of Covid-19
9	query	coronavirus in Canada
	question	how has COVID-19 affected Canada?
	narrative	seeking data related to infections (confirm, suspected, and projected) and health outcomes (symptoms, hospitalization, intensive care, mortality)
10	query	coronavirus social distancing impact
	question	has social distancing had an impact on slowing the spread of COVID-19?
	narrative	seeking specific information on studies that have measured COVID-19's transmission in one or more social distancing (or non-social distancing) approaches
11	query	coronavirus hospital rationing
	question	what are the guidelines for triaging patients infected with coronavirus?
	narrative	seeking information on any guidelines for prioritizing COVID-19 patients infected with coronavirus based on demographics, clinical signs, serology and other tests
12	query	coronavirus quarantine
	question	what are best practices in hospitals and at home in maintaining quarantine?
	narrative	seeking information on best practices for activities and duration of quarantine for those exposed and/ infected to COVID-19 virus
13	query	how does coronavirus spread
	question	what are the transmission routes of coronavirus?
	narrative	looking for information on all possible ways to contract COVID-19 from people, animals and objects
14	query	coronavirus super spreaders
	question	what evidence is there related to COVID-19 super spreaders?

Table B.1 continued from previous page

	narrative	seeking range of information related to the number and proportion of super spreaders, their patterns of behavior that lead to spread, and potential prevention strategies targeted specifically toward super spreaders
15	query	coronavirus outside body
	question	how long can the coronavirus live outside the body?
	narrative	seeking range of information on the SARS-CoV-2's virus's survival in different environments (surfaces, liquids, etc.) outside the human body while still being viable for transmission to another human
16	query	how long does coronavirus survive on surfaces
	question	how long does coronavirus remain stable on surfaces?
	narrative	studies of time SARS-CoV-2 remains stable after being deposited from an infected person on everyday surfaces in a household or hospital setting, such as through coughing or touching objects
17	query	coronavirus clinical trials
	question	are there any clinical trials available for the coronavirus?
	narrative	seeking specific COVID-19 clinical trials ranging from trials in recruitment to completed trials with results
18	query	masks prevent coronavirus
	question	what are the best masks for preventing infection by Covid-19?
	narrative	what types of masks should or should not be used to prevent infection by Covid-19?
19	query	what alcohol sanitizer kills coronavirus
	question	what type of hand sanitizer is needed to destroy Covid-19?
	narrative	studies assessing chemicals and their concentrations needed to destroy the Covid-19 virus
20	query	coronavirus and ACE inhibitors
	question	are patients taking Angiotensin-converting enzyme inhibitors (ACE) at increased risk for COVID-19?

Table B.1 continued from previous page

narrative	looking for information on interactions between coronavirus and angiotensin converting enzyme 2 (ACE2) receptors, risk for patients taking these medications, and recommendations for these patients

Table B.1: Search Topics

Appendix C

Post-task Questions

Topic 1:	**coronavirus origin**
Q1.	*What sort of origin is COVID-19 suggested to have?*
a)	Silicotic
b)	Zoonotic
c)	Zymotic
d)	Posthypnotic
e)	Siderotic
Q2.	*A number of arguments suggest that COVID-19 has a zoonotic origin. What species is suspected of being the "missing link" in its transmission to humans?*
a)	Rabbits
b)	Armadillos
c)	Pangolins
d)	Mosquitos
e)	Pigs
Q3.	*How does the H5N1 virus, also known as the "avian influenza", differ from COVID-19?*
a)	Avian and mammals are different
b)	Humans were never affected
c)	Coughing is not a symptom
d)	All of the above
e)	Both a and b
Topic 2:	**coronavirus response to weather changes**

Table C.1 continued from previous page

Q1.	*Different weather and climate is most likely to affect the ________ of COVID-19.*
a)	Symptoms
b)	Treatment
c)	Operation
d)	Transmission
e)	Placebo
Q2.	*The transmissibility of COVID-19 seems to be strongly correlated with certain features of weather and climate. Especially with ________.*
a)	Temperature
b)	Humidity
c)	Precipitation
d)	Wind speed
e)	Both a and b
Q3.	*The 2 metre of social distance recommended by the Center for Disease Control and Prevention (CDC) may be insufficient in certain environmental conditions. Especially in ________.*
a)	High precipitation
b)	High wind speed
c)	High temperature and high humidity
d)	High temperature and low humidity
e)	Low temperature and low humidity
Topic 3:	**coronavirus immunity**
Q1.	*Which of the following(s) is/are part of the seven coronaviruses associated with diseases in humans other than SARS-CoV-2, commonly known as COVID-19?*
a)	NIRS-Cov
b)	HARS-Cov
c)	MERS-CoV
d)	URA-Cov
e)	GBA-Cov
Q2.	*Receiving ________ from other coronavirus outbreaks is one possible explanation for some populations experiencing milder symptoms than others.*
a)	ADE (adverse drug events)
b)	ADE (antibody dependent enhancement)
c)	ADE (absorption-desorption equilibrium)

Table C.1 continued from previous page

d)	ADE (after death experiences)
e)	None of the above
Q3.	*Antibody-dependent enhancement (ADE) is an atypical immunological paradox commonly associated with dengue virus re-infection. Which of the following have been hypothesized to be a result(s) of receiving ADE from other coronaviruses?*
a)	Severe clinical manifestations of SARS-CoV-2
b)	Milder symptoms correlated to cross protection
c)	Guaranteed immunity to Covid-19
d)	Both a and b
e)	None of the above
Topic 4:	**how do people die from the coronavirus**
Q1.	*COVID-19 patients with severe ________ is often found to have fatal outcomes.*
a)	Acute Kidney Injury (AKI)
b)	Acute Schististmus (AS)
c)	Acute Glycacal Injury (AGI)
d)	Acute Oligectasia (AO)
e)	Acute Pachyad Injury (API)
Q2.	*Which of the following features are correlated with higher mortality from Covid-19?*
a)	Being male
b)	Older age
c)	Having comorbidities
d)	All of the above
e)	Both b and c
Q3.	*Patients with cancer are more likely to ________*
a)	be asymptomatic when infected with Covid-19
b)	die from Covid-19
c)	recover from Covid-19
d)	Both a and b
e)	Both b and c
Topic 5:	**animal models of COVID-19**
Q1.	*A promising target for both biagnosis and therapeutics treatments of the new COVID-19 is the corona virus (CoV) spike (S) ________.*
a)	Fibronectin

Table C.1 continued from previous page

b)	Pikachurin
c)	Glycoprotein
d)	Rhodopsin
e)	Hemoglobin
Q2.	*________ in SARS-CoV-2 S protein have been found to bind strongly to human and bat ________ receptors.*
a)	receptor binding domain (RBD), angiotensin-converting enzyme 2 (ACE2)
b)	receptor binding domain (RBD), molecular dynamics (MD)
c)	molecular dynamics (MD), angiotensin-converting enzyme 2 (ACE2)
d)	angiotensin-converting enzyme 2 (ACE2), receptor binding domain (RBD)
e)	None of the above
Q3.	*SARS-CoV-2 and SARS-CoV are known to both recognize the same host cell receptor responsible for mediating infection. This means that SAR-CoV, the SARS pandemic that occurred in 2002, also ________.*
a)	has loss of olfactory senses as a symptom
b)	has the same level of severity once infected
c)	binds to angiotensin-converting enzyme 2 (ACE2)
d)	All of the above
e)	Both b and c
Topic 6:	**coronavirus test rapid testing**
Q1.	*Which of the following is a type of rapid testing for COVID-19?*
a)	Singlewave rapid detection tests
b)	Mayocarditis-linked rapid detection tests
c)	Antigen-based rapid detection tests
d)	Brachyium rapid detection tests
e)	Glycole-linked rapid detection tests
Q2.	*Using samples of nasopharyngeal swabs allows us to perform which of the following tests?*
a)	Molecular tests
b)	Antigen tests
c)	Antibody tests
d)	Both a and b
e)	Both b and c

Table C.1 continued from previous page

Q3.	*Which tests are helpful in identifying people who are capable of donating plasma to help patients currently fighting the infection.*
a)	Molecular tests
b)	Antibody tests
c)	Antigen tests
d)	Both a and b
e)	Both b and c
Topic 7:	**serological tests for coronavirus**
Q1.	*Serological methods can be used to detect specific antibodies of ________ classes.*
a)	HpA and HpC
b)	IgM and IgG
c)	TsG and TsB
d)	ArC and ArN
e)	CvK and CvG
Q2.	*A positive serological test result indicates ________.*
a)	Immunity to reinfection with the suspected pathogen
b)	Active infection with the suspected pathogen
c)	Recent exposure to the suspected pathogen
d)	Both a and c
e)	None of the above
Q3.	*Why may serological tests have limited use as a diagnostic method for active COVID-19 infections?*
a)	It detects antibodies
b)	Results can take up to a week
c)	It requires blood
d)	Both a and b
e)	None of the above
Topic 8:	**coronavirus under reporting**
Q1.	*A possible solution for sparsity of testing kits is to use ________.*
a)	Molecular testing
b)	Predictive testing
c)	Genetic testing
d)	Protoid testing
e)	Group testing

Table C.1 continued from previous page

Q2.	*Antibody tests are _______ to the under reporting of COVID-19 case since _______.*
a)	Helpful, it accounts for recovered patients
b)	Not helpful, it only accounts for recovered patients
c)	Helpful, it relies on blood samples
d)	Not helpful, it relies on blood samples
e)	Both b and d
Q3.	*To retrospectively help catch under reported cases of COVID-19, we should collect samples of _______ to test.*
a)	Blood
b)	Nasopharyngeal swab
c)	Oropharyngeal swab
d)	Both a and b
e)	Both b and c
Topic 9:	**coronavirus in Canada**
Q1.	*Which of the following are journals reporting on COVID-19 in Canada?*
a)	Canadian Health journal (CHJ)
b)	Canadian Health Care journal (CHCJ)
c)	Canadian Open journal (COJ)
d)	Canadian Medicine journal (CMJ)
e)	Canadian Medical Association journal (CMAJ)
Q2.	*Which of the following are computational/mathematical models used to achieve?*
a)	Projecting demand for critical care beds during COVID-19
b)	Estimation of CVOID-19-induced depletion of hospital resources
c)	Effects of physical-distancing interventions used to slow the spread of COVID-19
d)	All of the above
e)	Both a and b
Q3.	*Which of the following are non-pharmaceutical interventions Canadians can apply to reduce the burden on the healthcare system during the COVID-19 pandemic?*
a)	Vaccination
b)	Social distancing
c)	Remdesivir
d)	Opting for in person meetings

Table C.1 continued from previous page

e)	Both b and d
Topic 10:	**coronavirus social distancing impact**
Q1.	*Rigorous social distancing has led to dramatic declines in ________.*
a)	Confidentiality
b)	Mobility
c)	Bioavailability
d)	Monospecific
e)	None of the above
Q2.	*Social distancing commonly referred to as a ________ intervention on the transmission of COVID-19.*
a)	Non-pharmaceutical
b)	Parasitological
c)	Non-countercyclical
d)	Transhistorical
e)	Non-Neuroanatomical
Q3.	*Examples of social distancing includes ________.*
a)	Wearing a mask
b)	Working from home
c)	Washing your hands
d)	All of the above
e)	Both a and b
Topic 11:	**coronavirus hospital rationing**
Q1.	*Which of the following methods have been used for severity risk prediction and triage of COVID-19 patients?*
a)	Lacrimitis testing
b)	Ectacal ranking
c)	Unation rapid testing
d)	Machine learning
e)	Ostacal ranking
Q2.	*Which of the following features is/are the best predictor(s) of COVID-19?*
a)	Cough
b)	Headache
c)	Fever
d)	Sore throat
e)	Myalgia

Table C.1 continued from previous page

Q3.	*Which of the following are likely procedures used in triage of COVID-19 patients?*
a)	Checking for fever
b)	Collecting history of travel
c)	Collecting a blood sample
d)	All of the above
e)	Both a and b
Topic 12:	**coronavirus quarantine**
Q1.	*Quarantine is used to control ________ exposure among people.*
a)	Consubstantial
b)	Inquisitorial
c)	Social
d)	Salpingad
e)	Uncial
Q2.	*Quarantine is commonly referred to as a ________ intervention on the transmission of COVID-19.*
a)	Non-pharmaceutical
b)	Parasitological
c)	Non-countercyclical
d)	Transhistorical
e)	Non-Neuroanatomical
Q3.	*Why is quarantining for 14 days alone not enough? (i.e. we still need to wear masks, maintain physical distancing, get tested, etc.)*
a)	People can be asymptomatic
b)	Symptoms can develop after 14 days
c)	SEIR modeling of COVID-19 does not consider quarantine
d)	All of the above
e)	Both a and b
Topic 13:	**how does coronavirus spread**
Q1.	*Which of the following(s) is/are potential transmission routes?*
a)	Facies-nasal transmission
b)	Pyrostaxis transmission
c)	Morphic-prosoposis transmission
d)	Faecal-oral transmission
e)	Dynamectomy transmission

Table C.1 continued from previous page

Q2.	*Which of the following are confirmed methods of transmission of COVID-19?*
a)	Direct contact transmission
b)	Airborne water droplets (aerosol) transmission
c)	Faecal-oral (including waterborne) transmission
d)	Both a and b
e)	Both b and c
Q3.	*Patients can have positive faecal tests even after having negative nasopharyngeal swabs. This means that ________.*
a)	all patients can expect to have gastrointestinal symptoms in early stages of their infection
b)	currently, there are more confirmed cases of waterborne transmission than aerosol transmission
c)	wastewater surveillance of SARS-CoV-2 can be an effective tool in early detection of outbreak and determination of COVID-19 prevalence within a population
d)	All of the above
e)	Both a and c
Topic 14:	**coronavirus super spreaders**
Q1.	*Reconstruction of infection events, including that of super-spreading events, falls into the study of ________.*
a)	epidemiology
b)	semeiology
c)	soteriology
d)	apiology
e)	trachology
Q2.	*Which of the following are most commonly used to infer the presence or absence of super-spreading events during the early phases of these outbreaks?*
a)	Reproductive number R(0)
b)	Dispersion factor (k)
c)	Probability of death (P)
d)	Both a and b
e)	Both b and c
Q3.	*If MERS had an estimated k of 0.2 and SARS-CoV-2 has an estimated k of 0.08, we can deduce that super spreaders ________.*

Table C.1 continued from previous page

a)	are more prominent in SARS-CoV-2 than MERS
b)	are more prominent in MERS than SARS-CoV-2
c)	played a key role in the early stages of both events
d)	Both a and c
e)	Both b and c
Topic 15:	**coronavirus outside body**
Q1.	*Common place(s) for Covid-19 virus survival outside of the body is/are in ________.*
a)	saliva droplets
b)	hair shedding
c)	skin shedding
d)	Both a and b
e)	Both a and c
Q2.	*What are some environmental factors which affect how long SARS-CoV-2 can survive on common touch surfaces outside the body?*
a)	Temperature
b)	Humidity
c)	Surface texture
d)	All of the above
e)	Both a and b
Q3.	*Reason(s) why SARS-CoV-2 and other animal CoVs have remarkably short persistence on copper and latex surfaces compared to stainless steel, plastic, and glass is/are due to their difference in ________.*
a)	thermal conductivity
b)	density
c)	porosity
d)	All of the above
e)	Both a and c
Topic 16:	**how long does coronavirus survive on surfaces**
Q1.	*Common place(s) for Covid-19 virus survival outside of the body is/are in ________.*
a)	saliva droplets
b)	hair shedding
c)	skin shedding
d)	Both a and b
e)	Both a and c

Table C.1 continued from previous page

Q2.	*What are some environmental factors which affect how long SARS-CoV-2 can survive on common touch surfaces outside the body?*
a)	Temperature
b)	Humidity
c)	Surface texture
d)	All of the above
e)	Both a and b
Q3.	*Reason(s) why SARS-CoV-2 and other animal CoVs have remarkably short persistence on copper and latex surfaces compared to stainless steel, plastic, and glass is/are due to their difference in ________.*
a)	porosity
b)	thermal conductivity
c)	density
d)	All of the above
e)	Both a and b
Topic 17:	**coronavirus clinical trials**
Q1.	*A common site(s) used for clinical trial registration is ________.*
a)	covidtrials.gov
b)	clinicaltrials.gov
c)	cdc.gov/clinical-trials
d)	biomedcentral.com/clinical-trials
e)	medicaltrials.com
Q2.	*Which of the following are drug repurposing clinical trials?*
a)	Hydroxychloroquine
b)	Radiotherapy
c)	Remdesivir
d)	Plasma
e)	Both a and c
Q3.	*Which of the following are reasons which makes Chloroquine a good candidate for drug repurposing clinical trials against COVID-19.*
a)	Like hydroxychloroquine, chloroquine also affect endosomal function
b)	Unlike hydroxychloroquine, chloroquine does not block autophagosome-lysosome fusion
c)	Unlike hydroxychloroquine, chloroquine is a broad-spectrum antibiotic
d)	Both a and b
e)	None of the above

Table C.1 continued from previous page

Topic 18:	**masks prevent coronavirus**
Q1.	*Which of the following is a possible risk of prolonged mask wearing?*
a)	Ciliphilia
b)	Cytopexy
c)	Orostenosis
d)	Skin damage
e)	Carbon dioxide poisoning
Q2.	*Which of the following are helpful in slowing the spread of COVID-19 in normal everyday settings?*
a)	Surgical grade N95 masks
b)	Surgical masks other than N95 masks
c)	Cloth masks
d)	All of the above
e)	Both a and b
Q3.	*Are surgical grade N95 masks capable of completely filtering all SARS-CoV-2 virions?*
a)	Yes, its filtration efficiency for sub-300nm particles is 100%
b)	Yes, virions smaller than 300nm particles are not infectious for humans
c)	Yes, if worn properly
d)	No, its filtration efficiency for sub-300nm particles is not 100%
e)	No, since SARS-CoV-2 virions are oil-based
Topic 19:	**what alcohol sanitizer kills coronavirus**
Q1.	*Which of the following are ingredients frequently used in alcohol hand sanitizers?*
a)	Methanol
b)	Hydrogen monoxide
c)	Sodium oxide
d)	Propanol
e)	Ethanol
Q2.	*Which of the following is the recommended alcohol concentration in alcohol hand sanitizers to be effective against COVID-19?*
a)	30% or more
b)	40% or more
c)	50% or more
d)	60% or more
e)	None of the above

Table C.1 continued from previous page

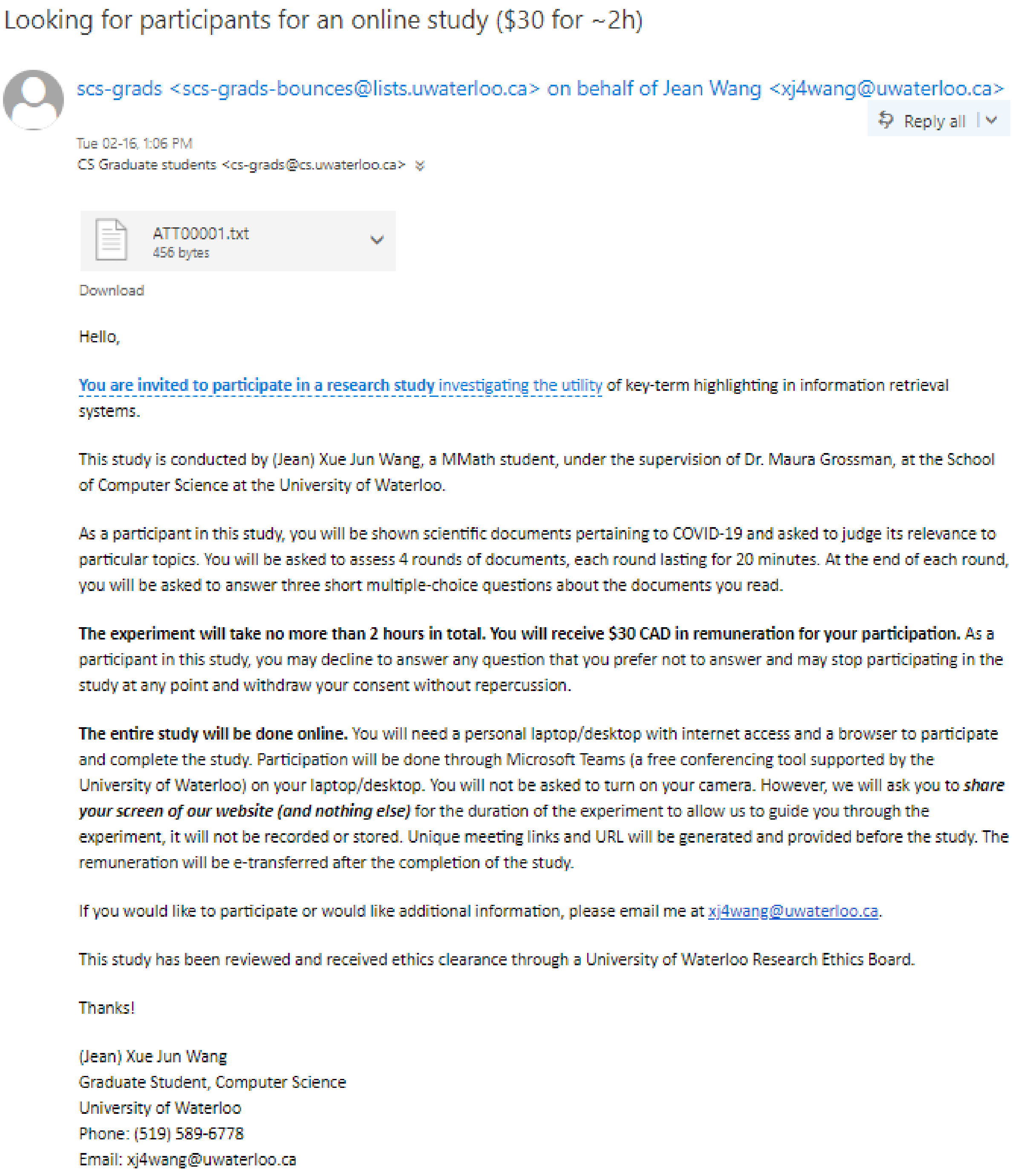
Q3.	*Hand sanitizer ________ many microbes, making it ________ effective than washing your hands with soap and water at reducing the transmission of COVID-19.*
a)	eliminates, more
b)	removes, less
c)	removes, more
d)	inactivates, less
e)	inactivates, more
Topic 20:	**coronavirus and ACE inhibitors**
Q1.	*Which of the following is angiotensin converting enzyme 2 (ACE2) a regulator of?*
a)	Renin-angiotensin system
b)	Pyad-papilliarty system
c)	Corics-oligoid system
d)	Enteresophageal system
e)	Myringiatry-variceal system
Q2.	*Which of the following are treated using angiotensin-converting enzyme inhibitors?*
a)	Hypertension
b)	Diabetes
c)	SARS-CoV-2
d)	Both a and b
e)	Both a and c
Q3.	*Why are patients taking Angiotensin-converting enzyme inhibitors (ACE) potentially at an increased risk for COVID-19?*
a)	ACE2 enzyme is the SARS-CoV-2 receptor
b)	It increases oxygen demand from the heart
c)	It increases the production of angiotensin II
d)	Both a and b
e)	Both a and c

Table C.1: Post-task questions

Appendix D

Ethics

D.1 Recruitment E-mail



Looking for participants for an online study ($30 for ~2h)

scs-grads <scs-grads-bounces@lists.uwaterloo.ca> on behalf of Jean Wang <xj4wang@uwaterloo.ca>

Tue 02-16, 1:06 PM

CS Graduate students <cs-grads@cs.uwaterloo.ca>

Hello,

You are invited to participate in a research study investigating the utility of key-term highlighting in information retrieval systems.

This study is conducted by (Jean) Xue Jun Wang, a MMath student, under the supervision of Dr. Maura Grossman, at the School of Computer Science at the University of Waterloo.

As a participant in this study, you will be shown scientific documents pertaining to COVID-19 and asked to judge its relevance to particular topics. You will be asked to assess 4 rounds of documents, each round lasting for 20 minutes. At the end of each round, you will be asked to answer three short multiple-choice questions about the documents you read.

The experiment will take no more than 2 hours in total. You will receive $30 CAD in remuneration for your participation. As a participant in this study, you may decline to answer any question that you prefer not to answer and may stop participating in the study at any point and withdraw your consent without repercussion.

The entire study will be done online. You will need a personal laptop/desktop with internet access and a browser to participate and complete the study. Participation will be done through Microsoft Teams (a free conferencing tool supported by the University of Waterloo) on your laptop/desktop. You will not be asked to turn on your camera. However, we will ask you to ***share your screen of our website (and nothing else)*** for the duration of the experiment to allow us to guide you through the experiment, it will not be recorded or stored. Unique meeting links and URL will be generated and provided before the study. The remuneration will be e-transferred after the completion of the study.

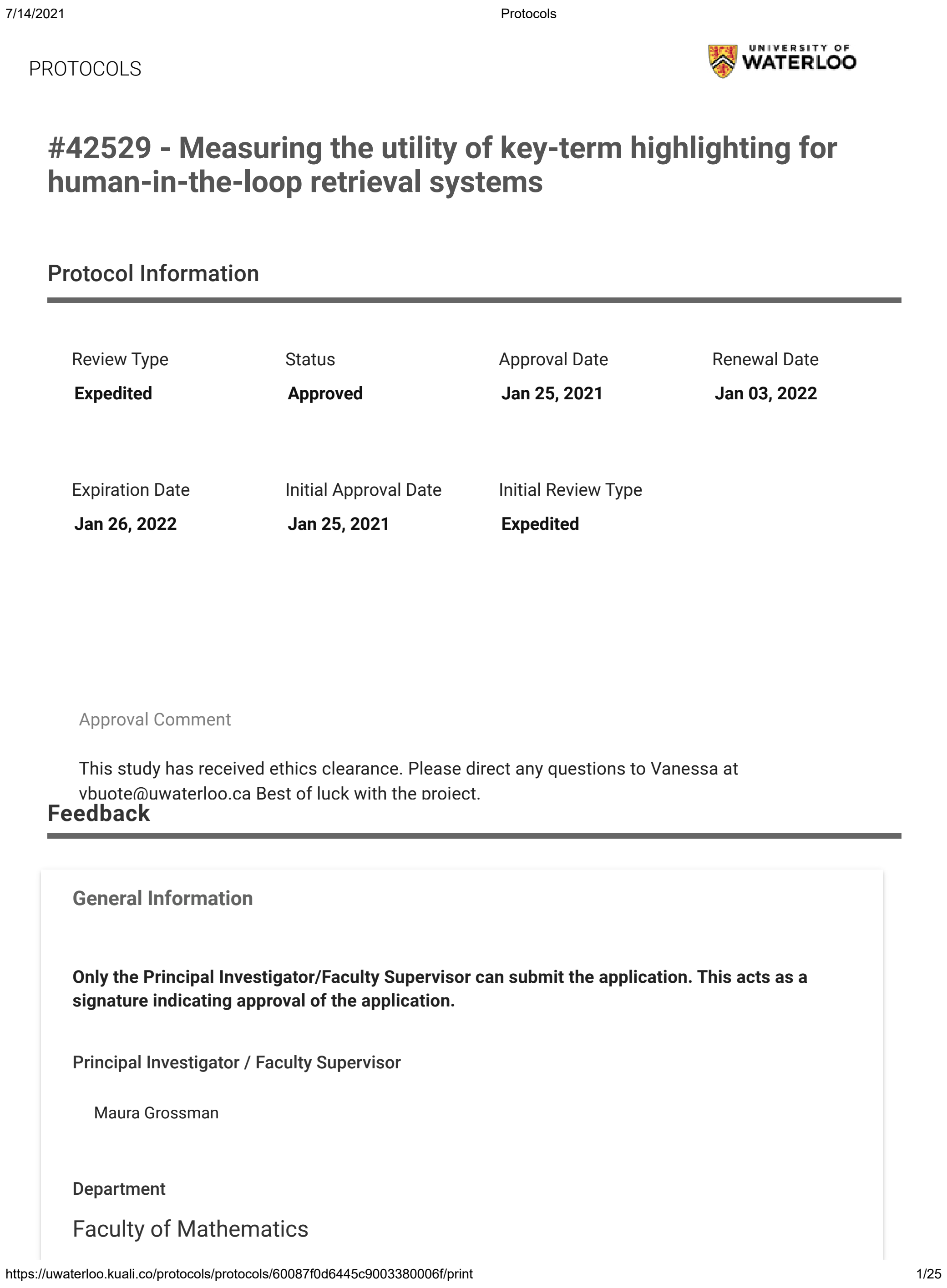
If you would like to participate or would like additional information, please email me at xj4wang@uwaterloo.ca.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board.

Thanks!

(Jean) Xue Jun Wang
Graduate Student, Computer Science
University of Waterloo
Phone: (519) 589-6778
Email: xj4wang@uwaterloo.ca

D.2 Ethics Form



PROTOCOLS

UNIVERSITY OF WATERLOO

#42529 - Measuring the utility of key-term highlighting for human-in-the-loop retrieval systems

Protocol Information

Review Type	Status	Approval Date	Renewal Date
Expedited	**Approved**	**Jan 25, 2021**	**Jan 03, 2022**

Expiration Date	Initial Approval Date	Initial Review Type
Jan 26, 2022	**Jan 25, 2021**	**Expedited**

Approval Comment

This study has received ethics clearance. Please direct any questions to Vanessa at vbuote@uwaterloo.ca Best of luck with the project.

Feedback

General Information

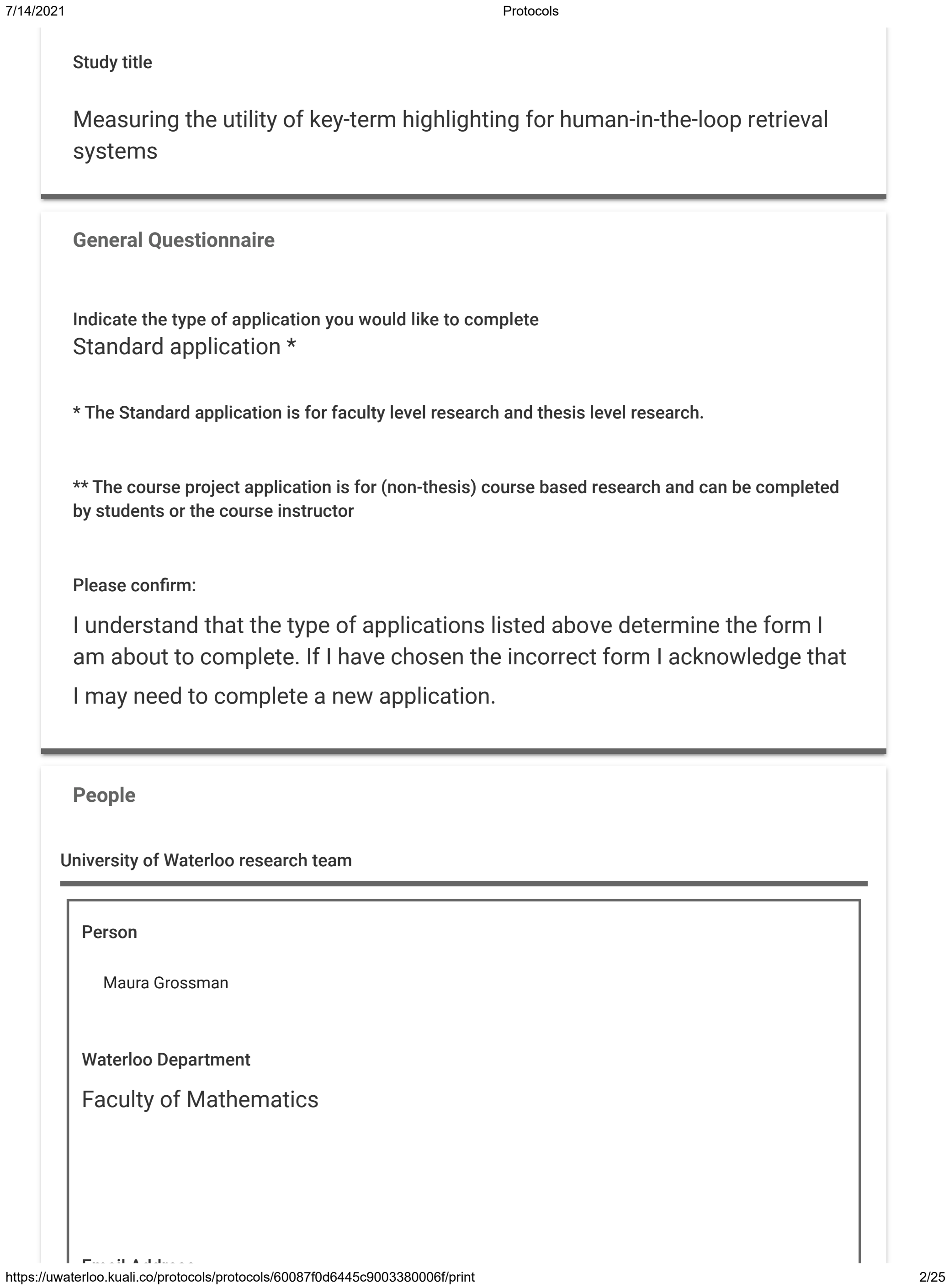
Only the Principal Investigator/Faculty Supervisor can submit the application. This acts as a signature indicating approval of the application.

Principal Investigator / Faculty Supervisor

Maura Grossman

Department

Faculty of Mathematics



7/14/2021 Protocols

Study title

Measuring the utility of key-term highlighting for human-in-the-loop retrieval systems

General Questionnaire

Indicate the type of application you would like to complete
Standard application *

*** The Standard application is for faculty level research and thesis level research.**

**** The course project application is for (non-thesis) course based research and can be completed by students or the course instructor**

Please confirm:

I understand that the type of applications listed above determine the form I am about to complete. If I have chosen the incorrect form I acknowledge that I may need to complete a new application.

People

University of Waterloo research team

Person

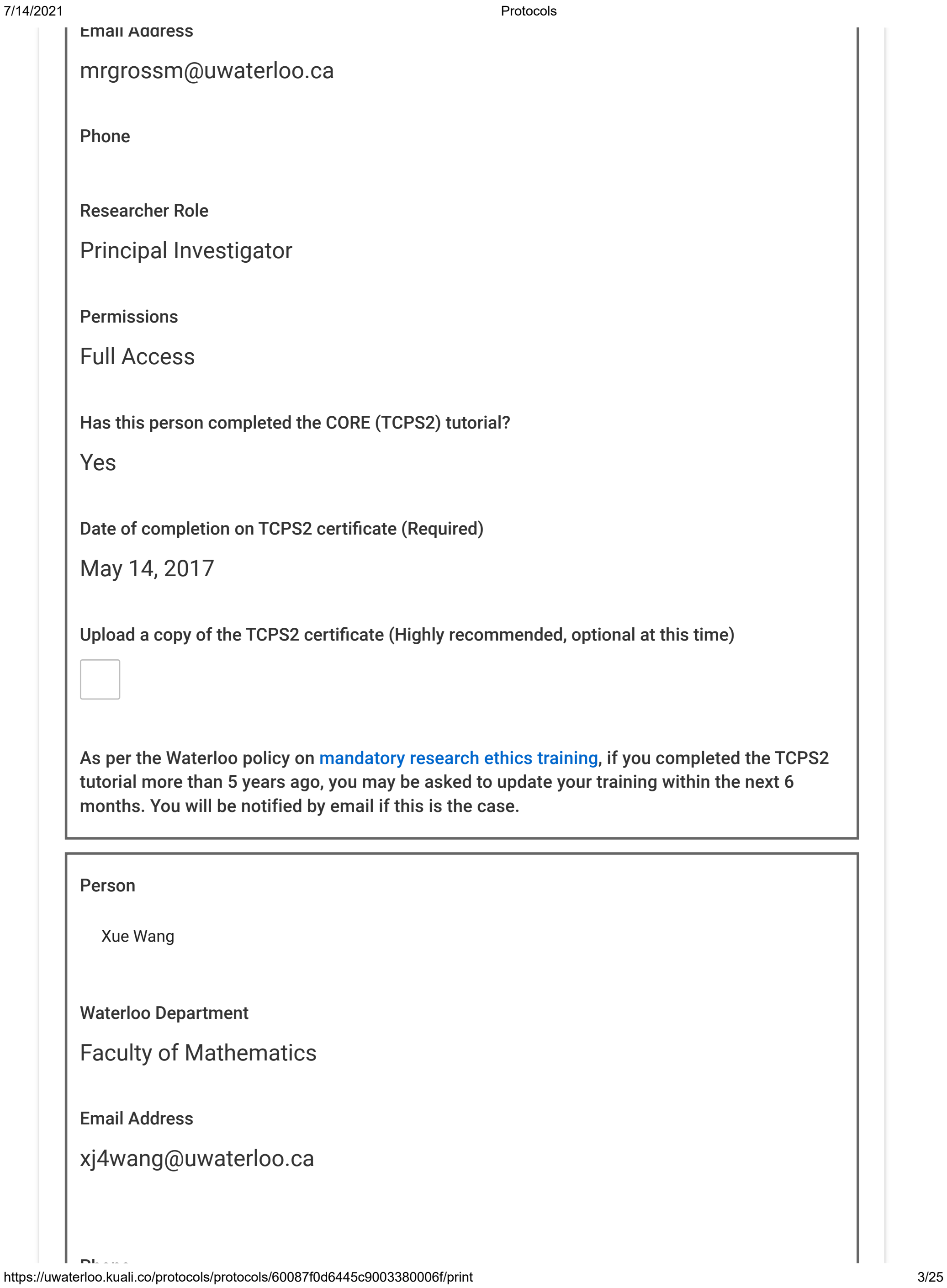
Maura Grossman

Waterloo Department

Faculty of Mathematics

Email Address

https://uwaterloo.kuali.co/protocols/protocols/60087f0d6445c9003380006f/print 2/25



7/14/2021 Protocols

Email Address

mrgrossm@uwaterloo.ca

Phone

Researcher Role

Principal Investigator

Permissions

Full Access

Has this person completed the CORE (TCPS2) tutorial?

Yes

Date of completion on TCPS2 certificate (Required)

May 14, 2017

Upload a copy of the TCPS2 certificate (Highly recommended, optional at this time)

□

As per the Waterloo policy on mandatory research ethics training, if you completed the TCPS2 tutorial more than 5 years ago, you may be asked to update your training within the next 6 months. You will be notified by email if this is the case.

Person

Xue Wang

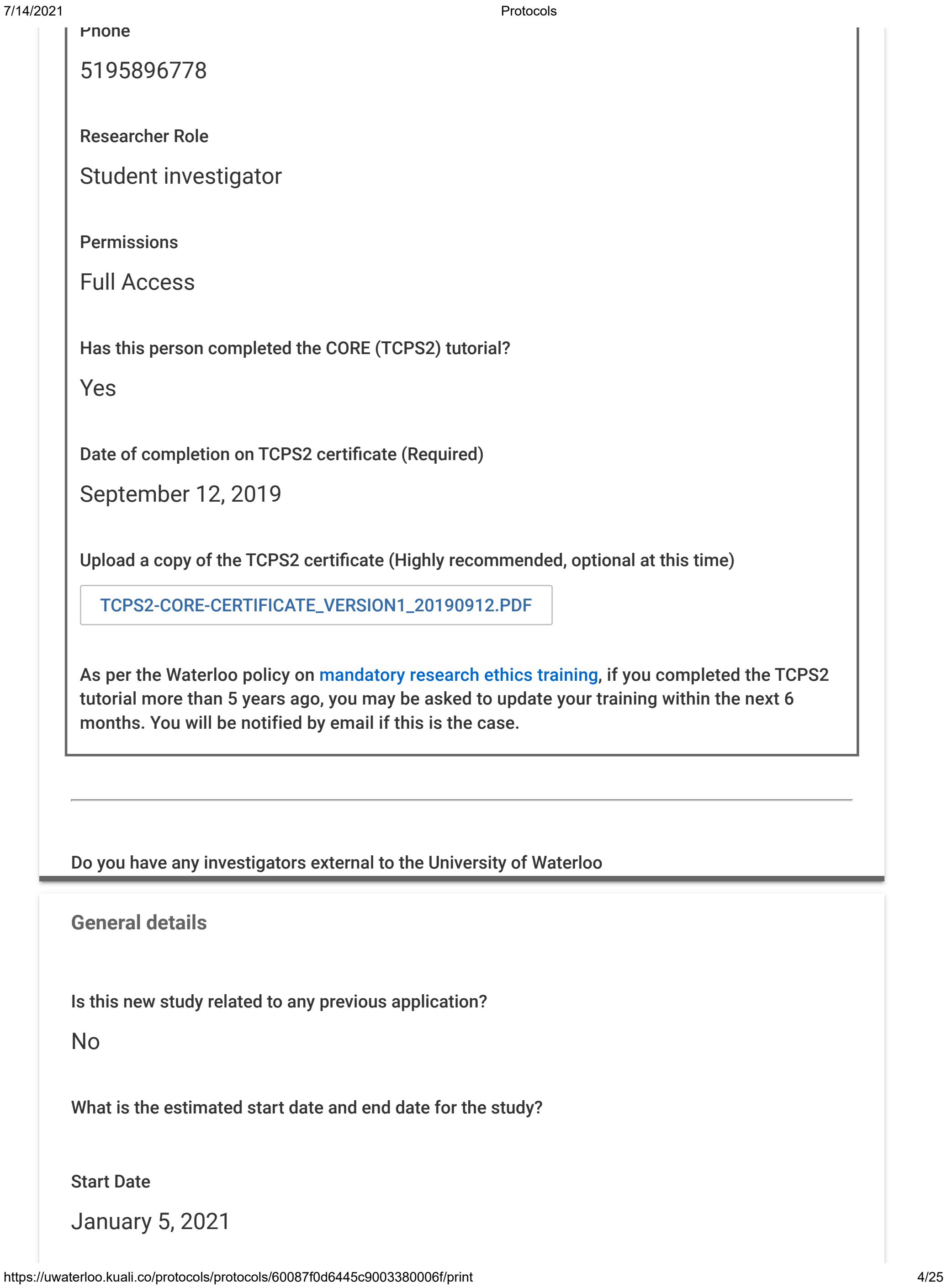
Waterloo Department

Faculty of Mathematics

Email Address

xj4wang@uwaterloo.ca

https://uwaterloo.kuali.co/protocols/protocols/60087f0d6445c9003380006f/print 3/25



Phone

5195896778

Researcher Role

Student investigator

Permissions

Full Access

Has this person completed the CORE (TCPS2) tutorial?

Yes

Date of completion on TCPS2 certificate (Required)

September 12, 2019

Upload a copy of the TCPS2 certificate (Highly recommended, optional at this time)

TCPS2-CORE-CERTIFICATE_VERSION1_20190912.PDF

As per the Waterloo policy on mandatory research ethics training, if you completed the TCPS2 tutorial more than 5 years ago, you may be asked to update your training within the next 6 months. You will be notified by email if this is the case.

Do you have any investigators external to the University of Waterloo

General details

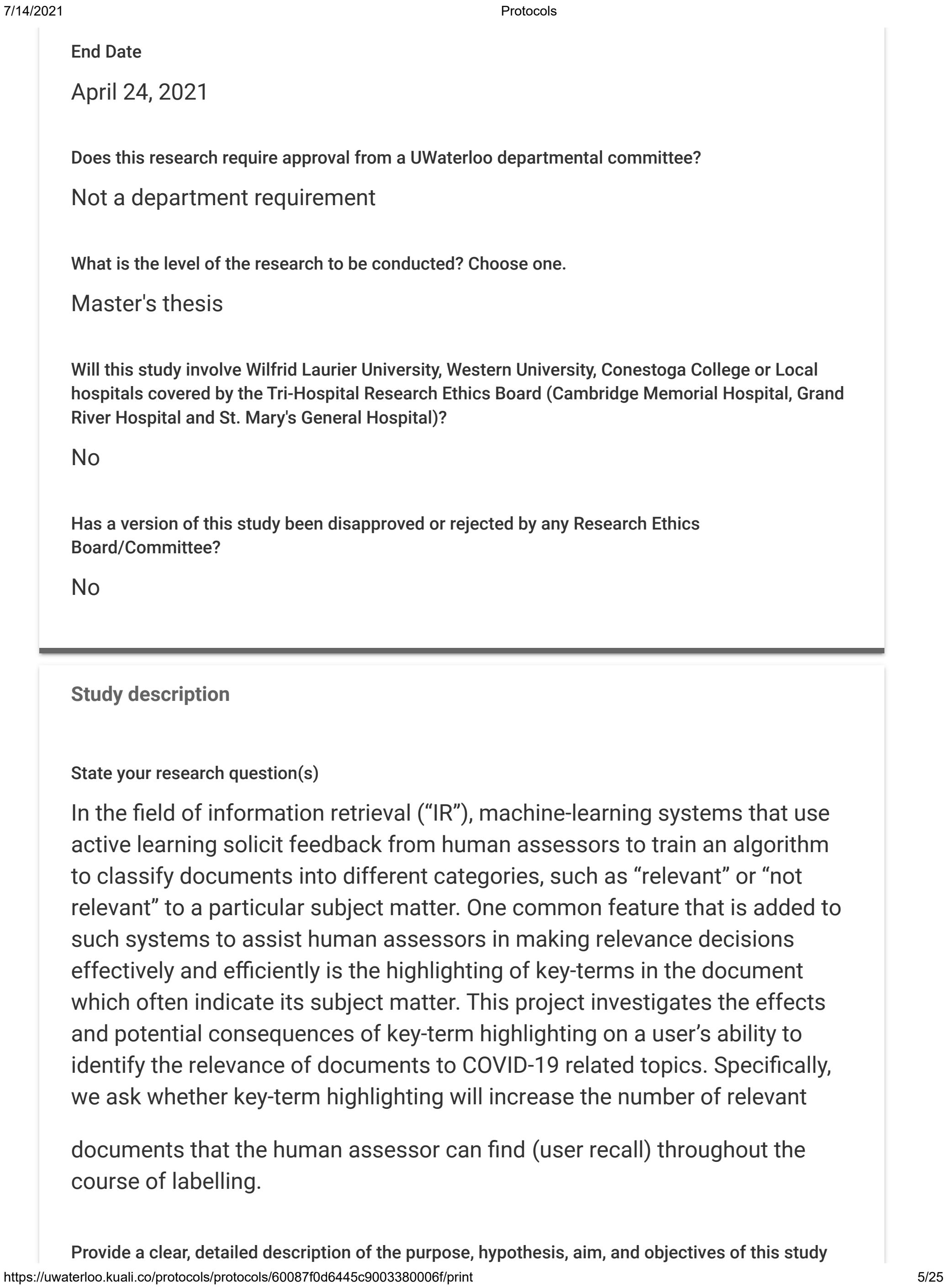
Is this new study related to any previous application?

No

What is the estimated start date and end date for the study?

Start Date

January 5, 2021



End Date

April 24, 2021

Does this research require approval from a UWaterloo departmental committee?

Not a department requirement

What is the level of the research to be conducted? Choose one.

Master's thesis

Will this study involve Wilfrid Laurier University, Western University, Conestoga College or Local hospitals covered by the Tri-Hospital Research Ethics Board (Cambridge Memorial Hospital, Grand River Hospital and St. Mary's General Hospital)?

No

Has a version of this study been disapproved or rejected by any Research Ethics Board/Committee?

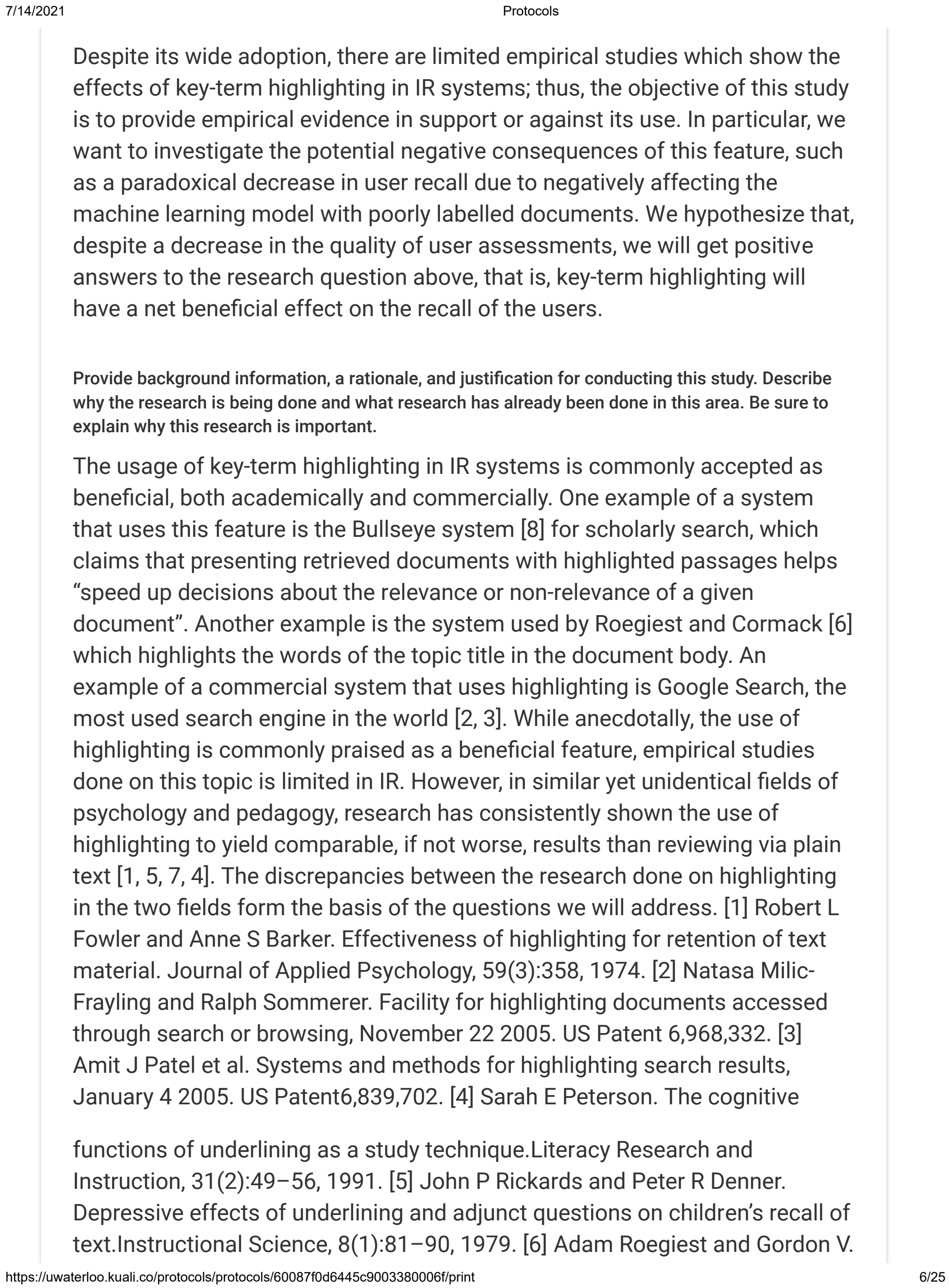
No

Study description

State your research question(s)

In the field of information retrieval (“IR”), machine-learning systems that use active learning solicit feedback from human assessors to train an algorithm to classify documents into different categories, such as “relevant” or “not relevant” to a particular subject matter. One common feature that is added to such systems to assist human assessors in making relevance decisions effectively and efficiently is the highlighting of key-terms in the document which often indicate its subject matter. This project investigates the effects and potential consequences of key-term highlighting on a user’s ability to identify the relevance of documents to COVID-19 related topics. Specifically, we ask whether key-term highlighting will increase the number of relevant documents that the human assessor can find (user recall) throughout the course of labelling.

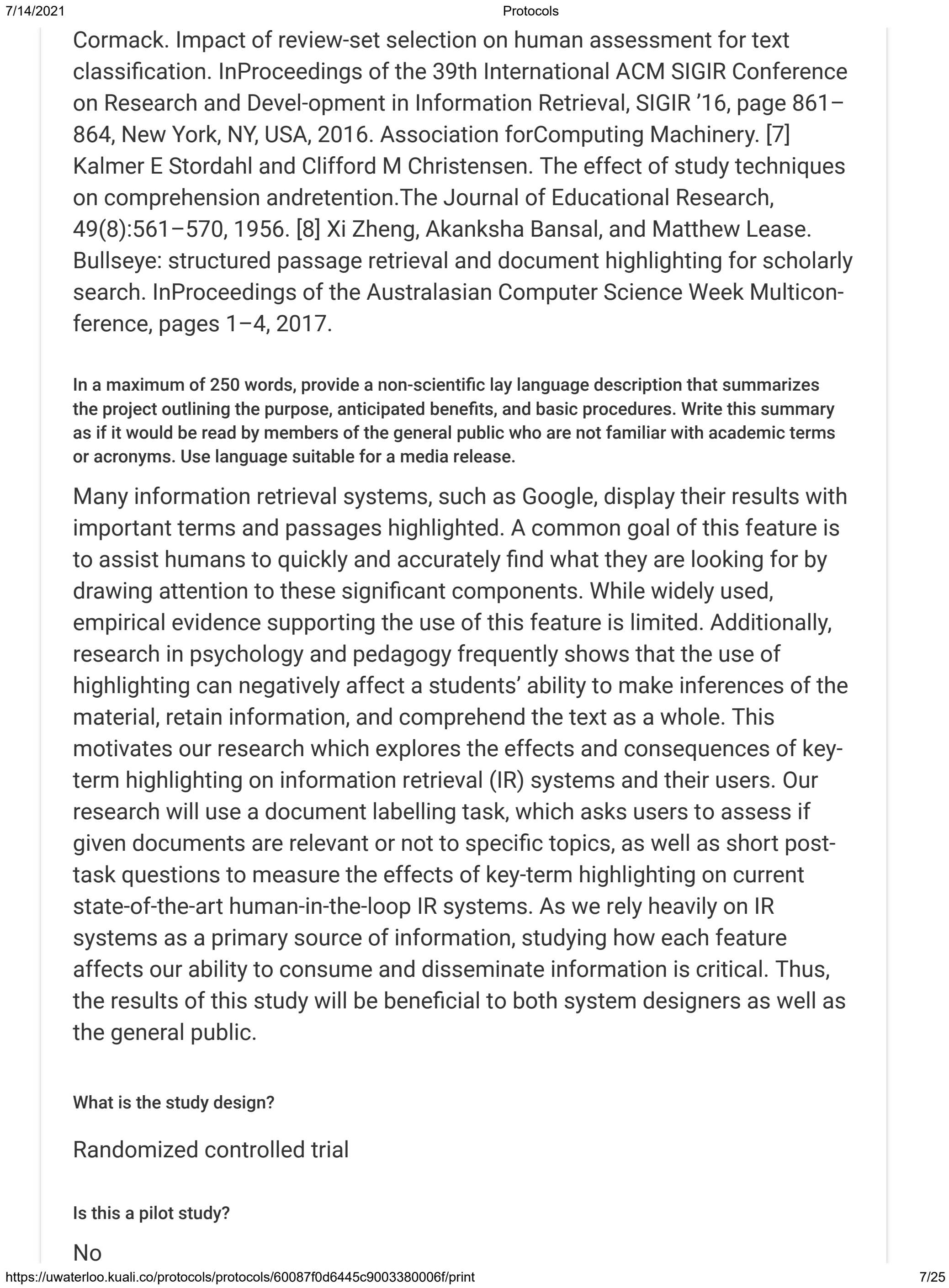
Provide a clear, detailed description of the purpose, hypothesis, aim, and objectives of this study



Despite its wide adoption, there are limited empirical studies which show the effects of key-term highlighting in IR systems; thus, the objective of this study is to provide empirical evidence in support or against its use. In particular, we want to investigate the potential negative consequences of this feature, such as a paradoxical decrease in user recall due to negatively affecting the machine learning model with poorly labelled documents. We hypothesize that, despite a decrease in the quality of user assessments, we will get positive answers to the research question above, that is, key-term highlighting will have a net beneficial effect on the recall of the users.

Provide background information, a rationale, and justification for conducting this study. Describe why the research is being done and what research has already been done in this area. Be sure to explain why this research is important.

The usage of key-term highlighting in IR systems is commonly accepted as beneficial, both academically and commercially. One example of a system that uses this feature is the Bullseye system [8] for scholarly search, which claims that presenting retrieved documents with highlighted passages helps “speed up decisions about the relevance or non-relevance of a given document”. Another example is the system used by Roegiest and Cormack [6] which highlights the words of the topic title in the document body. An example of a commercial system that uses highlighting is Google Search, the most used search engine in the world [2, 3]. While anecdotally, the use of highlighting is commonly praised as a beneficial feature, empirical studies done on this topic is limited in IR. However, in similar yet unidentical fields of psychology and pedagogy, research has consistently shown the use of highlighting to yield comparable, if not worse, results than reviewing via plain text [1, 5, 7, 4]. The discrepancies between the research done on highlighting in the two fields form the basis of the questions we will address. [1] Robert L Fowler and Anne S Barker. Effectiveness of highlighting for retention of text material. Journal of Applied Psychology, 59(3):358, 1974. [2] Natasa Milic-Frayling and Ralph Sommerer. Facility for highlighting documents accessed through search or browsing, November 22 2005. US Patent 6,968,332. [3] Amit J Patel et al. Systems and methods for highlighting search results, January 4 2005. US Patent6,839,702. [4] Sarah E Peterson. The cognitive functions of underlining as a study technique.Literacy Research and Instruction, 31(2):49–56, 1991. [5] John P Rickards and Peter R Denner. Depressive effects of underlining and adjunct questions on children’s recall of text.Instructional Science, 8(1):81–90, 1979. [6] Adam Roegiest and Gordon V.



Cormack. Impact of review-set selection on human assessment for text classification. InProceedings of the 39th International ACM SIGIR Conference on Research and Devel-opment in Information Retrieval, SIGIR ’16, page 861–864, New York, NY, USA, 2016. Association forComputing Machinery. [7] Kalmer E Stordahl and Clifford M Christensen. The effect of study techniques on comprehension andretention.The Journal of Educational Research, 49(8):561–570, 1956. [8] Xi Zheng, Akanksha Bansal, and Matthew Lease. Bullseye: structured passage retrieval and document highlighting for scholarly search. InProceedings of the Australasian Computer Science Week Multiconference, pages 1–4, 2017.

In a maximum of 250 words, provide a non-scientific lay language description that summarizes the project outlining the purpose, anticipated benefits, and basic procedures. Write this summary as if it would be read by members of the general public who are not familiar with academic terms or acronyms. Use language suitable for a media release.

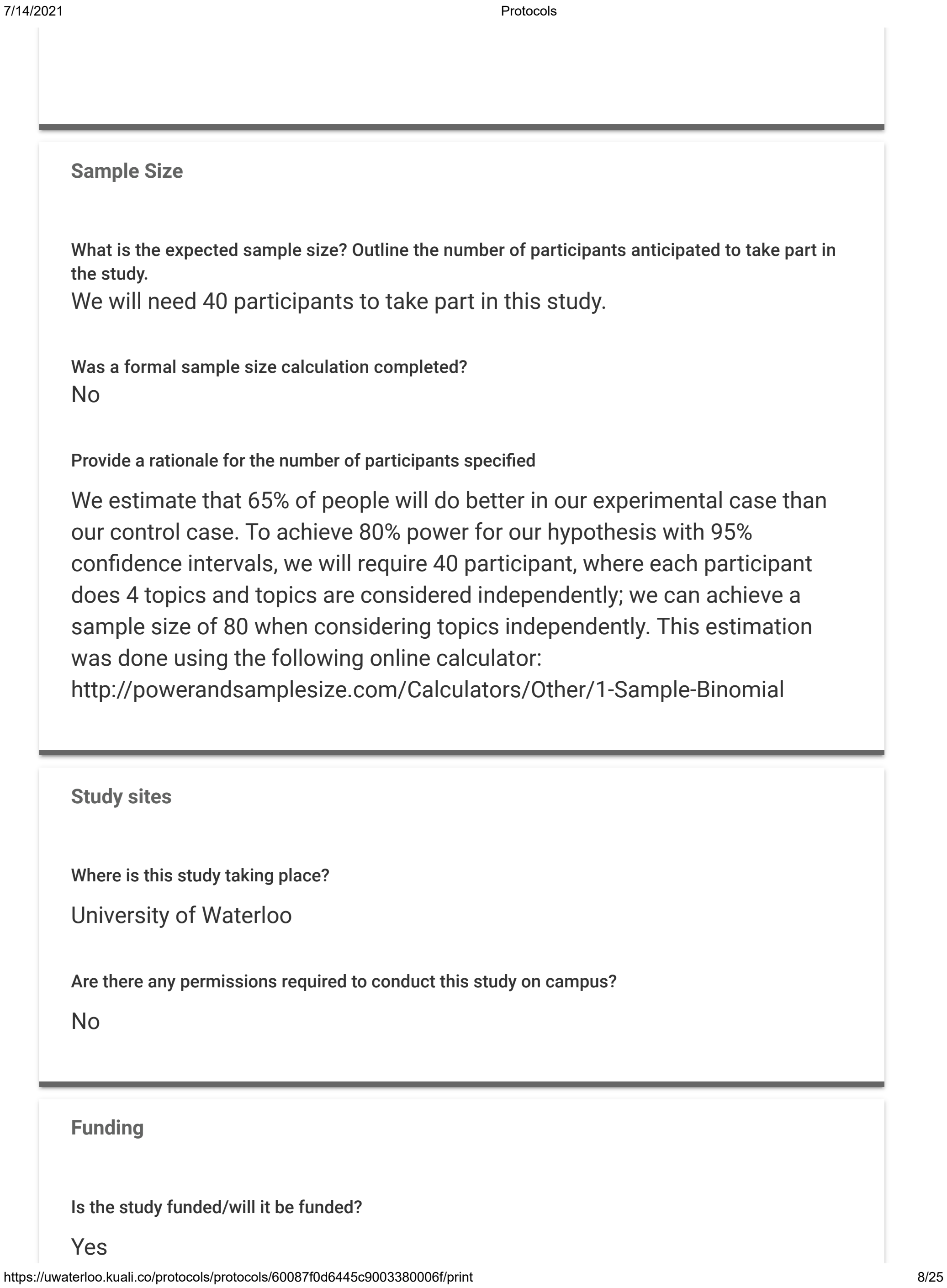
Many information retrieval systems, such as Google, display their results with important terms and passages highlighted. A common goal of this feature is to assist humans to quickly and accurately find what they are looking for by drawing attention to these significant components. While widely used, empirical evidence supporting the use of this feature is limited. Additionally, research in psychology and pedagogy frequently shows that the use of highlighting can negatively affect a students’ ability to make inferences of the material, retain information, and comprehend the text as a whole. This motivates our research which explores the effects and consequences of key-term highlighting on information retrieval (IR) systems and their users. Our research will use a document labelling task, which asks users to assess if given documents are relevant or not to specific topics, as well as short post-task questions to measure the effects of key-term highlighting on current state-of-the-art human-in-the-loop IR systems. As we rely heavily on IR systems as a primary source of information, studying how each feature affects our ability to consume and disseminate information is critical. Thus, the results of this study will be beneficial to both system designers as well as the general public.

What is the study design?

Randomized controlled trial

Is this a pilot study?

No



Sample Size

What is the expected sample size? Outline the number of participants anticipated to take part in the study.

We will need 40 participants to take part in this study.

Was a formal sample size calculation completed?

No

Provide a rationale for the number of participants specified

We estimate that 65% of people will do better in our experimental case than our control case. To achieve 80% power for our hypothesis with 95% confidence intervals, we will require 40 participant, where each participant does 4 topics and topics are considered independently; we can achieve a sample size of 80 when considering topics independently. This estimation was done using the following online calculator: http://powerandsamplesize.com/Calculators/Other/1-Sample-Binomial

Study sites

Where is this study taking place?

University of Waterloo

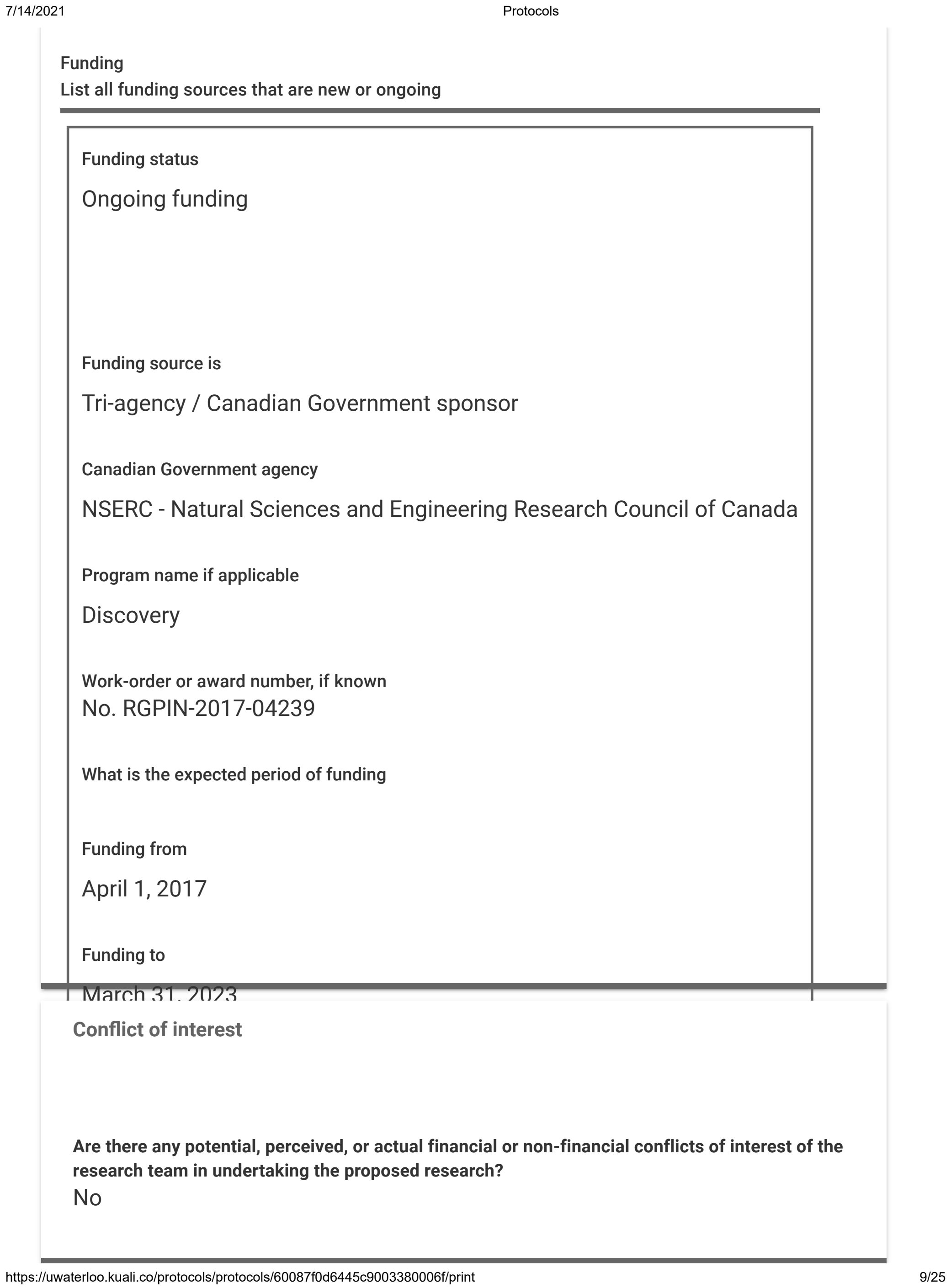
Are there any permissions required to conduct this study on campus?

No

Funding

Is the study funded/will it be funded?

Yes



Funding
List all funding sources that are new or ongoing

Funding status

Ongoing funding

Funding source is

Tri-agency / Canadian Government sponsor

Canadian Government agency

NSERC - Natural Sciences and Engineering Research Council of Canada

Program name if applicable

Discovery

Work-order or award number, if known
No. RGPIN-2017-04239

What is the expected period of funding

Funding from

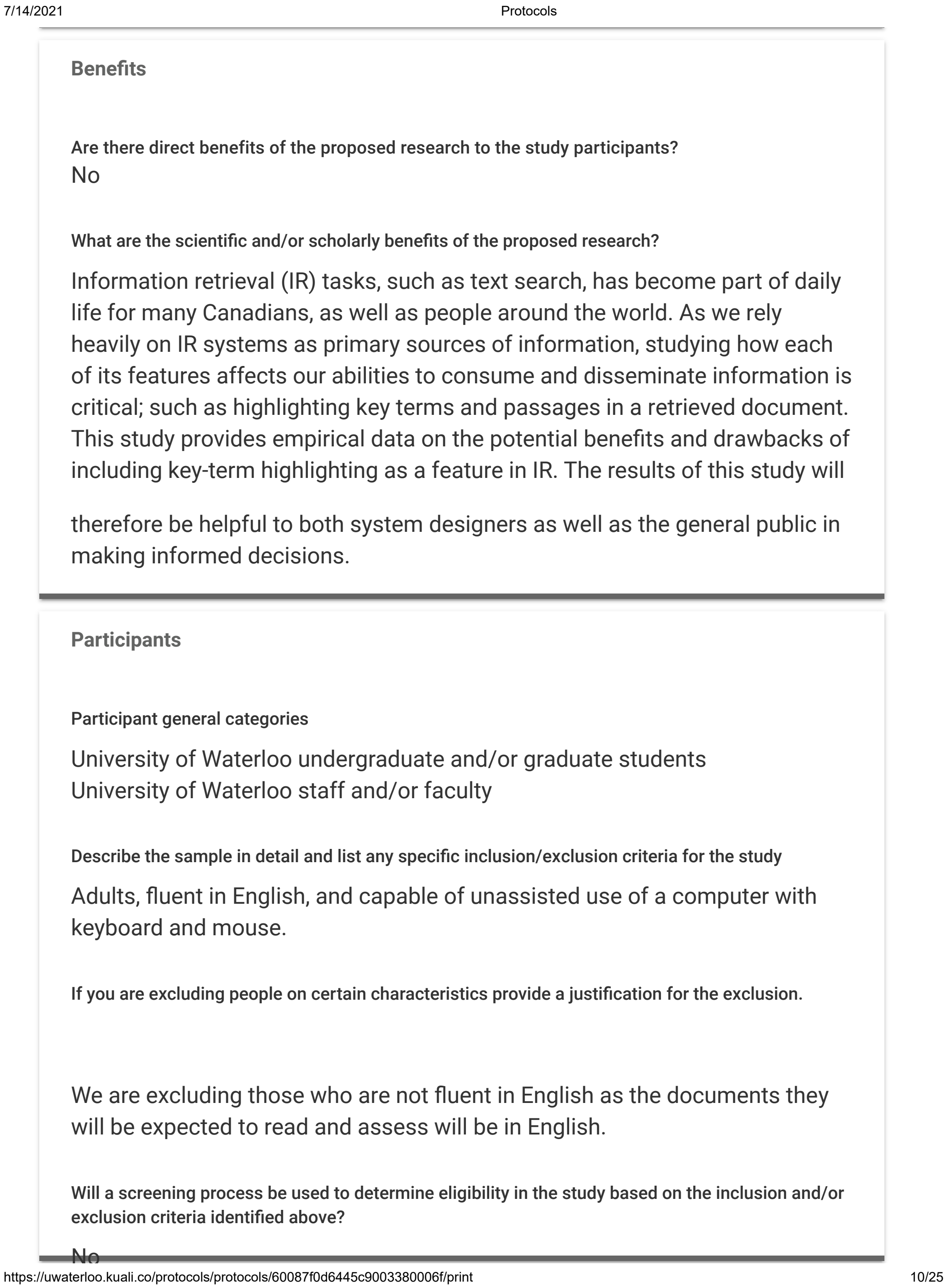
April 1, 2017

Funding to

March 31, 2023

Conflict of interest

Are there any potential, perceived, or actual financial or non-financial conflicts of interest of the research team in undertaking the proposed research?
No



7/14/2021 Protocols

Benefits

Are there direct benefits of the proposed research to the study participants?

No

What are the scientific and/or scholarly benefits of the proposed research?

Information retrieval (IR) tasks, such as text search, has become part of daily life for many Canadians, as well as people around the world. As we rely heavily on IR systems as primary sources of information, studying how each of its features affects our abilities to consume and disseminate information is critical; such as highlighting key terms and passages in a retrieved document. This study provides empirical data on the potential benefits and drawbacks of including key-term highlighting as a feature in IR. The results of this study will

therefore be helpful to both system designers as well as the general public in making informed decisions.

Participants

Participant general categories

University of Waterloo undergraduate and/or graduate students
University of Waterloo staff and/or faculty

Describe the sample in detail and list any specific inclusion/exclusion criteria for the study

Adults, fluent in English, and capable of unassisted use of a computer with keyboard and mouse.

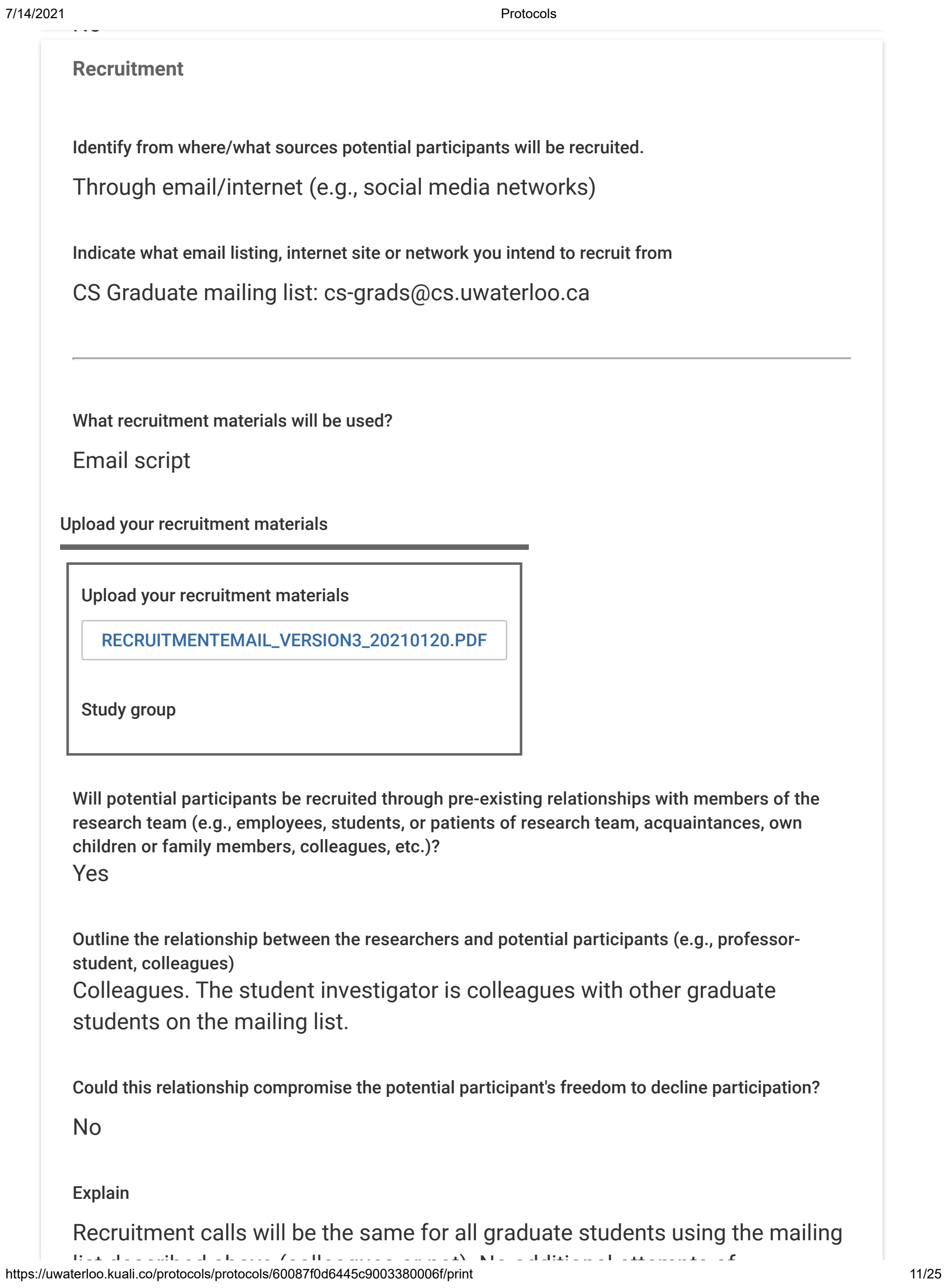
If you are excluding people on certain characteristics provide a justification for the exclusion.

We are excluding those who are not fluent in English as the documents they will be expected to read and assess will be in English.

Will a screening process be used to determine eligibility in the study based on the inclusion and/or exclusion criteria identified above?

No

https://uwaterloo.kuali.co/protocols/protocols/60087f0d6445c9003380006f/print 10/25



Recruitment

Identify from where/what sources potential participants will be recruited.

Through email/internet (e.g., social media networks)

Indicate what email listing, internet site or network you intend to recruit from

CS Graduate mailing list: cs-grads@cs.uwaterloo.ca

What recruitment materials will be used?

Email script

Upload your recruitment materials

Upload your recruitment materials

RECRUITMENTEMAIL_VERSION3_20210120.PDF

Study group

Will potential participants be recruited through pre-existing relationships with members of the research team (e.g., employees, students, or patients of research team, acquaintances, own children or family members, colleagues, etc.)?

Yes

Outline the relationship between the researchers and potential participants (e.g., professor-student, colleagues)

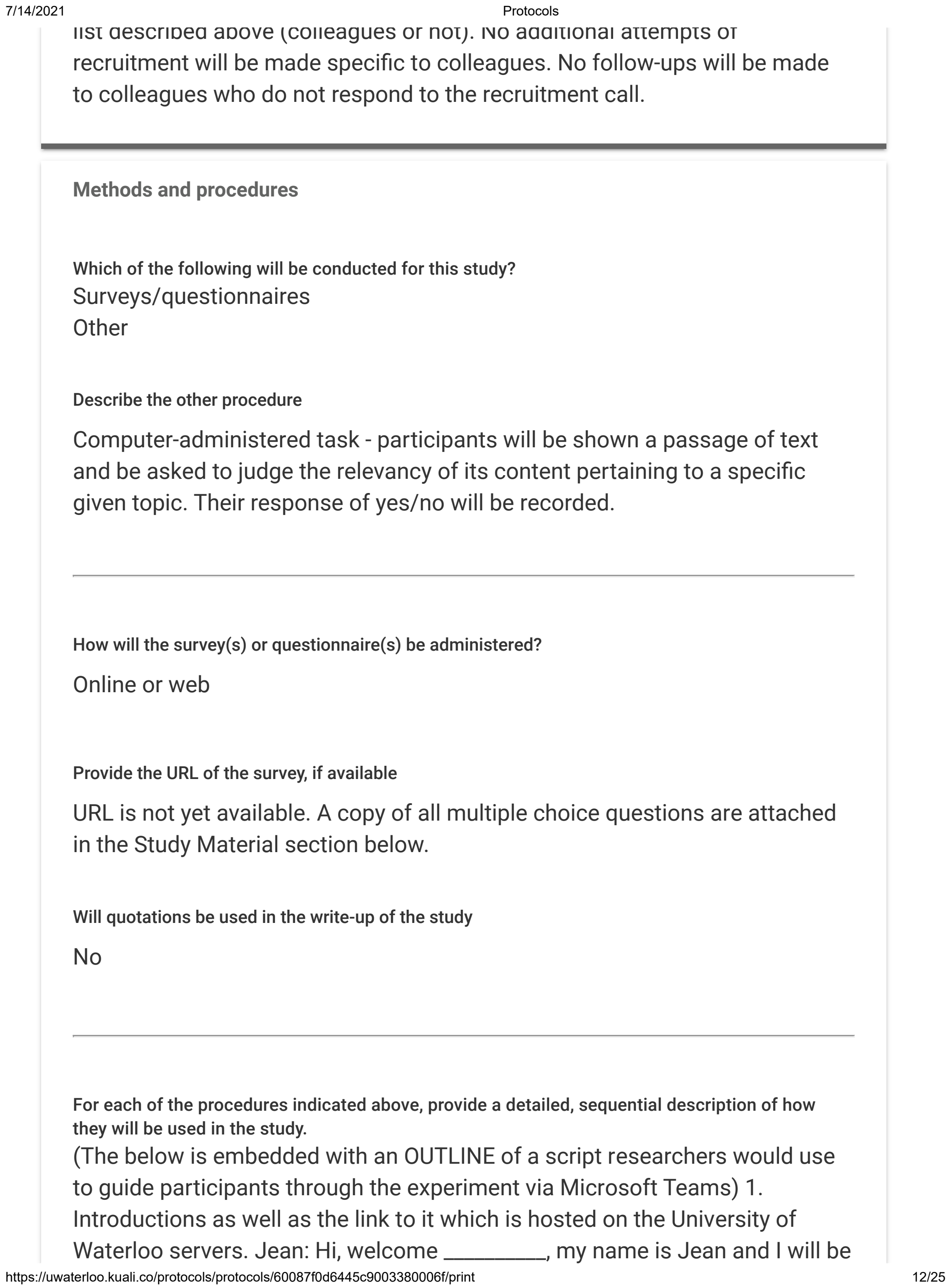
Colleagues. The student investigator is colleagues with other graduate students on the mailing list.

Could this relationship compromise the potential participant's freedom to decline participation?

No

Explain

Recruitment calls will be the same for all graduate students using the mailing list described above (colleagues or not). No additional attempts of



list described above (colleagues or not). No additional attempts of recruitment will be made specific to colleagues. No follow-ups will be made to colleagues who do not respond to the recruitment call.

Methods and procedures

Which of the following will be conducted for this study?

Surveys/questionnaires
Other

Describe the other procedure

Computer-administered task - participants will be shown a passage of text and be asked to judge the relevancy of its content pertaining to a specific given topic. Their response of yes/no will be recorded.

How will the survey(s) or questionnaire(s) be administered?

Online or web

Provide the URL of the survey, if available

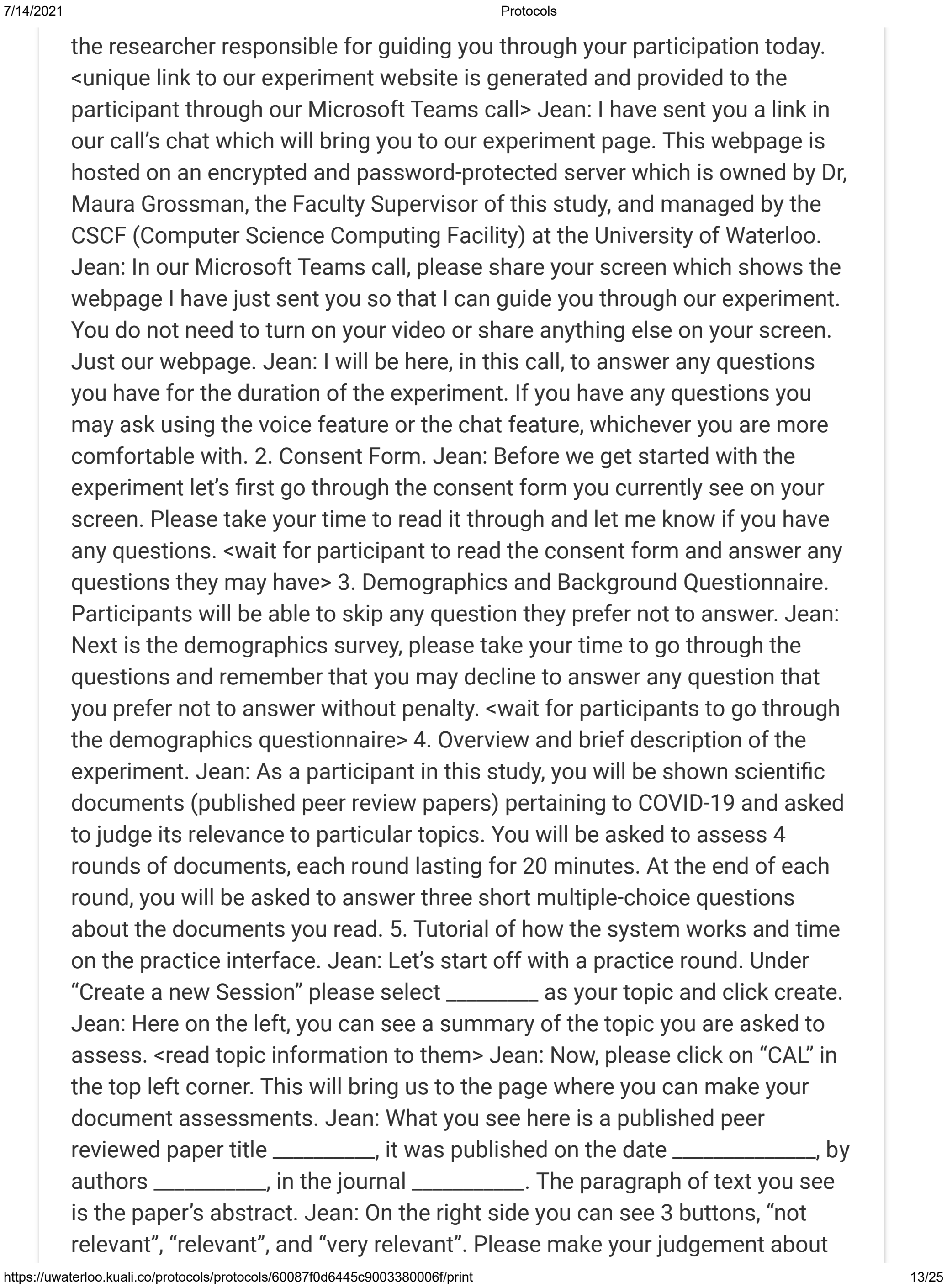
URL is not yet available. A copy of all multiple choice questions are attached in the Study Material section below.

Will quotations be used in the write-up of the study

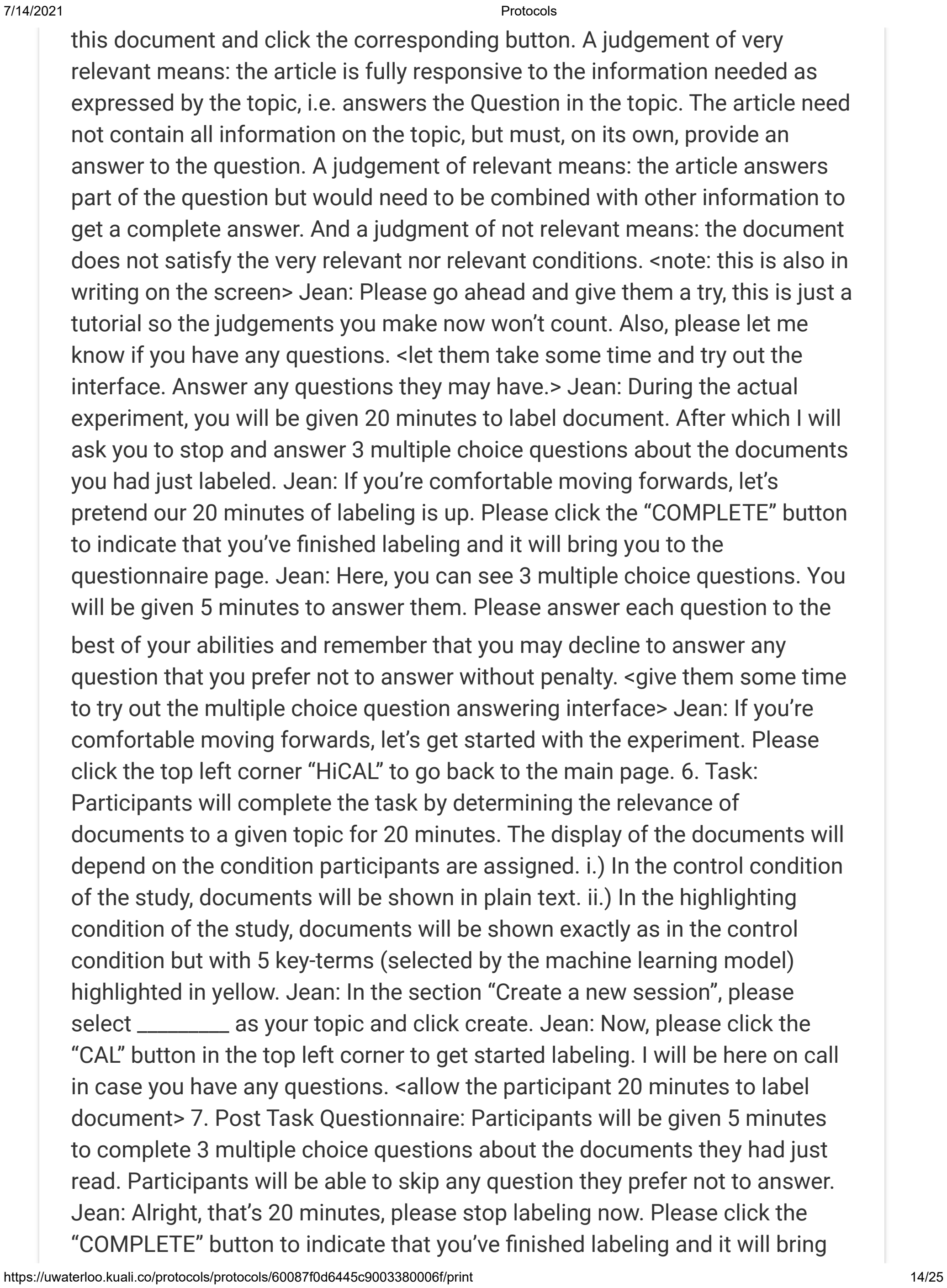
No

For each of the procedures indicated above, provide a detailed, sequential description of how they will be used in the study.

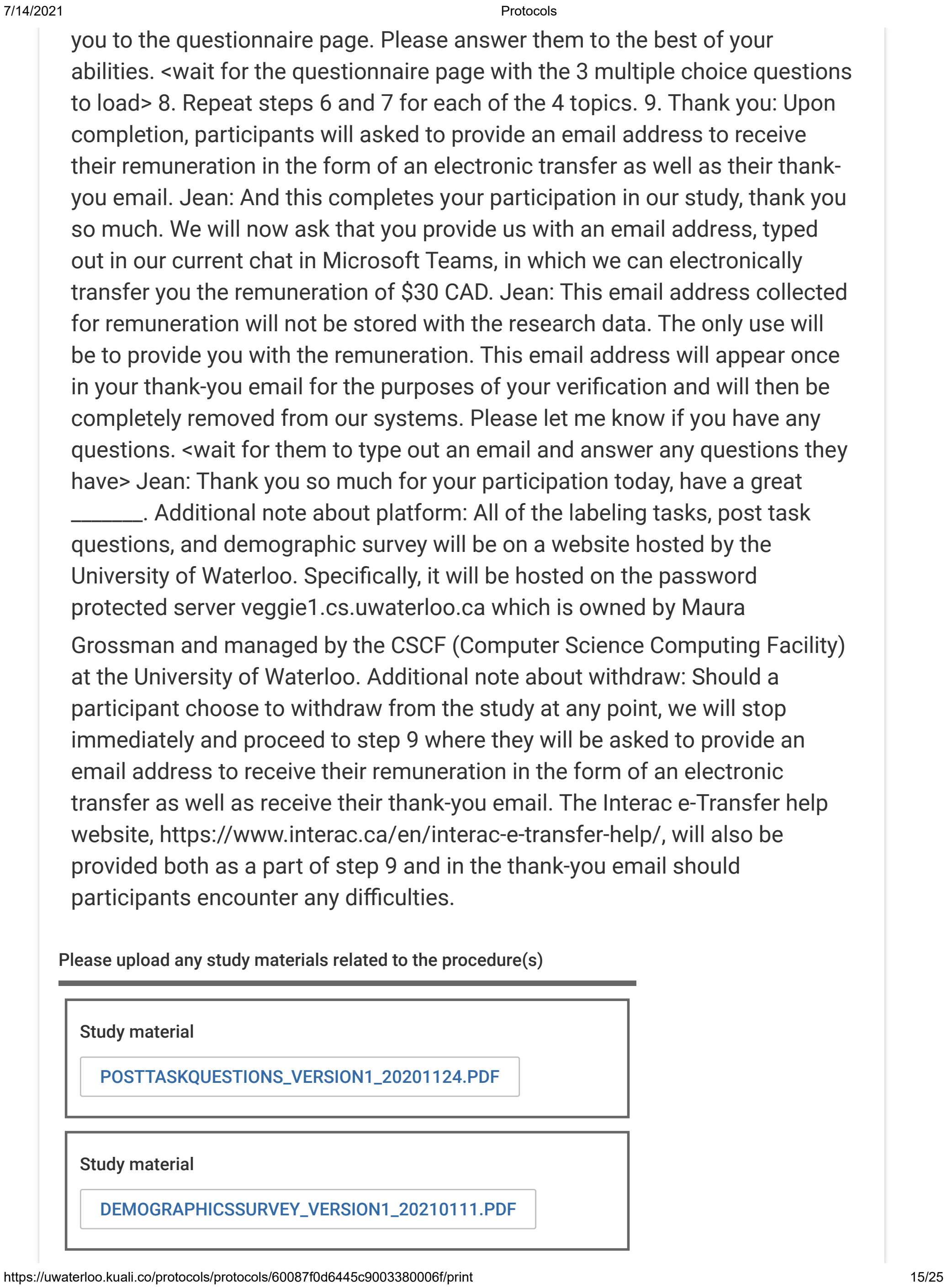
(The below is embedded with an OUTLINE of a script researchers would use to guide participants through the experiment via Microsoft Teams) 1. Introductions as well as the link to it which is hosted on the University of Waterloo servers. Jean: Hi, welcome __________, my name is Jean and I will be



the researcher responsible for guiding you through your participation today. <unique link to our experiment website is generated and provided to the participant through our Microsoft Teams call> Jean: I have sent you a link in our call's chat which will bring you to our experiment page. This webpage is hosted on an encrypted and password-protected server which is owned by Dr, Maura Grossman, the Faculty Supervisor of this study, and managed by the CSCF (Computer Science Computing Facility) at the University of Waterloo. Jean: In our Microsoft Teams call, please share your screen which shows the webpage I have just sent you so that I can guide you through our experiment. You do not need to turn on your video or share anything else on your screen. Just our webpage. Jean: I will be here, in this call, to answer any questions you have for the duration of the experiment. If you have any questions you may ask using the voice feature or the chat feature, whichever you are more comfortable with. 2. Consent Form. Jean: Before we get started with the experiment let's first go through the consent form you currently see on your screen. Please take your time to read it through and let me know if you have any questions. <wait for participant to read the consent form and answer any questions they may have> 3. Demographics and Background Questionnaire. Participants will be able to skip any question they prefer not to answer. Jean: Next is the demographics survey, please take your time to go through the questions and remember that you may decline to answer any question that you prefer not to answer without penalty. <wait for participants to go through the demographics questionnaire> 4. Overview and brief description of the experiment. Jean: As a participant in this study, you will be shown scientific documents (published peer review papers) pertaining to COVID-19 and asked to judge its relevance to particular topics. You will be asked to assess 4 rounds of documents, each round lasting for 20 minutes. At the end of each round, you will be asked to answer three short multiple-choice questions about the documents you read. 5. Tutorial of how the system works and time on the practice interface. Jean: Let's start off with a practice round. Under "Create a new Session" please select ________ as your topic and click create. Jean: Here on the left, you can see a summary of the topic you are asked to assess. <read topic information to them> Jean: Now, please click on "CAL" in the top left corner. This will bring us to the page where you can make your document assessments. Jean: What you see here is a published peer reviewed paper title _________, it was published on the date ____________, by authors __________, in the journal __________. The paragraph of text you see is the paper's abstract. Jean: On the right side you can see 3 buttons, "not relevant", "relevant", and "very relevant". Please make your judgement about



this document and click the corresponding button. A judgement of very relevant means: the article is fully responsive to the information needed as expressed by the topic, i.e. answers the Question in the topic. The article need not contain all information on the topic, but must, on its own, provide an answer to the question. A judgement of relevant means: the article answers part of the question but would need to be combined with other information to get a complete answer. And a judgment of not relevant means: the document does not satisfy the very relevant nor relevant conditions. <note: this is also in writing on the screen> Jean: Please go ahead and give them a try, this is just a tutorial so the judgements you make now won't count. Also, please let me know if you have any questions. <let them take some time and try out the interface. Answer any questions they may have.> Jean: During the actual experiment, you will be given 20 minutes to label document. After which I will ask you to stop and answer 3 multiple choice questions about the documents you had just labeled. Jean: If you're comfortable moving forwards, let's pretend our 20 minutes of labeling is up. Please click the "COMPLETE" button to indicate that you've finished labeling and it will bring you to the questionnaire page. Jean: Here, you can see 3 multiple choice questions. You will be given 5 minutes to answer them. Please answer each question to the best of your abilities and remember that you may decline to answer any question that you prefer not to answer without penalty. <give them some time to try out the multiple choice question answering interface> Jean: If you're comfortable moving forwards, let's get started with the experiment. Please click the top left corner "HiCAL" to go back to the main page. 6. Task: Participants will complete the task by determining the relevance of documents to a given topic for 20 minutes. The display of the documents will depend on the condition participants are assigned. i.) In the control condition of the study, documents will be shown in plain text. ii.) In the highlighting condition of the study, documents will be shown exactly as in the control condition but with 5 key-terms (selected by the machine learning model) highlighted in yellow. Jean: In the section "Create a new session", please select _________ as your topic and click create. Jean: Now, please click the "CAL" button in the top left corner to get started labeling. I will be here on call in case you have any questions. <allow the participant 20 minutes to label document> 7. Post Task Questionnaire: Participants will be given 5 minutes to complete 3 multiple choice questions about the documents they had just read. Participants will be able to skip any question they prefer not to answer. Jean: Alright, that's 20 minutes, please stop labeling now. Please click the "COMPLETE" button to indicate that you've finished labeling and it will bring



you to the questionnaire page. Please answer them to the best of your abilities. <wait for the questionnaire page with the 3 multiple choice questions to load> 8. Repeat steps 6 and 7 for each of the 4 topics. 9. Thank you: Upon completion, participants will asked to provide an email address to receive their remuneration in the form of an electronic transfer as well as their thank-you email. Jean: And this completes your participation in our study, thank you so much. We will now ask that you provide us with an email address, typed out in our current chat in Microsoft Teams, in which we can electronically transfer you the remuneration of $30 CAD. Jean: This email address collected for remuneration will not be stored with the research data. The only use will be to provide you with the remuneration. This email address will appear once in your thank-you email for the purposes of your verification and will then be completely removed from our systems. Please let me know if you have any questions. <wait for them to type out an email and answer any questions they have> Jean: Thank you so much for your participation today, have a great _______. Additional note about platform: All of the labeling tasks, post task questions, and demographic survey will be on a website hosted by the University of Waterloo. Specifically, it will be hosted on the password protected server veggie1.cs.uwaterloo.ca which is owned by Maura Grossman and managed by the CSCF (Computer Science Computing Facility) at the University of Waterloo. Additional note about withdraw: Should a participant choose to withdraw from the study at any point, we will stop immediately and proceed to step 9 where they will be asked to provide an email address to receive their remuneration in the form of an electronic transfer as well as receive their thank-you email. The Interac e-Transfer help website, https://www.interac.ca/en/interac-e-transfer-help/, will also be provided both as a part of step 9 and in the thank-you email should participants encounter any difficulties.

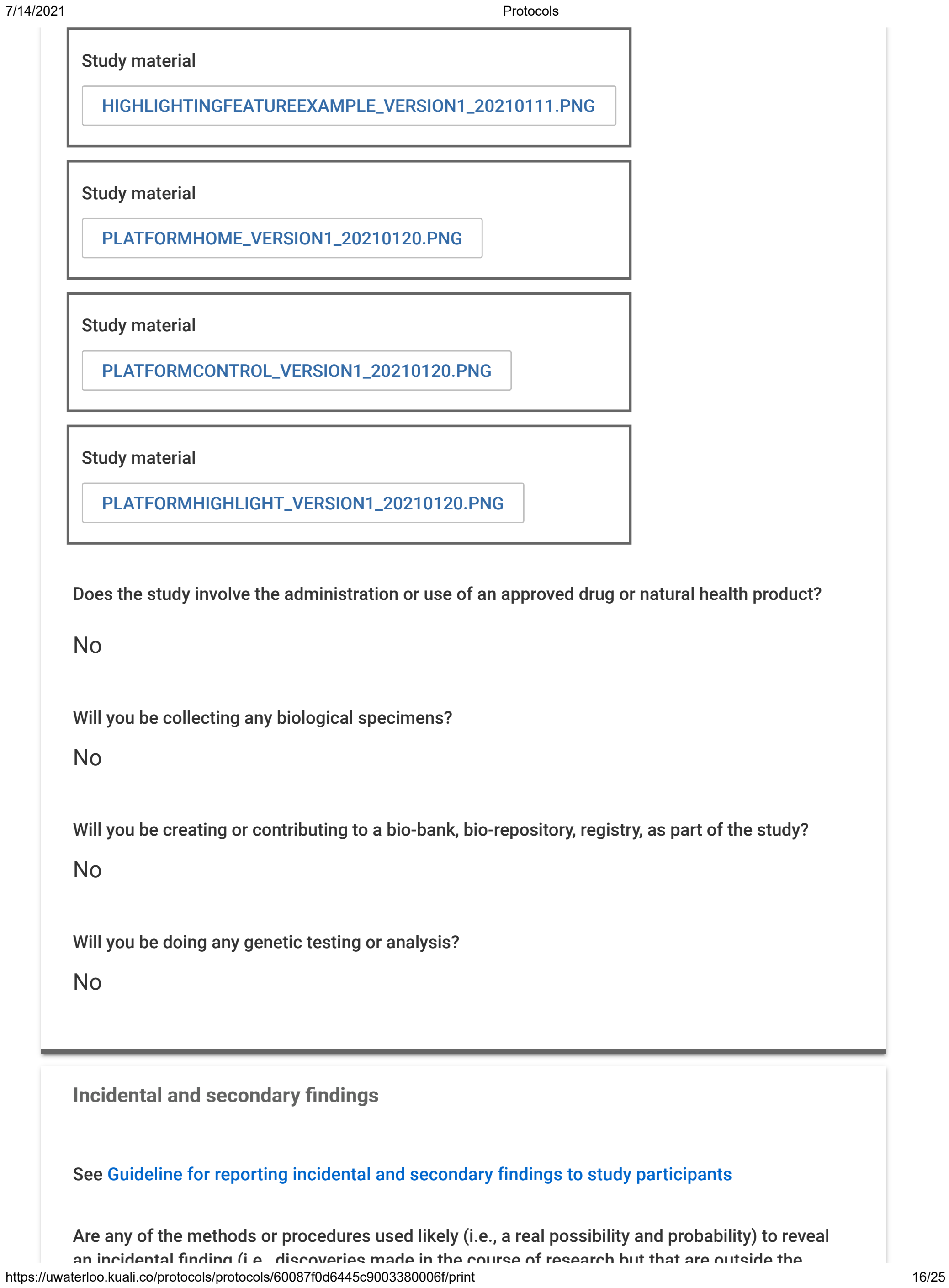
Please upload any study materials related to the procedure(s)

Study material

POSTTASKQUESTIONS_VERSION1_20201124.PDF

Study material

DEMOGRAPHICSSURVEY_VERSION1_20210111.PDF



7/14/2021 Protocols

Study material

HIGHLIGHTINGFEATUREEXAMPLE_VERSION1_20210111.PNG

Study material

PLATFORMHOME_VERSION1_20210120.PNG

Study material

PLATFORMCONTROL_VERSION1_20210120.PNG

Study material

PLATFORMHIGHLIGHT_VERSION1_20210120.PNG

Does the study involve the administration or use of an approved drug or natural health product?

No

Will you be collecting any biological specimens?

No

Will you be creating or contributing to a bio-bank, bio-repository, registry, as part of the study?

No

Will you be doing any genetic testing or analysis?

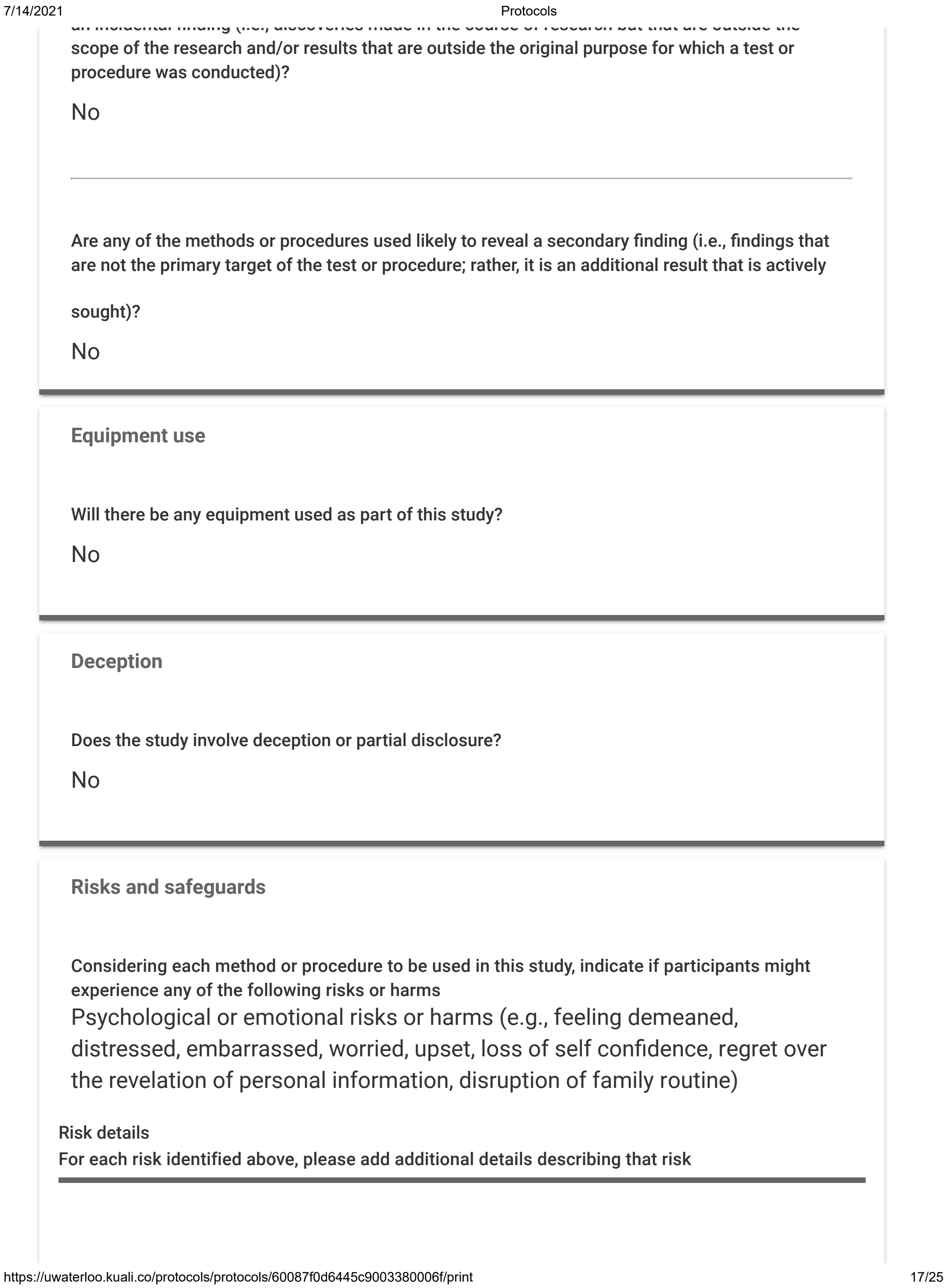
No

Incidental and secondary findings

See Guideline for reporting incidental and secondary findings to study participants

Are any of the methods or procedures used likely (i.e., a real possibility and probability) to reveal an incidental finding (i.e., discoveries made in the course of research but that are outside the

https://uwaterloo.kuali.co/protocols/protocols/60087f0d6445c9003380006f/print 16/25



an incidental finding (i.e., discoveries made in the course of research but that are outside the scope of the research and/or results that are outside the original purpose for which a test or procedure was conducted)?

No

Are any of the methods or procedures used likely to reveal a secondary finding (i.e., findings that are not the primary target of the test or procedure; rather, it is an additional result that is actively

sought)?

No

Equipment use

Will there be any equipment used as part of this study?

No

Deception

Does the study involve deception or partial disclosure?

No

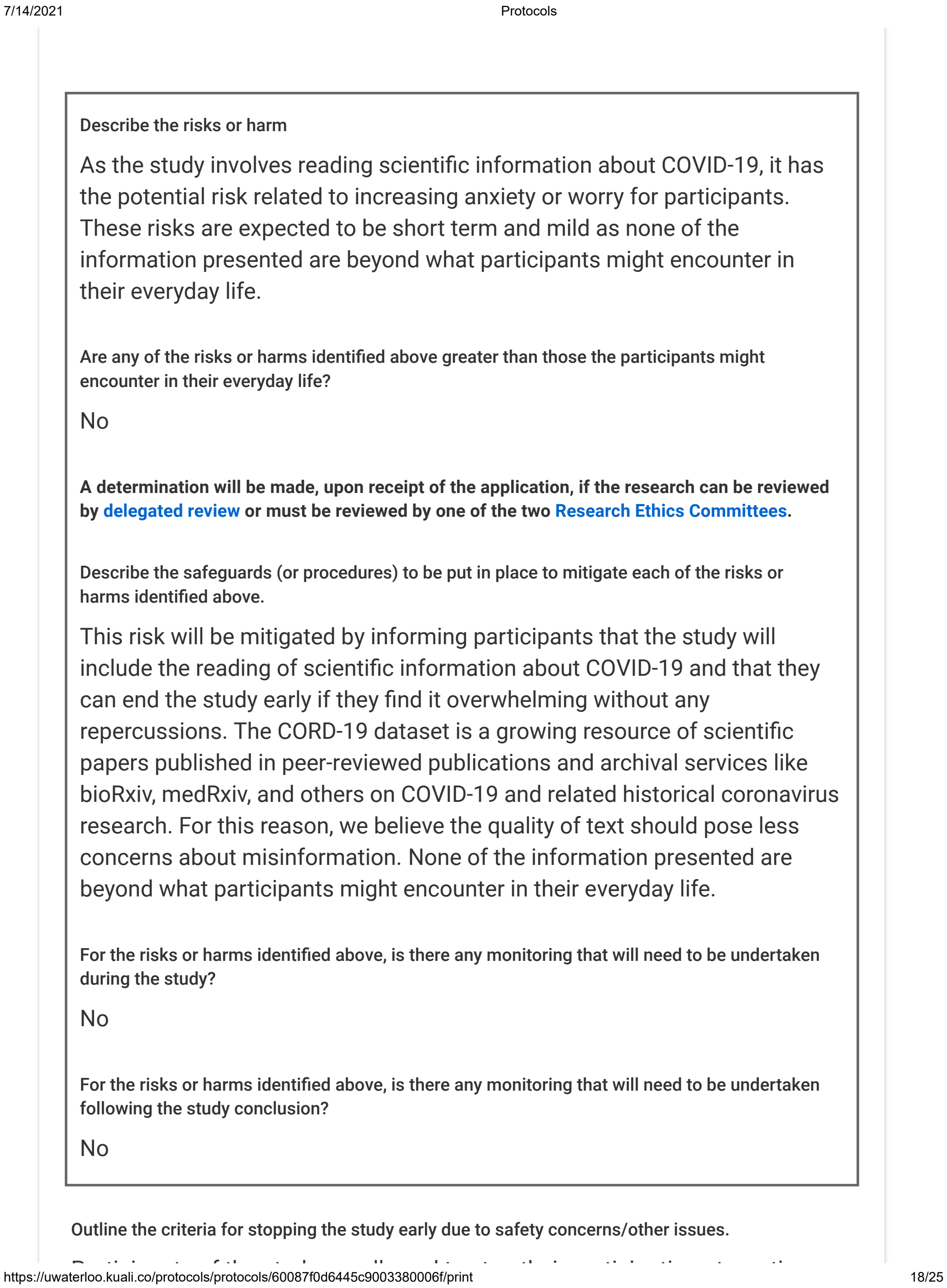
Risks and safeguards

Considering each method or procedure to be used in this study, indicate if participants might experience any of the following risks or harms

Psychological or emotional risks or harms (e.g., feeling demeaned, distressed, embarrassed, worried, upset, loss of self confidence, regret over the revelation of personal information, disruption of family routine)

Risk details

For each risk identified above, please add additional details describing that risk



Describe the risks or harm

As the study involves reading scientific information about COVID-19, it has the potential risk related to increasing anxiety or worry for participants. These risks are expected to be short term and mild as none of the information presented are beyond what participants might encounter in their everyday life.

Are any of the risks or harms identified above greater than those the participants might encounter in their everyday life?

No

A determination will be made, upon receipt of the application, if the research can be reviewed by delegated review or must be reviewed by one of the two Research Ethics Committees.

Describe the safeguards (or procedures) to be put in place to mitigate each of the risks or harms identified above.

This risk will be mitigated by informing participants that the study will include the reading of scientific information about COVID-19 and that they can end the study early if they find it overwhelming without any repercussions. The CORD-19 dataset is a growing resource of scientific papers published in peer-reviewed publications and archival services like bioRxiv, medRxiv, and others on COVID-19 and related historical coronavirus research. For this reason, we believe the quality of text should pose less concerns about misinformation. None of the information presented are beyond what participants might encounter in their everyday life.

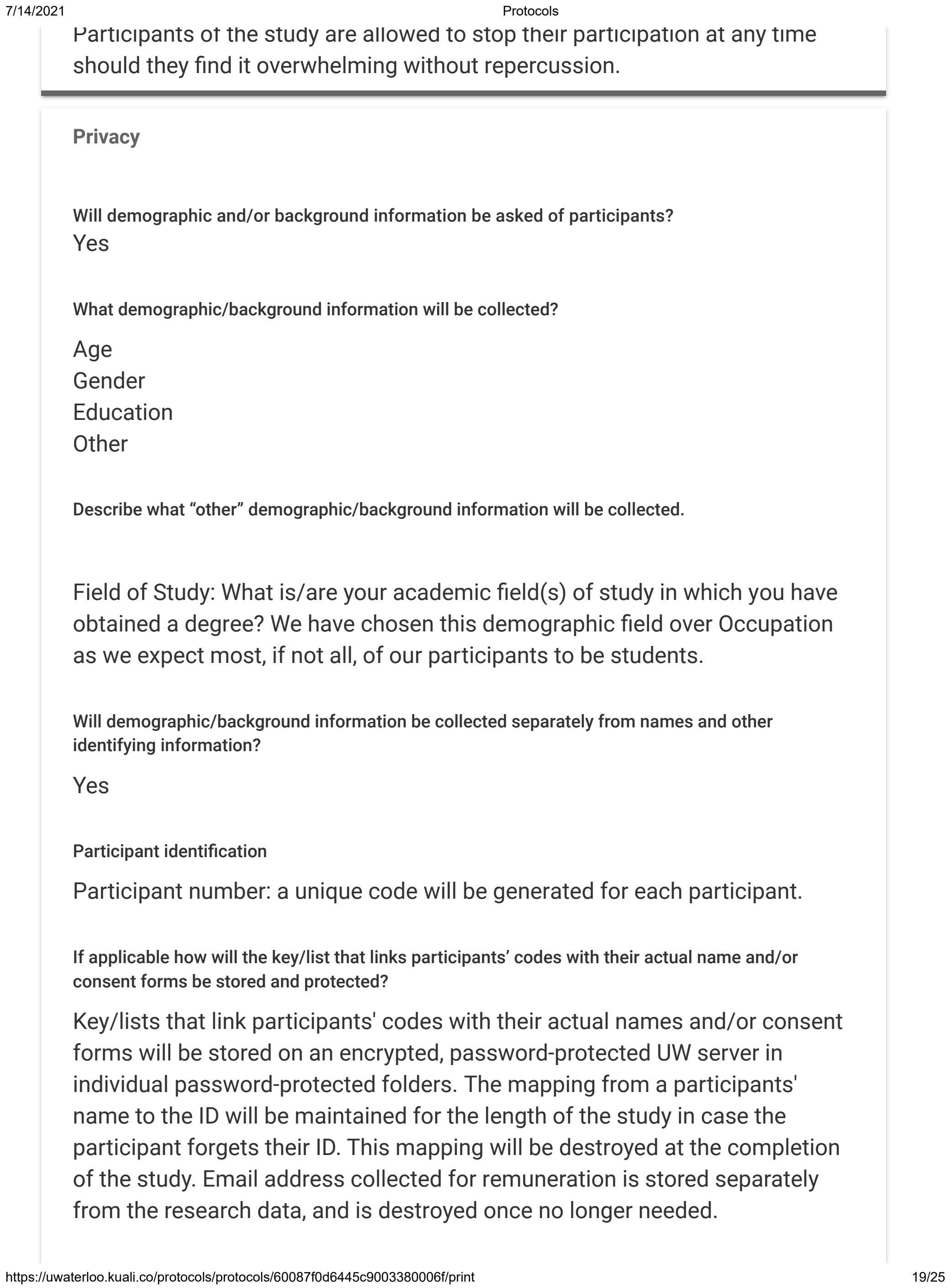
For the risks or harms identified above, is there any monitoring that will need to be undertaken during the study?

No

For the risks or harms identified above, is there any monitoring that will need to be undertaken following the study conclusion?

No

Outline the criteria for stopping the study early due to safety concerns/other issues.



Participants of the study are allowed to stop their participation at any time should they find it overwhelming without repercussion.

Privacy

Will demographic and/or background information be asked of participants?

Yes

What demographic/background information will be collected?

Age
Gender
Education
Other

Describe what “other” demographic/background information will be collected.

Field of Study: What is/are your academic field(s) of study in which you have obtained a degree? We have chosen this demographic field over Occupation as we expect most, if not all, of our participants to be students.

Will demographic/background information be collected separately from names and other identifying information?

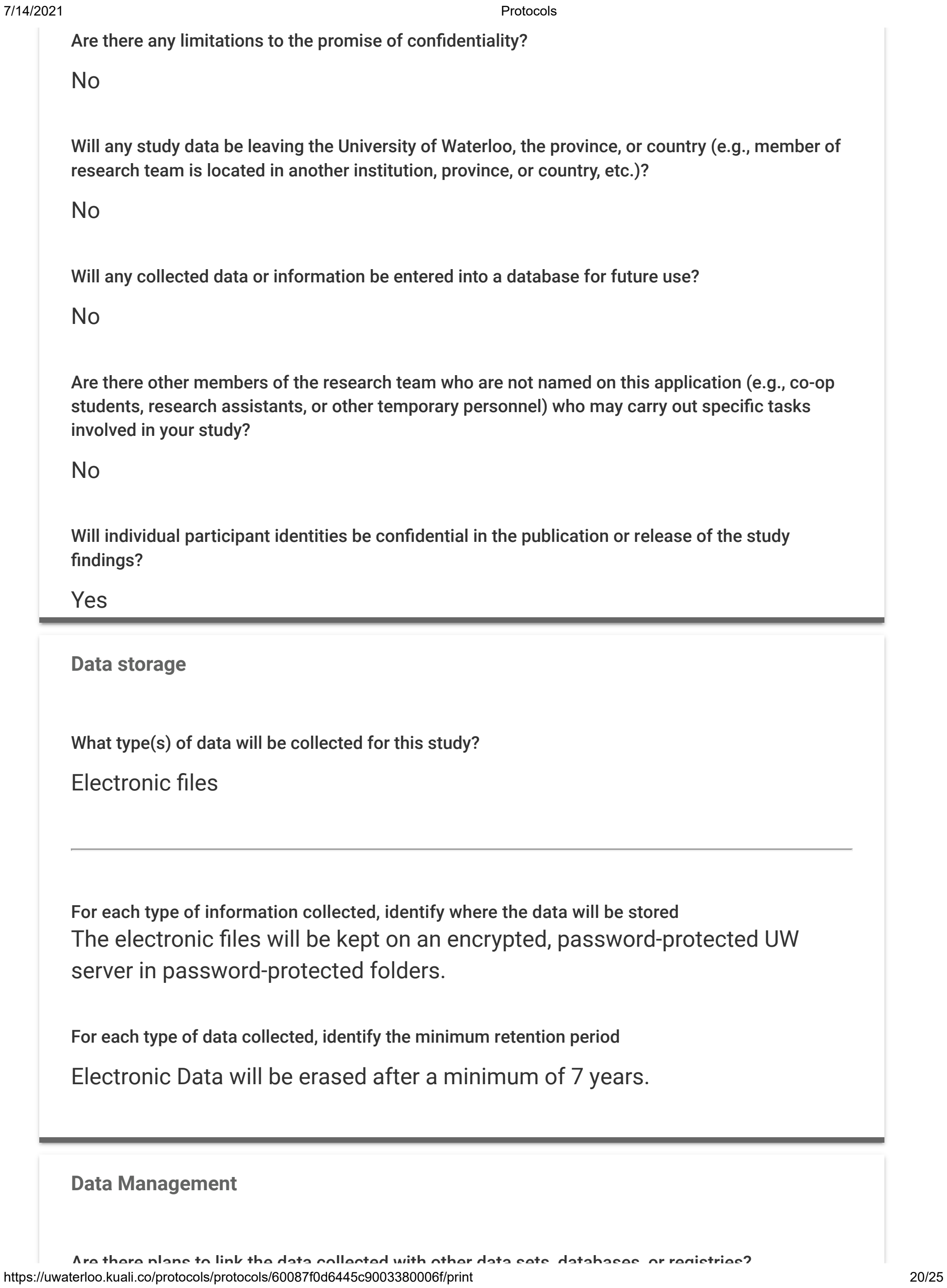
Yes

Participant identification

Participant number: a unique code will be generated for each participant.

If applicable how will the key/list that links participants’ codes with their actual name and/or consent forms be stored and protected?

Key/lists that link participants' codes with their actual names and/or consent forms will be stored on an encrypted, password-protected UW server in individual password-protected folders. The mapping from a participants' name to the ID will be maintained for the length of the study in case the participant forgets their ID. This mapping will be destroyed at the completion of the study. Email address collected for remuneration is stored separately from the research data, and is destroyed once no longer needed.



Are there any limitations to the promise of confidentiality?

No

Will any study data be leaving the University of Waterloo, the province, or country (e.g., member of research team is located in another institution, province, or country, etc.)?

No

Will any collected data or information be entered into a database for future use?

No

Are there other members of the research team who are not named on this application (e.g., co-op students, research assistants, or other temporary personnel) who may carry out specific tasks involved in your study?

No

Will individual participant identities be confidential in the publication or release of the study findings?

Yes

Data storage

What type(s) of data will be collected for this study?

Electronic files

For each type of information collected, identify where the data will be stored

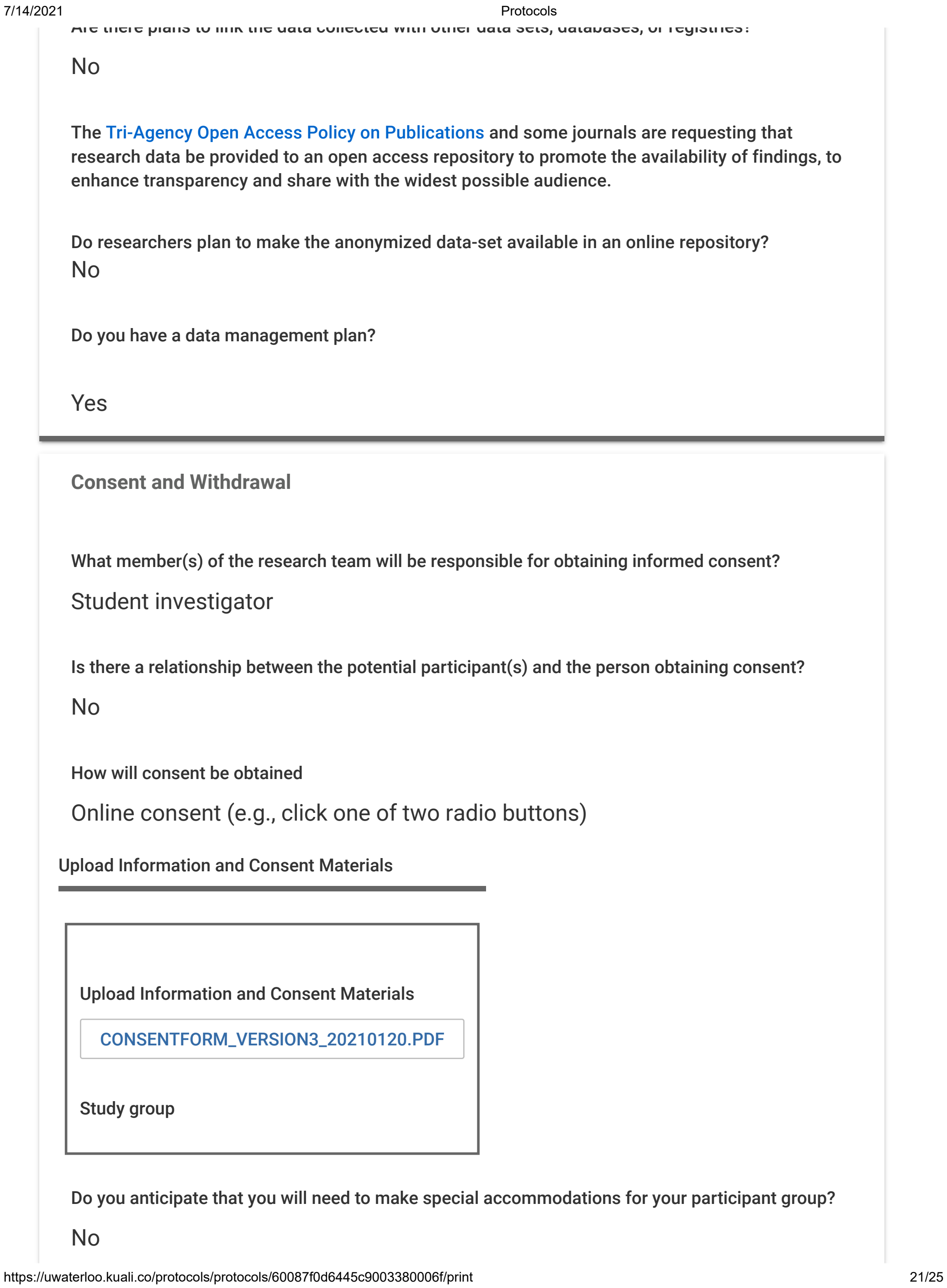
The electronic files will be kept on an encrypted, password-protected UW server in password-protected folders.

For each type of data collected, identify the minimum retention period

Electronic Data will be erased after a minimum of 7 years.

Data Management

Are there plans to link the data collected with other data sets, databases, or registries?



7/14/2021 Protocols

Are there plans to link the data collected with other data sets, databases, or registries?

No

The Tri-Agency Open Access Policy on Publications and some journals are requesting that research data be provided to an open access repository to promote the availability of findings, to enhance transparency and share with the widest possible audience.

Do researchers plan to make the anonymized data-set available in an online repository?

No

Do you have a data management plan?

Yes

Consent and Withdrawal

What member(s) of the research team will be responsible for obtaining informed consent?

Student investigator

Is there a relationship between the potential participant(s) and the person obtaining consent?

No

How will consent be obtained

Online consent (e.g., click one of two radio buttons)

Upload Information and Consent Materials

Upload Information and Consent Materials

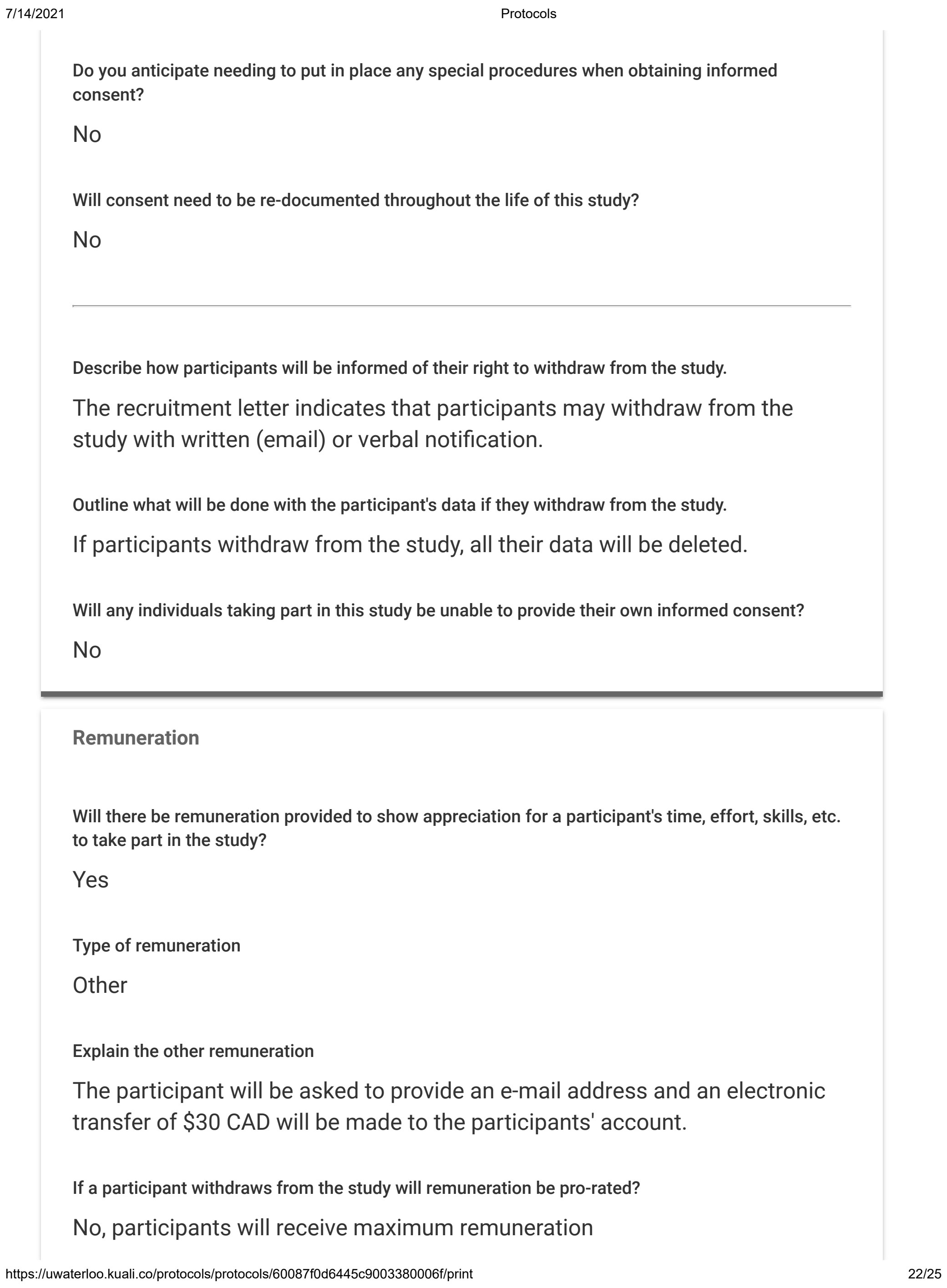
CONSENTFORM_VERSION3_20210120.PDF

Study group

Do you anticipate that you will need to make special accommodations for your participant group?

No

https://uwaterloo.kuali.co/protocols/protocols/60087f0d6445c9003380006f/print 21/25



Do you anticipate needing to put in place any special procedures when obtaining informed consent?

No

Will consent need to be re-documented throughout the life of this study?

No

Describe how participants will be informed of their right to withdraw from the study.

The recruitment letter indicates that participants may withdraw from the study with written (email) or verbal notification.

Outline what will be done with the participant's data if they withdraw from the study.

If participants withdraw from the study, all their data will be deleted.

Will any individuals taking part in this study be unable to provide their own informed consent?

No

Remuneration

Will there be remuneration provided to show appreciation for a participant's time, effort, skills, etc. to take part in the study?

Yes

Type of remuneration

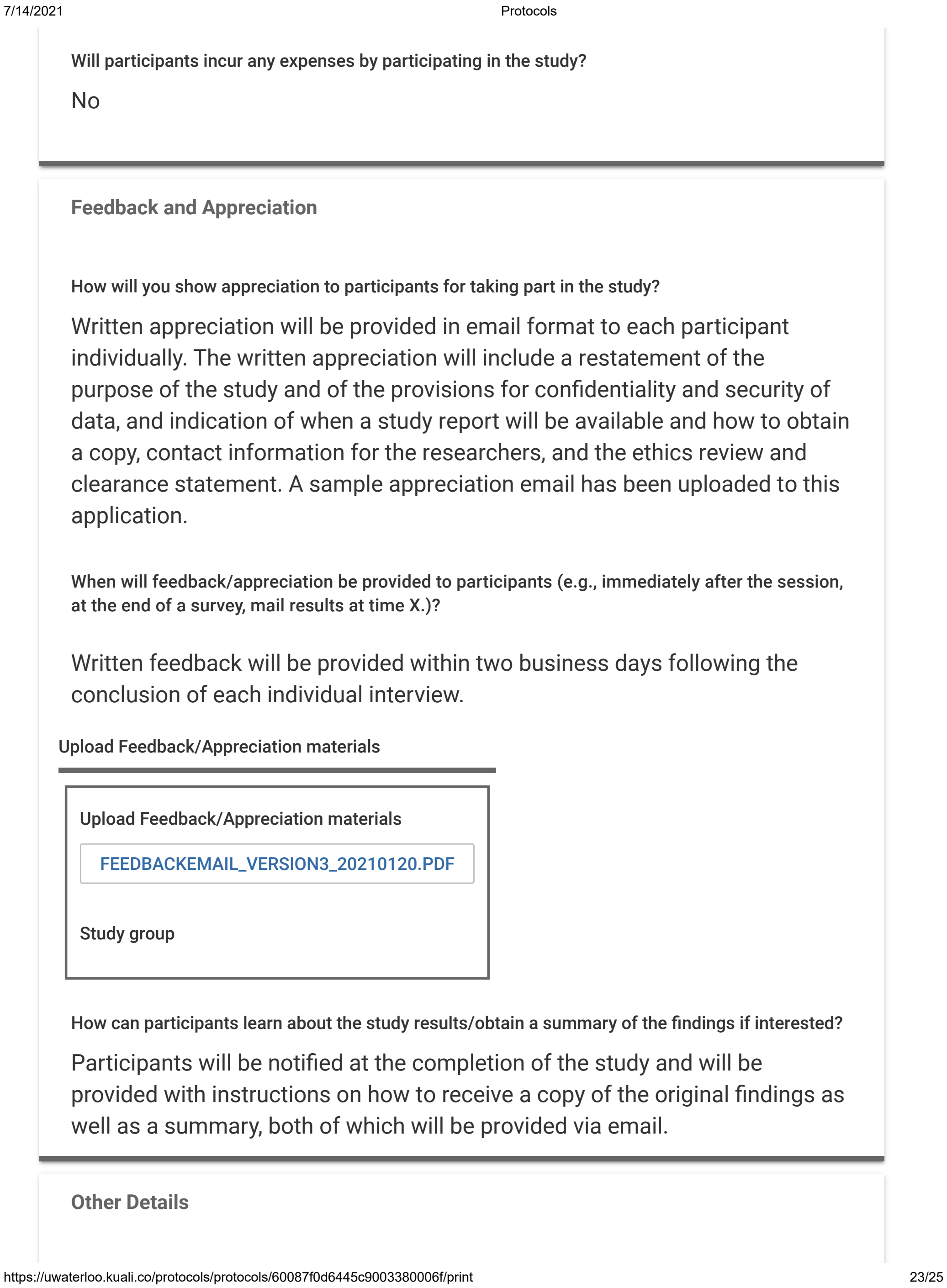
Other

Explain the other remuneration

The participant will be asked to provide an e-mail address and an electronic transfer of $30 CAD will be made to the participants' account.

If a participant withdraws from the study will remuneration be pro-rated?

No, participants will receive maximum remuneration



Will participants incur any expenses by participating in the study?

No

Feedback and Appreciation

How will you show appreciation to participants for taking part in the study?

Written appreciation will be provided in email format to each participant individually. The written appreciation will include a restatement of the purpose of the study and of the provisions for confidentiality and security of data, and indication of when a study report will be available and how to obtain a copy, contact information for the researchers, and the ethics review and clearance statement. A sample appreciation email has been uploaded to this application.

When will feedback/appreciation be provided to participants (e.g., immediately after the session, at the end of a survey, mail results at time X.)?

Written feedback will be provided within two business days following the conclusion of each individual interview.

Upload Feedback/Appreciation materials

Upload Feedback/Appreciation materials

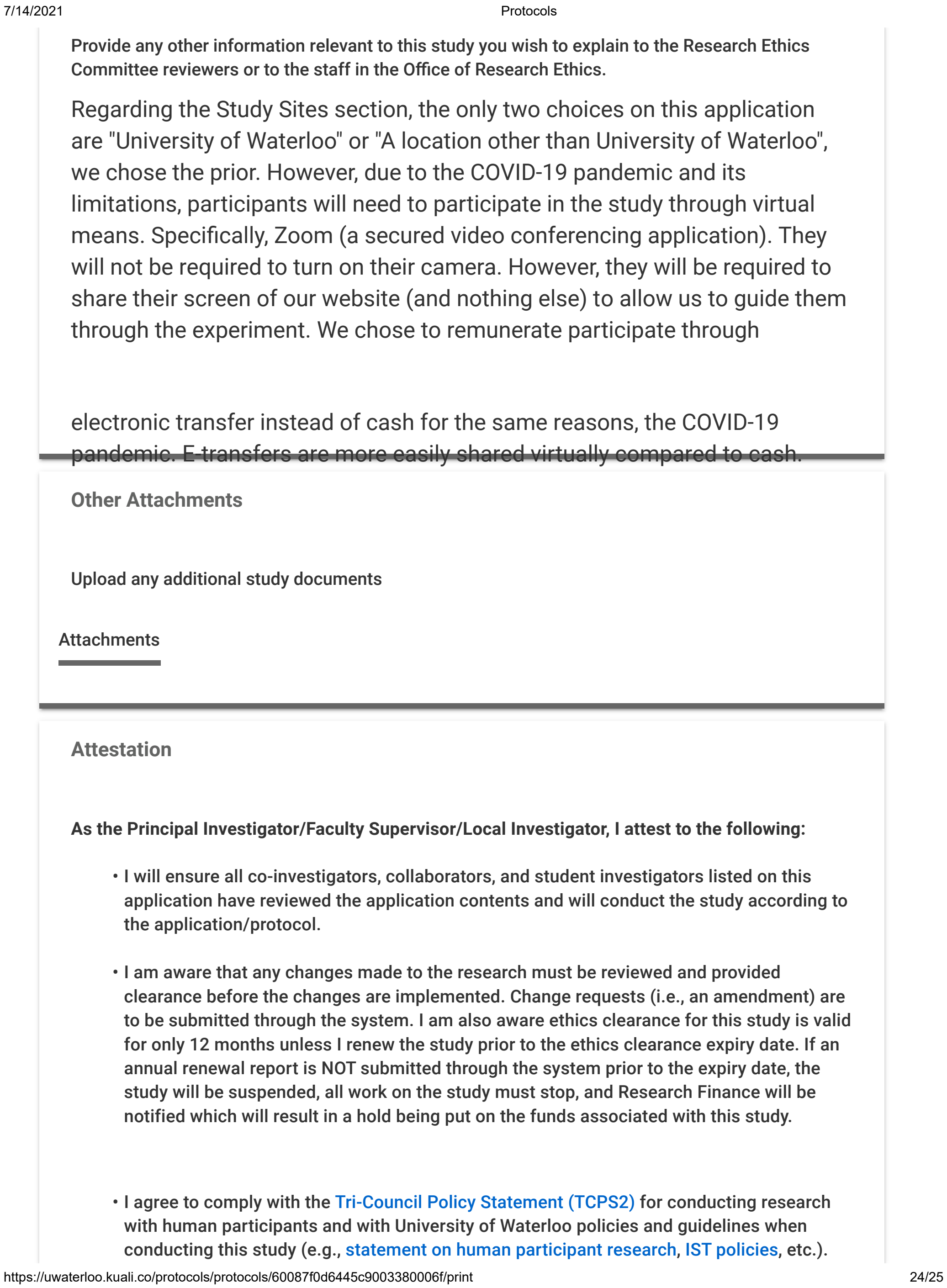
FEEDBACKEMAIL_VERSION3_20210120.PDF

Study group

How can participants learn about the study results/obtain a summary of the findings if interested?

Participants will be notified at the completion of the study and will be provided with instructions on how to receive a copy of the original findings as well as a summary, both of which will be provided via email.

Other Details



Provide any other information relevant to this study you wish to explain to the Research Ethics Committee reviewers or to the staff in the Office of Research Ethics.

Regarding the Study Sites section, the only two choices on this application are "University of Waterloo" or "A location other than University of Waterloo", we chose the prior. However, due to the COVID-19 pandemic and its limitations, participants will need to participate in the study through virtual means. Specifically, Zoom (a secured video conferencing application). They will not be required to turn on their camera. However, they will be required to share their screen of our website (and nothing else) to allow us to guide them through the experiment. We chose to remunerate participate through electronic transfer instead of cash for the same reasons, the COVID-19 pandemic. E-transfers are more easily shared virtually compared to cash.

Other Attachments

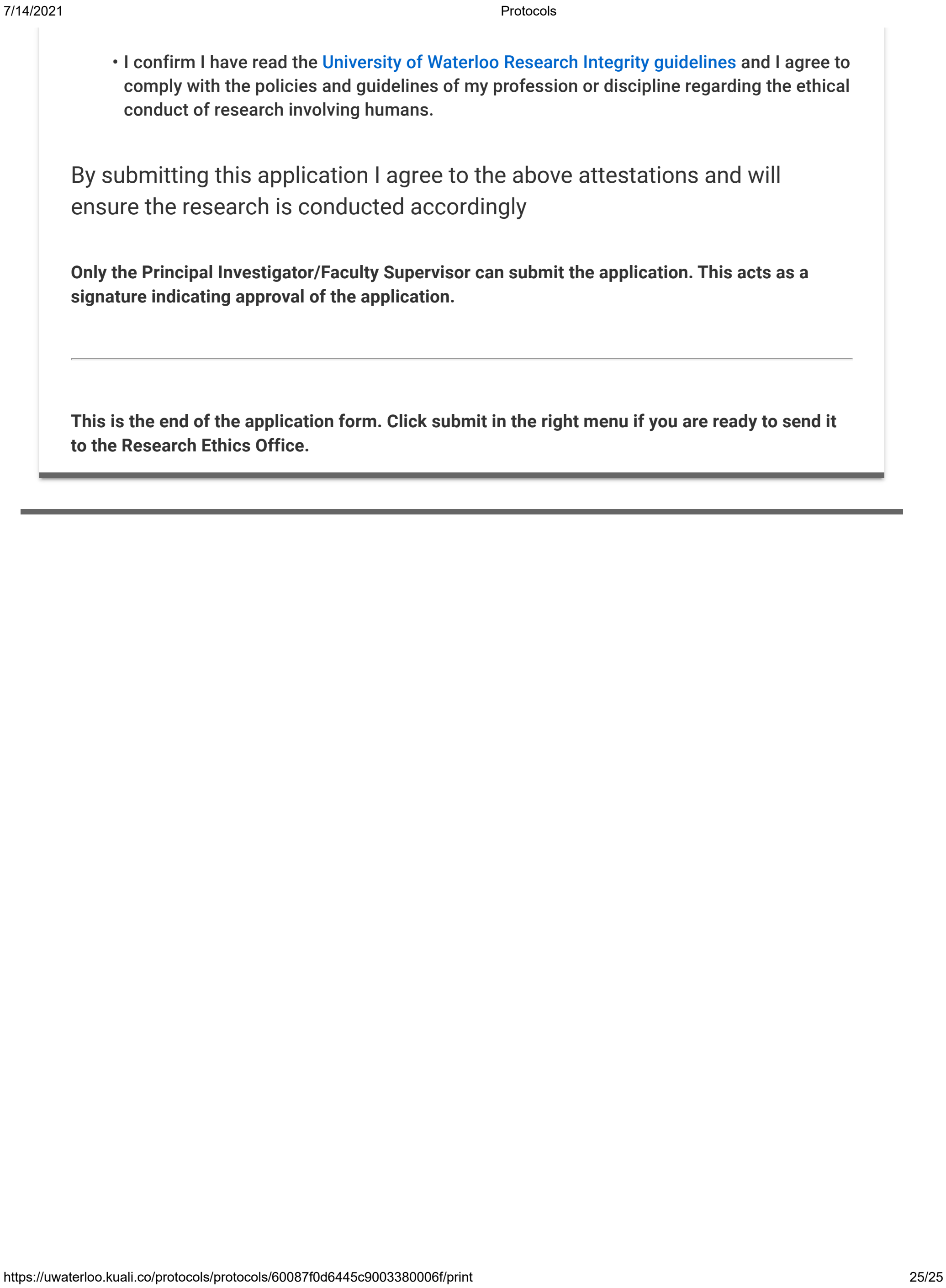
Upload any additional study documents

Attachments

Attestation

As the Principal Investigator/Faculty Supervisor/Local Investigator, I attest to the following:

- I will ensure all co-investigators, collaborators, and student investigators listed on this application have reviewed the application contents and will conduct the study according to the application/protocol.
- I am aware that any changes made to the research must be reviewed and provided clearance before the changes are implemented. Change requests (i.e., an amendment) are to be submitted through the system. I am also aware ethics clearance for this study is valid for only 12 months unless I renew the study prior to the ethics clearance expiry date. If an annual renewal report is NOT submitted through the system prior to the expiry date, the study will be suspended, all work on the study must stop, and Research Finance will be notified which will result in a hold being put on the funds associated with this study.
- I agree to comply with the Tri-Council Policy Statement (TCPS2) for conducting research with human participants and with University of Waterloo policies and guidelines when conducting this study (e.g., statement on human participant research, IST policies, etc.).



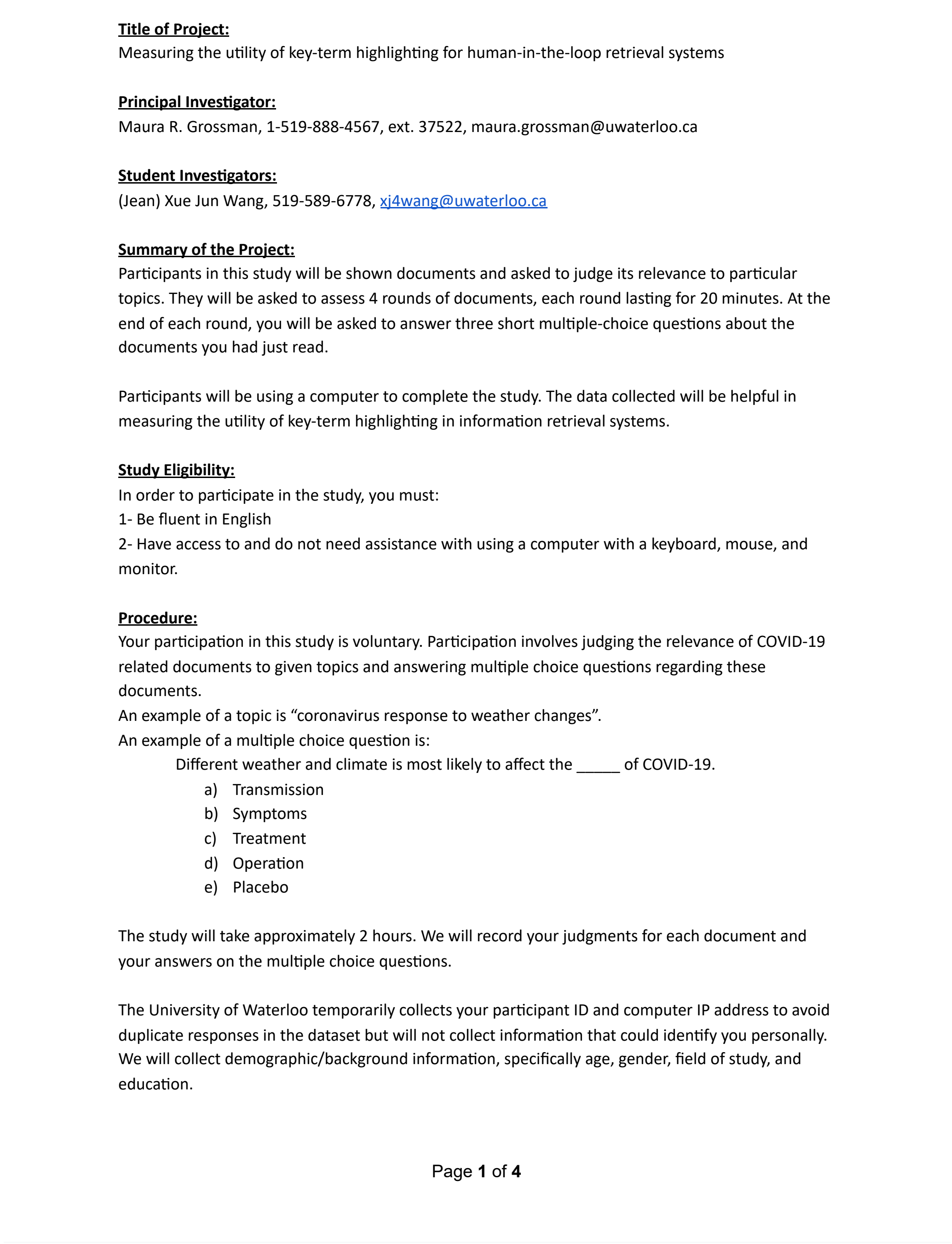
- I confirm I have read the University of Waterloo Research Integrity guidelines and I agree to comply with the policies and guidelines of my profession or discipline regarding the ethical conduct of research involving humans.

By submitting this application I agree to the above attestations and will ensure the research is conducted accordingly

Only the Principal Investigator/Faculty Supervisor can submit the application. This acts as a signature indicating approval of the application.

This is the end of the application form. Click submit in the right menu if you are ready to send it to the Research Ethics Office.

D.3 Consent Form



Title of Project:
Measuring the utility of key-term highlighting for human-in-the-loop retrieval systems

Principal Investigator:
Maura R. Grossman, 1-519-888-4567, ext. 37522, maura.grossman@uwaterloo.ca

Student Investigators:
(Jean) Xue Jun Wang, 519-589-6778, xj4wang@uwaterloo.ca

Summary of the Project:
Participants in this study will be shown documents and asked to judge its relevance to particular topics. They will be asked to assess 4 rounds of documents, each round lasting for 20 minutes. At the end of each round, you will be asked to answer three short multiple-choice questions about the documents you had just read.

Participants will be using a computer to complete the study. The data collected will be helpful in measuring the utility of key-term highlighting in information retrieval systems.

Study Eligibility:
In order to participate in the study, you must:
1- Be fluent in English
2- Have access to and do not need assistance with using a computer with a keyboard, mouse, and monitor.

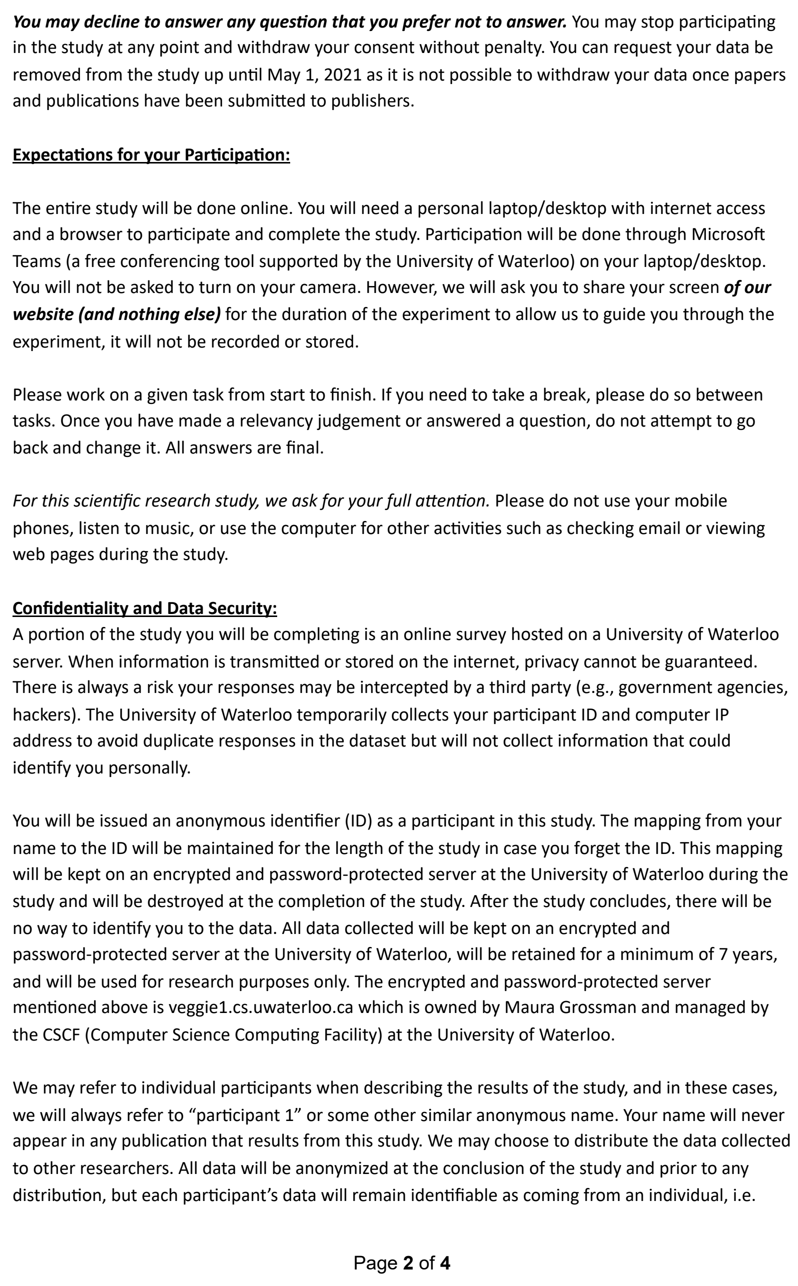
Procedure:
Your participation in this study is voluntary. Participation involves judging the relevance of COVID-19 related documents to given topics and answering multiple choice questions regarding these documents.
An example of a topic is "coronavirus response to weather changes".
An example of a multiple choice question is:

Different weather and climate is most likely to affect the _____ of COVID-19.

a) Transmission
b) Symptoms
c) Treatment
d) Operation
e) Placebo

The study will take approximately 2 hours. We will record your judgments for each document and your answers on the multiple choice questions.

The University of Waterloo temporarily collects your participant ID and computer IP address to avoid duplicate responses in the dataset but will not collect information that could identify you personally. We will collect demographic/background information, specifically age, gender, field of study, and education.



You may decline to answer any question that you prefer not to answer. You may stop participating in the study at any point and withdraw your consent without penalty. You can request your data be removed from the study up until May 1, 2021 as it is not possible to withdraw your data once papers and publications have been submitted to publishers.

Expectations for your Participation:

The entire study will be done online. You will need a personal laptop/desktop with internet access and a browser to participate and complete the study. Participation will be done through Microsoft Teams (a free conferencing tool supported by the University of Waterloo) on your laptop/desktop. You will not be asked to turn on your camera. However, we will ask you to share your screen ***of our website (and nothing else)*** for the duration of the experiment to allow us to guide you through the experiment, it will not be recorded or stored.

Please work on a given task from start to finish. If you need to take a break, please do so between tasks. Once you have made a relevancy judgement or answered a question, do not attempt to go back and change it. All answers are final.

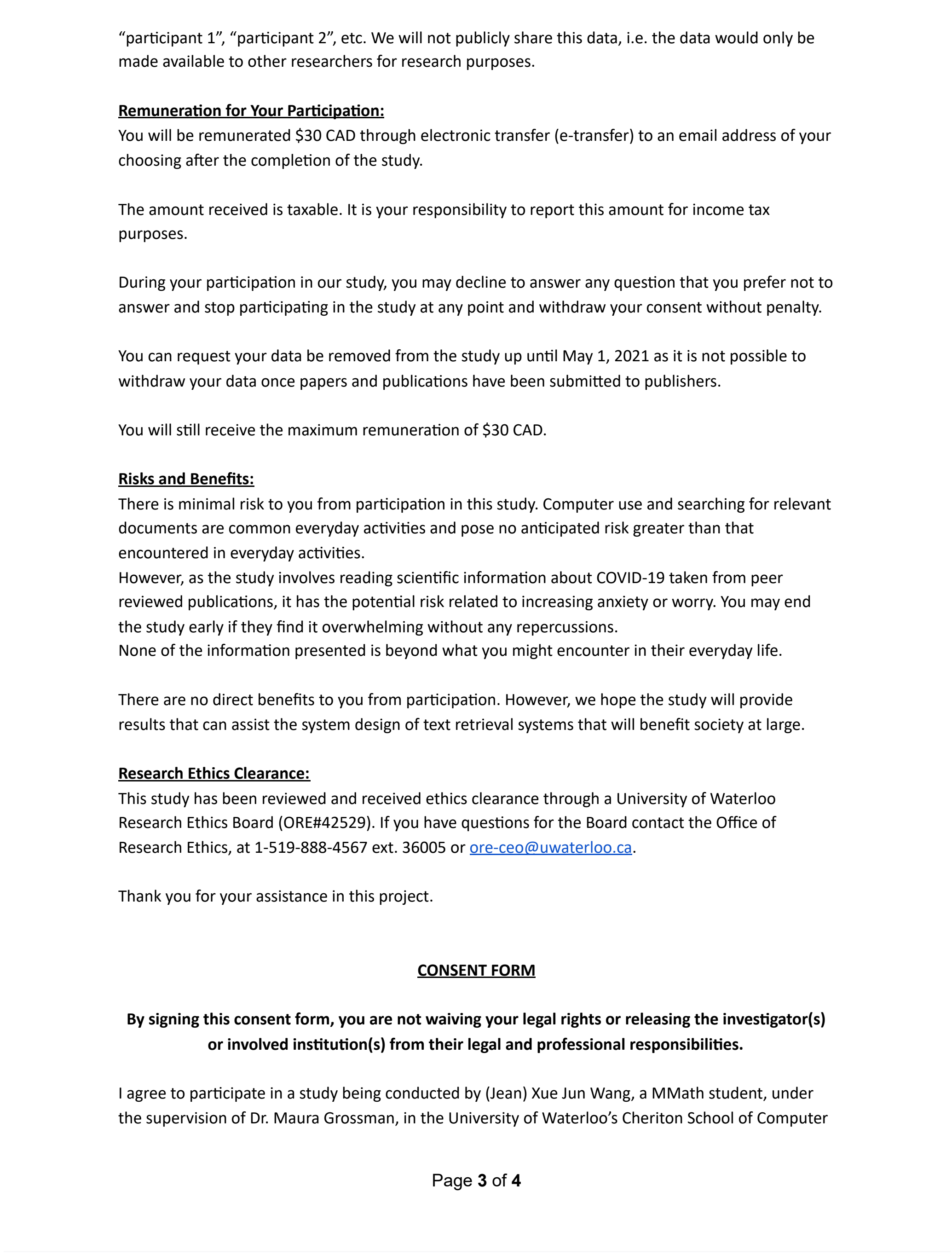
For this scientific research study, we ask for your full attention. Please do not use your mobile phones, listen to music, or use the computer for other activities such as checking email or viewing web pages during the study.

Confidentiality and Data Security:

A portion of the study you will be completing is an online survey hosted on a University of Waterloo server. When information is transmitted or stored on the internet, privacy cannot be guaranteed. There is always a risk your responses may be intercepted by a third party (e.g., government agencies, hackers). The University of Waterloo temporarily collects your participant ID and computer IP address to avoid duplicate responses in the dataset but will not collect information that could identify you personally.

You will be issued an anonymous identifier (ID) as a participant in this study. The mapping from your name to the ID will be maintained for the length of the study in case you forget the ID. This mapping will be kept on an encrypted and password-protected server at the University of Waterloo during the study and will be destroyed at the completion of the study. After the study concludes, there will be no way to identify you to the data. All data collected will be kept on an encrypted and password-protected server at the University of Waterloo, will be retained for a minimum of 7 years, and will be used for research purposes only. The encrypted and password-protected server mentioned above is veggie1.cs.uwaterloo.ca which is owned by Maura Grossman and managed by the CSCF (Computer Science Computing Facility) at the University of Waterloo.

We may refer to individual participants when describing the results of the study, and in these cases, we will always refer to “participant 1” or some other similar anonymous name. Your name will never appear in any publication that results from this study. We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participant’s data will remain identifiable as coming from an individual, i.e.



“participant 1”, “participant 2”, etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

Remuneration for Your Participation:
You will be remunerated $30 CAD through electronic transfer (e-transfer) to an email address of your choosing after the completion of the study.

The amount received is taxable. It is your responsibility to report this amount for income tax purposes.

During your participation in our study, you may decline to answer any question that you prefer not to answer and stop participating in the study at any point and withdraw your consent without penalty.

You can request your data be removed from the study up until May 1, 2021 as it is not possible to withdraw your data once papers and publications have been submitted to publishers.

You will still receive the maximum remuneration of $30 CAD.

Risks and Benefits:
There is minimal risk to you from participation in this study. Computer use and searching for relevant documents are common everyday activities and pose no anticipated risk greater than that encountered in everyday activities.
However, as the study involves reading scientific information about COVID-19 taken from peer reviewed publications, it has the potential risk related to increasing anxiety or worry. You may end the study early if they find it overwhelming without any repercussions.
None of the information presented is beyond what you might encounter in their everyday life.

There are no direct benefits to you from participation. However, we hope the study will provide results that can assist the system design of text retrieval systems that will benefit society at large.

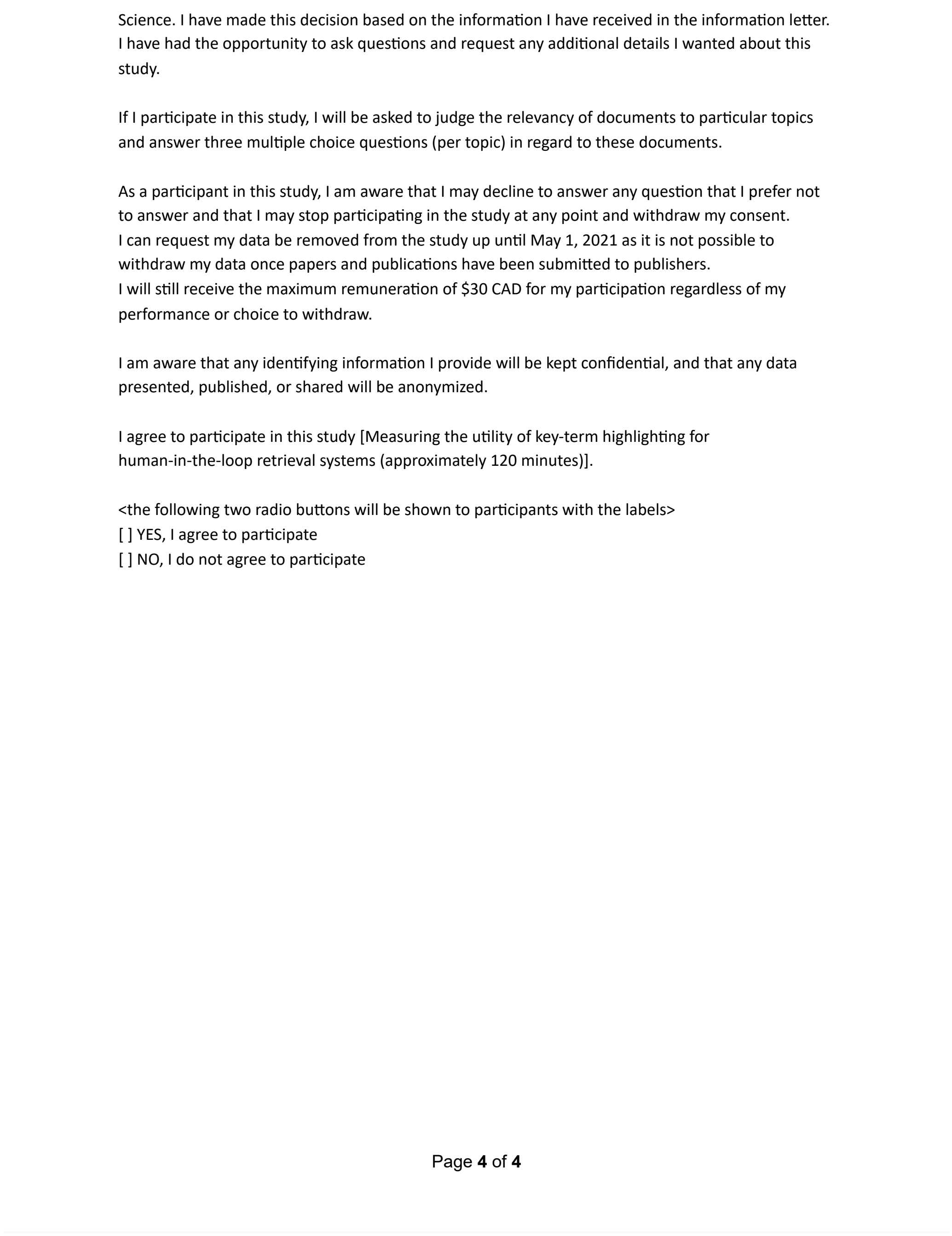
Research Ethics Clearance:
This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (ORE#42529). If you have questions for the Board contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

Thank you for your assistance in this project.

CONSENT FORM

By signing this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

I agree to participate in a study being conducted by (Jean) Xue Jun Wang, a MMath student, under the supervision of Dr. Maura Grossman, in the University of Waterloo’s Cheriton School of Computer



Science. I have made this decision based on the information I have received in the information letter. I have had the opportunity to ask questions and request any additional details I wanted about this study.

If I participate in this study, I will be asked to judge the relevancy of documents to particular topics and answer three multiple choice questions (per topic) in regard to these documents.

As a participant in this study, I am aware that I may decline to answer any question that I prefer not to answer and that I may stop participating in the study at any point and withdraw my consent.
I can request my data be removed from the study up until May 1, 2021 as it is not possible to withdraw my data once papers and publications have been submitted to publishers.
I will still receive the maximum remuneration of $30 CAD for my participation regardless of my performance or choice to withdraw.

I am aware that any identifying information I provide will be kept confidential, and that any data presented, published, or shared will be anonymized.

I agree to participate in this study [Measuring the utility of key-term highlighting for human-in-the-loop retrieval systems (approximately 120 minutes)].

<the following two radio buttons will be shown to participants with the labels>
[ ] YES, I agree to participate
[ ] NO, I do not agree to participate