TU Delft

Delft University of Technology

Licence
CC BY

Citation (APA)
Roy, N., Balayn, A., Maxwell, D., & Hauff, C. (2023). Hear Me Out: A Study on the Use of the Voice Modality for Crowdsourced Relevance Assessments. In *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 718-728). (SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval). Association for Computing Machinery (ACM). https://doi.org/10.1145/3539618.3591694

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Hear Me Out: A Study on the Use of the Voice Modality for Crowdsourced Relevance Assessments

Nirmal Roy
Delft University of Technology
The Netherlands
n.roy@tudelft.nl

Agathe Balayn
Delft University of Technology
The Netherlands
a.m.a.balayn@tudelft.nl

David Maxwell
Delft University of Technology
The Netherlands
maxwelld90@acm.org

Claudia Hauff
Spotify & Delft University of Technology
The Netherlands
c.hauff@tudelft.nl

ABSTRACT

The creation of relevance assessments by human assessors (often nowadays crowdworkers) is a vital step when building IR test collections. Prior works have investigated assessor quality & behaviour, and tooling to support assessors in their task. We have few insights though into the impact of a document's *presentation modality* on assessor efficiency and effectiveness. Given the rise of voice-based interfaces, we investigate whether it is feasible for assessors to judge the relevance of text documents via a voice-based interface. We ran a user study (n = 49) on a crowdsourcing platform where participants judged the relevance of short and long documents—sampled from the *TREC Deep Learning* corpus—presented to them either in the `text` or `voice` modality. We found that: *(i)* participants are *equally* accurate in their judgements across both the `text` and `voice` modality; *(ii)* with increased document length it takes participants significantly longer (for documents of length > 120 words it takes almost twice as much time) to make relevance judgements in the `voice` condition; and *(iii)* the ability of assessors to ignore stimuli that are not relevant (i.e., *inhibition*) impacts the assessment quality in the `voice` modality—assessors with higher inhibition are significantly more accurate than those with lower inhibition. Our results indicate that we can reliably leverage the `voice` modality as a means to effectively collect relevance labels from crowdworkers.

CCS CONCEPTS

• **Information systems** → **Relevance assessment**; **Presentation of retrieval results**; *Answer ranking*; *Question answering*.

KEYWORDS

Relevance Assessment; Cognitive Ability; Crowdsourcing; Data Annotation; User Interfaces

This research has been supported by *NWO VIDI* project *SearchX* (639.022.722) and *NWO* project *Aspasia* (015.013.027).

This work is licensed under a Creative Commons Attribution International 4.0 License.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591694

ACM Reference Format:
Nirmal Roy, Agathe Balayn, David Maxwell, and Claudia Hauff. 2023. Hear Me Out: A Study on the Use of the Voice Modality for Crowdsourced Relevance Assessments. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591694

1 INTRODUCTION

Document relevance assessments by human assessors—with respect to a given set of *information needs*—is a vital step in the building of an *Information Retrieval (IR)* test collection [33, 67]. Depending on the corpus, documents are represented in a variety of forms—including text (the most common form at *TREC*), images [19, 48], or videos [25, 42]. Prior works have investigated assessor quality, their behaviour, and tooling to support assessors—most often in the context of text documents [6, 38, 56, 62, 63]. Given the prevalent nature of text corpora, we continue in this vein and focus on an aspect that has received little attention so far: the *presentation modality* of the text documents during the judging process.

Thanks to the development of voice-based conversational search systems, people have become accustomed to being presented search results that are read out to them, an approach that is very different from the presentation of text on-screen. We posit that by utilising such audio-based devices, we can increase the scope for collecting relevance judgements for text documents in a number of ways. For example, assessors can contribute by judging documents on their smartphones [3, 77], if they have visual impairments [55, 79, 86], or if they come from a low-resource background [5, 55].

Two important aspects of collecting relevance judgements are: *(i)* the quality of assessments [62]; and *(ii)* the time taken by assessors to make their judgements [69]. Since relevance judgements are used to train and evaluate *Learning to Rank (LtR)* systems, the quality of judgements impacts the effectiveness of such systems [15, 83]. The time taken by assessors to judge relevance may not only affect the quality of judgements, but also contribute to the cost of building (and maintaining) test collections. NIST assessors [16, 17] and crowdworkers [4, 39] are often paid by their time spent on a task (e.g., as on *Prolific*). The longer it takes assessors to judge, the costlier it becomes. There are a number of factors—not limited to topic difficulty [18, 62], document familiarity [63], or relevance judgement session length [63]—that have been shown to affect the quality of (and the time taken for) judging relevance.

In our work, we focus on two such factors in our pursuit to examine the feasibility of using the voice modality for text-document relevance assessments: *document length* [23, 59, 63, 68] and an assessor's *cognitive abilities* [60, 62] expressed in terms of working memory and inhibition. Our selection of factors is motivated by a range of prior works. The serial [40] and temporal [61] nature of the voice medium makes it more difficult for listeners to *"skim"* back and forth over a piece of information as compared to reading it on-screen [49, 82, 84]. Voice interfaces also demand greater cognitive load when compared to text interfaces for processing information [40, 54, 66]. These are exacerbated as the amount of information to be conveyed increases in size [51, 64]. Understanding how these factors affect the relevance judgement process can help us design tasks for assessors with a wide range of abilities and for different document presentation modalities. While there exists various measures for cognitive abilities, we selected two—*working memory* (someone's ability to hold information in short-term memory) [21] and *inhibition* (someone's ability to ignore or inhibit attention to stimuli that are not relevant) [21]—which have been shown to play an important role in speech understanding [26, 58, 71]. We posit that they will also be crucial in the relevance judgement process, especially when documents are presented in the voice modality. Taken together, we investigate the following research questions.

RQ1 *How does the modality of document presentation (text vs. voice) affect an assessor's relevance judgement in terms of accuracy, time taken, and perceived workload?*

RQ2 *How does the length of documents affect assessors' ability to judge relevance?* Specifically, we look into the main effect of document length and the effect of its interplay with presentation modality.

RQ3 *How do the cognitive abilities of an assessor (with respect to their working memory and inhibition) affect their ability to judge relevance?* Specifically, we look into the main effect of the cognitive abilities and the effect of their interplay with the presentation modality.

To answer these questions, we conducted a quantitative user study ($n = 49$) on the crowdsourcing platform Prolific. Participants judged the relevance of 40 short and long documents sampled from the passage retrieval task data of the 2019 & 2020 *TREC Deep Learning (DL) track* [16, 17]. Our findings are summarised as follows.

- Participants judging documents presented in the voice modality were *equally* accurate as those judging them in the text modality.
- As documents got longer, participants judging documents in voice modality took significantly longer than those in text modality. For documents of length greater than 120 words, the former took twice as much time with less reliable judgements.
- We also found that inhibition—or a participant's individual ability to ignore or inhibit attention to stimuli that are not relevant—impacts relevance judgements in voice modality. Indeed, those with higher inhibition were significantly more accurate than their lower inhibition counterparts.

Overall, our results indicate that we *can* leverage the voice modality to effectively collect relevance labels from crowdworkers.

2 RELATED WORK

2.1 Relevance Judgement Collection

The general approach for gathering relevance assessments for large document corpora (large enough that a full judgement of all corpus documents is not possible) was established by TREC in the early 1990s [28]. Given a set of information needs, a pooled set of documents based on the top-k results of (ideally) a wide range of retrieval runs are assessed by topic experts. This method is typically costly and does not scale up [4] once the number of information needs or k increases. In the last decade, creating test collections using crowdsourcing via platforms like Prolific or *Amazon Mechanical Turk (AMT)* have been shown to be a less costly yet reliable alternative [4, 39, 63, 85]. While the potential of crowdsourcing for more efficient relevance assessment has been acknowledged, concerns have been raised regarding its quality—as workers might be too inexperienced, lack the necessary topical expertise, or be paid an insufficient salary. In turn, these issues may lead them to completing the tasks to a low standard [36, 46, 53]. Aggregation methods (e.g., majority voting) can be used as effective countermeasures to improve the reliability of judgements [32, 34].

There are a number of factors that have been shown to affect the relevance judgement process. Scholer et al. [62] observed that participants exposed to non-relevant documents at the start of a judgement session assigned higher overall relevance scores to documents than when compared to those exposed to relevant documents. Damessie et al. [18] found that for easier topics, assessors processed documents more quickly, and spent less time overall. Document length was also shown to be an important factor for judgement reliability. Hagerty [27] found that the precision and recall of abstracts judged increased as the abstract lengths increased (30, 60, and 300 words). In a similar vein, Singhal et al. [68] observed that the likelihood of a document being judged relevant by an assessor increased with the document length. Chandar et al. [12] found that shorter documents that are easier to understand provoked higher disagreement, and that there was a weak relationship between document length and disagreement between the assessor. In terms of time spent for relevance judgement, Konstan et al. [37] and Shinoda [65] asserted that there is no significant correlation between time and document length. On the other hand, Smucker et al. [70] found participants took more time to read, as document length increased (from ~10s for 100 words, to ~25s for 1000 words).

2.2 Voice Modality

Voice-based crowdsourcing has been shown to be more accessible for people with visual impairments [79, 86], or those from low resource backgrounds [55]. It can also provide greater flexibility to crowdworkers by allowing them to work in brief sessions, enabling multitasking, reducing effort required to initiate tasks, and being reliable [31, 78]. However, information processing via voice is inherently different compared to when it is presented as text. The use of voice has been often shown to lead to a higher cognitive load [50, 80]. Individuals also exhibit different preferences. For example, Trippas et al. [75] observed that participants preferred longer summaries for text presentation. For voice however, shortened summaries were preferred when the queries were single-faceted. Although their study did not measure the accuracy of judgements against a ground

truth, what participants considered the most relevant was similar across both conditions (text vs. voice presentation). Furthermore, the voice modality can leverage its own unique characteristics for information presentation. For instance, Chuklin et al. [14] varied the prosody features (pauses, speech rate, pitch) of sentences containing answers to factoid questions. They found that emphasising the answer phrase with a lower speaking rate and higher pitch increased the perceived level of information conveyed.

Concerning the collection of relevance assessments, Tombros and Crestani [74] found in their lab study that participants were more accurate and faster in judging relevance when the list of documents (with respect to a query) were presented as text on screen as compared to when they were read out to the participants—either in person, or via telephone. It should however be noted that this work was conducted more than two decades ago—barely ten years after the invention of the Web, when the now common voice assistants and voice-enabled devices were long to be developed.

The work closest to ours is the study by Vtyurina et al. [80], who presented crowdworkers with five results of different ranks from *Google*—either in text or voice modality. The asked their participants to select the two most useful results and the least useful one. The relevance judgements of participants in the text condition were observed to be significantly more consistent with the true ranking of the results than those who were presented with five audio snippets. The ability to identify the most relevant result was however *not* different between the two cohorts. This study did not consider the effect of document length or cognitive abilities of participants on their relevance judgement performance, which is what we explore.

2.3 Cognitive Abilities

Prior works have explored how the cognitive abilities of assessors impact relevance judgements. Davidson [20] observed that openness to information—measured by a number of cognitive style variables such as open-mindedness, rigidity, and locus of control—accounted for approximately 30% of the variance in relevance assessments. Scholer et al. [62] found that assessors with a higher need for cognition (i.e., a predisposition to enjoy cognitively demanding activities) had higher agreement with *expert* assessors, and took longer to judge compared to their lower need for cognition counterparts. Our work focuses on *working memory* and *inhibition*.

Working Memory (WM) refers to *an individual's capacity for keeping information in short-term memory even when it is no longer perceptually present* [21]. This ability plays a role in higher-level tasks, such as reading comprehension [43] and problem solving [81]. MacFarlane et al. [44] observed that participants with dyslexia—a learning disorder characterised by low working memory—judged fewer text documents as non-relevant when compared to participants without the learning disorder. They posited that it might be cognitively more demanding to identify text documents as non-relevant for the cohort with dyslexia. With regards to processing speech, High **WM** has also been shown to be helpful in adapting to distortion of speech signals caused by background noise [26]. Rudner et al. [58] and Stenbäck [71] observed high **WM** individuals perceived less effort while recognising speech from noise.

Inhibition (IN) refers to the capacity to regulate attention, behaviour, thoughts, and/or emotions by overriding internal impulses

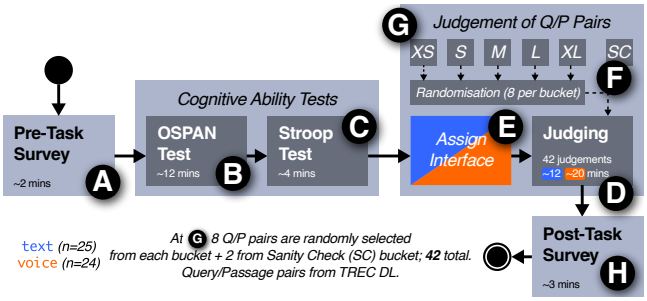

Figure 1: A high-level overview of the user study protocol, including approximate times for participants to complete each component. Refer to §3.1 for mappings to the letters highlighting key aspects of the study procedure.

or external *'lure'*—and maintaining focus on what is appropriate or needed [21]. To our knowledge, prior studies have not investigated the effect of **IN** on the relevance assessment process. High **IN** has been shown to help in speech recognition, especially in adverse conditions like the presence of background noise [71, 72].

A significant number of prior works have explored various aspects related to the process of relevance assessment. This work however considers the novel effect of document length and the cognitive abilities of assessors to explore the utility of the voice modality with regards to judging relevance.

3 METHODOLOGY

To address our three research questions outlined in §1, we conducted a crowdsourced user study. The study participants were asked to judge the relevance of *Query/Passage (Q/P)* pairings, where passages were presented either in the form of `text` (i.e., a piece of text) or `voice` (i.e., an audio clip). In our study, passage **presentation modality** is a *between-subjects* variable. We also controlled the **length of passages**; this is a *within-subjects variable* to ensure that participants judged passages of varying lengths. The *independent variables* **working memory** and **inhibition** allow us to estimate the impact of the cognitive abilities of the participants on the accuracy of their judgements, time taken and perceived workload.

3.1 Study Overview

Figure 1 presents an overview of the user study design.[1] The diagram highlights the main tasks that study participants undertook. Lasting approximately 32 minutes for `text` and 40 minutes for `voice`, the study consisted of four main parts: *(i)* the *pre-task survey* (§3.6); *(ii)* the *cognitive ability tests* (§3.3); *(iii)* the *judgements* (§3.4); and *(iv)* the *post-task survey* (§3.5).

After agreeing to the terms of the study, participants completed a pre-task survey (A). This survey included demographics questions, including questions about their familiarity with voice assistants—as reported in §3.6. Participants would then move onto two *psychometric tests*; as outlined in §3.3, these tests measured their cognitive abilities with respect to working memory (B) and inhibition (C). Participants undertook a short practice task to help them familiarise themselves with the interface for each test.

[1]Note that circles refer to superimposed labels on the illustration in Figure 1.

After the psychometric tests, participants moved to the main part of the study: judging Q/P pairings (D). The experimental system first assigned the participants to either `text` or `voice` randomly (E) (§3.4). Based on the assigned condition, participants then judged a total of 42 Q/P pairings presented to them in a random order to mitigate the effect of topic ordering [62, 63] (§3.2)—40 were selected from the *2019 and 2020 TREC Deep Learning (DL) track*, and the remaining two acted as a *sanity check* (**SC**) (F). The 40 passages belonged to different *answer length buckets* §3.2 (G). Finally, the participants would be taken to the post-task survey (H).

3.2 Query/Passage Pairings

As mentioned, we obtained the Q/P pairings from the 2019 and 2020 TREC DL track—specifically the passage retrieval task [16, 17, 45]. The test partition of the datasets contain 43 and 54 natural language queries with passages that are judged by *NIST* assessors. Using a graded relevance scale, passages for each query were judged by assessors as: *(i) perfectly relevant* when the passage is dedicated to the query, containing an exact answer; *(ii) related* when the passage appears somewhat related to the query, but does not answer it completely; or *(iii) non-relevant*, when the passage has nothing to do with the provided query [16, 17]. We note that an additional relevance category exists (*highly relevant*). However, we ignore judgements of this category in our work (similar to [39]) in order to have a clear distinction between the different categories.

Sampling Procedure. From the available test queries, we sampled 40 (due to budget constraints). As **RQ2** states we are interested in how passage length affects assessments, we next determined five different buckets of passage length: from *very short* to *very long* (more details follow below). We randomly assigned the 40 queries to these five buckets, leading to eight queries per passage length bucket. For each query, we sampled three passages from the QRELs, with the additional condition that the sampled passages must fall into the query's passage length bucket: one *perfectly relevant*, one *related*, and one *non-relevant* passage. And thus, each bucket contains 24 passages pertaining to eight queries. Table 2 demonstrates three Q/P examples, each coming from a different length bucket.

Sanity Check (SC). We also created two additional Q/P pairings to act as a sanity checks[2] in order to perform quality control of the relevance judgements by our participants, as suggested by Scholer et al. [63]. *We did not consider the **SC** Q/P pairs in our data analysis.*

Judgements per Participant. We presented all our participants with the *same* set of 40 queries + 2 **SC** queries in order to mitigate effects arising due to differences in queries [18]. Each participant judged one randomly sampled passage—out of the three available ones—for each of the 40 queries (ignoring the **SC** queries). We thus collected relevance judgements on a total of $40 \times 3 = 120$ Q/P pairs[3]. Each participant judged 13 passages per QREL.

Passage Length Buckets. To add more detail to our passage length bucketing procedure, we chose five types of length buckets:

Table 1: Overview of passage length buckets. Averages are reported together with the standard deviation.

Passage Length	Min-max #words	Avg. #words	Min-max audio clip length (s)	Avg. audio clip length (s)
Very Short	12 – 32	24.67(±5.3)	3 – 13	10.04(±2.4)
Short	33 – 53	41.67(±3.8)	14 – 19	17.04(±1.4)
Medium	54 – 74	63.17(±5.6)	20 – 30	25.17(±3.4)
Long	90 – 120	99.79(±6.9)	31 – 42	36.04(±3.04)
Very Long	121 – 151	139.96(±8.2)	48 – 70	54.58(±4.9)

XS *(Very Short)*; **S** *(Small)*; **M** *(Medium)*; **L** *(Long)*; and **XL** *(Very Long)*. They corresponded to the $0-5$, $5-50$, $50-75$, $75-99$ and $99+$%-ile of the lengths of all judged passages of the 97 test queries in our TREC-DL datasets. We selected the percentiles to have a range of 20 to 30 words per passage length bucket. The concrete word ranges for each passage length bucket can be found in Table 1.

From Text Passage to Audio Clip. We processed the passages to remove any unwanted punctuation, leading and trailing whitespace, and corrected a few spelling errors. These cleaning steps were necessary as we did not want the participants to be distracted by unclean text, and to create legible audio clips for the `voice` interface. We used *Amazon Polly*[4]—an open-source text to speech system with an array of options for language and voice types—to generate the audio clips for the voice results. Specifically, we chose `Matthew`, a male US English voice, with a speed of 95% as the authors unanimously agreed that this particular setting (among other evaluated voice options) had the clearest pronunciation, in particular of difficult words[5] that might appear in the passages. Lastly, we ran a pilot study ($n = 5$) where participants were asked to rate the pace, accent, and length of our generated audio clips on a seven-point scale. They reported an average score of 6.3, confirming the high quality of the audio clips for our task. Table 1 shows the minimum, maximum, and average length of the audio clips in seconds for the passages belonging to the five length buckets.[6]

3.3 Cognitive Ability Tests

In order to measure the cognitive abilities of our participants with relation to judging the presented Q/P pairings, we chose two established psychometric tests that examine both an individual's working memory and their inhibition. Prior work [26, 58, 71] has shown that working memory and inhibition play an important role in speech understanding.

Working Memory. To measure working memory capacity, we used the *Operation-word-SPAN (OSPAN)* test [76] that has also been used in prior *Interactive IR (IIR)* work [13]. The OSPAN test measures an individual's ability to recall letters displayed in sequence, while concurrently completing simple secondary tasks. Participants completed eight trials of varying lengths. During each trial, participants were shown a sequence of 3 – 7 letters, and were then asked to recall the letters in their original order from a grid display.

[2] The sanity check questions were: *(i) Who was the lead vocalist of Queen?*, with the answer passage being perfectly relevant; and *(ii) What is the difference between powerlifting and weightlifting?*, with the answer passage being non-relevant.
[3] The list of collected Q/P pairs are available here.

[4] https://aws.amazon.com/polly/
[5] Difficult words in this context include words from languages other than English (e.g., *"..and include Gruyère, Emmental, Tête De Moine, Sbrinz.."*), words specific to a domain (e.g., *"...the manubrium, sternebrae, and xiphoid cartilage."*), etc.
[6] Audio clips for all the passages are released here.

Table 2: Examples of *Query/Passage (Q/P)* pairs for different passage length categories. The (Qid) is taken from the TREC datasets. We also provide links to [audio] clips of the respective passages.

Passage Length	Query (Qid)	Ground Truth Relevance & Passage
Very Short (XS)	*What metal are hip replacements made of?* (877809)	RELEVANT Some prosthesis, like hip and knee joints made of cobalt chrome, contain some trace of nickel and for patients with allergies to this may have to go with Titanium joints. [Audio]
Short (S)	*Who has the highest career passer rating in the nfl?* (1056416)	SOMEWHAT-RELEVANT Wilson is the only quarterback in NFL history to post a 100-plus passer rating in each of his first two seasons, and he's already won a Super Bowl. Dan Marino is really the only quarterback you could argue was better out of the gate. [Audio]
Long (L)	*What is the appearance of granulation tissue?* (1133579)	NON-RELEVANT The protective outer layer of the plant. Everything needs skin, or at least some sort of a covering, for plants, it's a system of dermal tissue. Which covers the outside of a plant and it protects the plant in a variety of ways. Dermal tissue called epidermis is made up of live parenchyma cells in the non-woody parts of plants. Epidermal cells can secrete a wax-coated substance on leaves and stems, which becomes the cuticle. Dermal tissue that is made up of dead parenchyma cells is what makes up the outer bark in woody plants. [Audio]

Additionally, during each trial, participants completed simple mathematical problems between each letter shown in sequence (e.g., *"is 8+6=15?"*). The final score was equal to the sum of sequence lengths of all trials perfectly recalled. A higher score in the OSPAN test indicates a participant's greater ability to hold information (the letter sequence in correct order) in short-term memory when it is no longer perceptually present.

Inhibition. To measure inhibition, we used the *Stroop test* which was first introduced in 1935 [73]. As an example, the Stroop test has been used to measure inhibitory attention control in learning [24, 35] and speech processing [71]. We used a computerised version of the test that was also used in the IIR study undertaken by Arguello and Choi [7]. During the Stroop test, participants were shown a sequence of words indicating one of four colours: red, green, yellow, or blue. Some of the words displayed are congruent (e.g., the word "blue" displayed in blue font), and others are incongruent (e.g., the word "blue" displayed in red font). For each word, participants had to indicate the *font colour* of the word as quickly as possible by clicking on the correct option presented as a list (the trial continued until the correct colour was chosen). Participants had to complete 48 correct trials (similar to the study by Arguello and Choi [7]), of which 24 are congruent and 24 are incongruent. The final score is equal to the participant's average response time (in milliseconds) for the incongruent trials, minus the average response time for the congruent trials. Response times are typically slower for the incongruent trials, an effect referred to as the *Stroop effect*. Lower scores are better for the Stroop test, with higher scores indicating a greater difficulty in focusing on the relevant stimulus (the colour of the word) and ignoring the non-relevant stimulus (the word itself).

3.4 Assessor Interface

Our study interface is shown in Figure 2, as a composition of both the `text` and `voice` interfaces. The `text`-specific components are highlighted in blue; `voice`-specific ones in orange. For each Q/P pairing they were required to judge, participants were presented with a static query box ❶ which could not be altered; it displayed the query for which the participant was to judge the passage for. Only one passage was shown ❷; depending on the condition, this was either presented as text (for `text`), or a series of buttons to control the audio clip (for `voice`). In the case of `voice`, the participant had to press the `Play Answer` button to listen to the audio clip. They could also pause and restart the audio clip by pressing the `Pause Answer` and `Restart Answer` buttons respectively.

Once they had read or listened to the answer passage, participants then moved to the underlying form located at ❸ to provide their judgement of the passage. Participants could choose between *'Relevant'*, *'Somewhat relevant'*, *'Non relevant'*, and *'I do not know'*. We included the final option to ensure that participants were not forced to make a relevance decision in the case that they were not sure as it has been shown that assessors are not always certain of their judgements [1]. We did not provide the participants with the option to skip parts of the audio clip or adjust the speed. Certain checks were in place to ensure reliability of relevance judgements of participants, *in addition* to the two **SC** pairings as outlined in §3.2. For `text`, the form for marking relevance ❸ appeared after five seconds. For `voice`, the form for marking relevance ❸ appeared after 50% of the audio clip had been played. Participants could also proceed to judge the next query/passage pair by clicking the *Next Query* button ❹ which was enabled only after a participant made their judgement. Once participants moved on to the next pairing, they could not go back to revise earlier judgements. No time limit was imposed on participants during the judging process.

3.5 Outcome Measures

In addition to the use of the two psychometric tests outlined in §3.3, we used interaction logging apparatus and additional surveys to capture both behavioural and experience data respectively.

Measuring Participant Behaviours. We added the *JavaScript* library `LogUI` [47] into our web-based judgement interface; it allowed us to capture a variety of different behaviours and events such as: *(i)* when the page was loaded; *(ii)* clicks on the form to record the judgement made by a participant; and *(iii)* clicks on the `Play`/`Pause`/`Restart` buttons (for `voice`). From these events, we could compute the amount of time taken for an individual to make a judgement—that is, from when the page loaded (showing the query/passage pairing) to when the `Next Query` button was clicked ❹ (Figure 2). In turn, this allowed us to compute the *time per relevance judgement*, as reported in our results.

Measuring Participant Experiences. After completing the relevance judgements, participants completed the post-task survey. Participants were asked about their perceived workload based *only* on their perceived experiences of the relevance judgement tasks. To measure workload, we used five questions from the raw *NASA TLX*

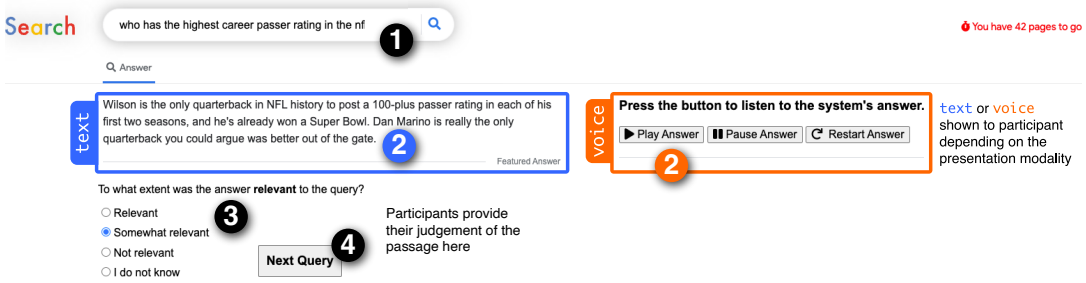

Figure 2: Composition screenshot of both the `text` and `voice` interfaces used by participants for judging query-passage pairs. Circled numbers correspond to the same in the narrative, found in §3.4.

survey, as proposed by Hart and Staveland [29]. This instrument has been used (in slightly different forms) in several prior IIR studies (e.g., [7, 8, 57]). The five selected questions from the NASA TLX are designed to measure perceived: *(i)* mental demand; *(ii)* effort; *(iii)* temporal demand; *(iv)* frustration; and *(iv)* performance. We omitted the *'physical demand'* question from the survey as it was not relevant to our task.[7] Participants responded to the five NASA TLX questions using a seven-point scale (from *"poor"* to *"good"* for performance and from *"low"* to *"high"* for the remaining four).

Measuring Participant Performance. We also computed the *accuracy* of our participants in the relevance judgement tasks. Accuracy was calculated in terms of how many Q/P pairs participants judged *correctly*—that is, their relevance judgement matching the ground truth from the QRELs. We also aggregated relevance judgements of participants on each Q/P pairing based on majority voting, as done by Kutlu et al. [39] to observe if collective judgements are more accurate. We used Krippendorff's alpha (α) to measure inter-annotator agreement (as used by Damessie et al. [18]). Lastly, we calculated Cohen's kappa (κ) [9–11] which measures the agreement of judgements with ground truths by considering chance.

3.6 Participant Demographics

We conducted an *a-priori* power analysis using *G-power* [22] to determine the minimum sample size required to test our **RQ**s. The results indicated that the required sample size—to achieve 95% power for detecting an effect of 0.25, with two groups (modality) and five measurements (passage length)—is 46. As such, we recruited 50 participants from the Prolific platform. We disqualified one participant as they failed to correctly judge our sanity check Q/P pairs (§3.2). Our n = 49 (25 for `text`, 24 for `voice`) participants were native English speakers, with a 98% approval rate on the platform—a minimum of 250 prior successful task submissions, and self-declared as having no issues in seeing colour. Participants were required to use a desktop/laptop device in order to control for variables that might affect results of the Stroop and OSPAN tests on other (smaller) devices. From our participants, 22 identified as female, 24 as male, with 3 declining to disclose this information. The mean age of our participants was 38 (*min.* 22, *max.* 69). With respect to the highest completed education level, 28 possessed a Bachelors (or equivalent), nine has a Masters (or equivalent), ten had a high school degree, and two had a PhD (or equivalent). We also asked participants how often they used a smart speaker to search for information, and listening to the provided answer—to which 13 reported daily usage, 20 said usage on a weekly basis, and 16 said never. Participants were paid GBP £11/hour.

[7] This was also done in prior studies, such as the study reported by Vtyurina et al. [80]

4 RESULTS AND DISCUSSION

This section presents the results of our experiments pertaining to our three **RQ**s. First, we provide details on the statistical tests we conducted, and how we utilised the cognitive ability tests to divide participants into *low-* and *high-ability* groups.

Statistical Tests. For our analyses[8], we conducted a series of independent sample t-tests with Bonferroni correction (α = 0.05) to observe if the modality of presentation has a significant effect on our dependent variables—accuracy of relevance judgements, the time taken to judge, and the perceived workload (**RQ1**). We also conducted a series of mixed factorial ANOVA tests (where modality of presentation is a ***between-subjects*** variable, and passage length is a ***within subjects*** variable) to observe if presentation modality, passage length, or the interaction between them have a significant effect on accuracy of relevance judgement and time taken (**RQ2**). Lastly, we conducted a series of three-way ANOVA tests to observe if the two user dispositions—working memory and inhibition–or their interaction with modality of presentation have a significant effect on the three dependent variables (**RQ3**). For **RQ2** and **RQ3**, we followed up the ANOVA with pairwise Tukey tests with Bonferroni correction (α = 0.05) to observe where significant differences lay. In the case where no significant difference was observed between the two conditions, we used equivalence testing between conditions through the *two one-sided t-tests (TOST)* procedure. The upper and lower bounds for the TOST was set at 7.5% (-ΔL = ΔU = 7.5) for accuracy, as Xu et al. [83] observed that LtR models were robust to errors of up to 10% in the dataset (we used 7.5% for *conservativeness*). For each scale of NASA-TLX, we set -ΔL = ΔU = 2.04, following Lee et al. [41], who used a bound of ±18 on a 100-point NASA TLX. For our seven-point scale, it translates to ±2.08 according to the formula of Hertzum [30].

Cognitive Ability Scores and High vs. Low Ability Groups. To examine the effect of a participant's cognitive abilities on relevance judgement accuracy (**RQ3**), we performed a median split of the scores obtained by the participants in the OSPAN (*min.* 0, *max.* 50, *mean* = 25.4(±12), *median* = 22) and Stroop test (*min.* = −300,

[8] All data and code pertaining to our analyses are released.

Table 3: RQ1: Effect of modality of passage presentation on accuracy of relevance judgement, time taken per judgement in seconds and perceived workload (IV-VIII) per participant. We also report Krippendorff's α and Cohen's κ for accuracy. † indicates significant difference in between the two conditions according to independent sample t-test. ⋆ indicates the corresponding metric is equivalent for both conditions based on the TOST procedure.

	Metrics	text	voice
I	**Accuracy ⋆**	68.40(±9.15)%	65.94(±8.56)%
	α, κ	0.41, 0.61	0.37, 0.54
II	**Majority Voting Acc.**	79.1%	75.8%
	κ	0.76	0.71
III	**Time/Rel. Judge. (sec.) †**	17.56(±9.08)	29.54(±7.85)
IV	**Mental demand⋆**	4.68(±1.60)	4.83(±1.37)
V	**Effort⋆**	4.88(±1.88)	4.00(±1.50)
VI	**Temporal Demand⋆**	4.04(±1.86)	3.08(±1.82)
VII	**Frustration†**	3.96(±2.07)	1.83(±0.82)
VIII	**Performance†**	4.16(±1.93)	5.67(±0.70)

max. = 650, *mean* = 171.25(±184), *median* = 170) respectively. The mean scores of our participants for working memory and inhibition were within one standard deviation of the reference mean scores as reported in [7], validating our methodology. Participants were thus divided into a high- and low-ability group for each of working memory (based on OSPAN test scores) and inhibition (based on Stroop test scores). Note that for inhibition, a low test score indicates high ability. Prior studies have also analysed the effects of different cognitive abilities by dividing participants into low/high ability groups using a median split [2, 7, 13, 62].

4.1 RQ1: Modality of Passage Presentation

Table 3 presents the main results for **RQ1**. There was no significant difference in judgement accuracy (row **I**, Table 3) between participants in `text` and those in `voice` ($t(47) = 0.97, p = 0.33$). TOST revealed that accuracy of judgements across both conditions were *equivalent* ($p = 0.02$). The inter-annotator agreement (α) was slightly higher in `text`. When using majority voting to aggregate relevance judgements (on average we had eight judgements per Q/P pair in each condition), we found that the accuracy increased from 68% and 66% to 79% and 76% respectively for `text` and `voice` (**II**, Table 3). This observation is in line with prior work [39], which shows that aggregating judgements from several assessors is more reliable than a single untrained assessor. Cohen's κ also increased with majority voting for both experimental conditions, indicating an increase in judgement reliability. Participants also showed similar trends of relevance judgement accuracy per relevance label category for both experimental conditions. As shown in Figure 3, participants in both conditions were most accurate in judging *'relevant'* passages (in line with findings by Alonso and Mizzaro [4]), followed by 'non-relevant' passages. 'Somewhat relevant' passages were most difficult to judge as participants in both conditions judged them correctly about half the time. With respect to the time taken to judge (**III**, Table 3), judgements in `text` were made significantly faster ($t(47) = -4.93, p < 0.001$) than in `voice`.

Figure 3: Accuracy of relevance judgements per label category for both `text` and `voice`. Diagonals represent percentage of time the true labels were *correctly* predicted by participants. Here, `R` = `RELEVANT`, `SR` = `SOMEWHAT-RELEVANT`, `NR` = `NON-RELEVANT` and `IDK` = *I do not know.*

In terms of workload measured using NASA-TLX, there was no significant difference in averages between the two cohorts in terms of perceived mental demand, effort, and temporal demand (**IV-VI**, Table 3). The TOST procedure revealed equivalent scores ($p < 0.05$) provided by participants for these three items of the NASA-TLX scale. For the other dimensions of NASA-TLX questionnaire, participants in `text` reported they felt significantly more frustrated (**VII**, Table 3) while performing the task than those in `voice` ($t(47) = 4.69, p < 0.001$). Participants in `voice` also reported significantly higher perceived performance (**VIII**, Table 3) when compared to the former ($t(47) = -3.60, p < 0.001$).

Overall, we found that participants listening to voice passages were equally accurate to their text counterparts. Vtyurina et al. [80] also observed that the probability of participants to identify the most relevant document was the same for both text and voice conditions. However, the authors implemented a different task design to ours. Their participants were presented with a list of results, and were significantly better at identifying the correct order of relevance when the summaries were presented in text modality. Insofar as to acknowledging the difference in task design, our observations with regards to the accuracy of participants with respect to relevance judgements across modalities are found to be partially in line with those of Vtyurina et al. [80]. We also observed that `voice` participants perceived a lower or equal workload when compared to those of `text`, in contrast to the other study's findings [80]. This can be attributed to their study setup. Contrary to ours, their presentation modality was a *within-subjects* variable. Our results indicate the proficiency of participants with both modalities for the given design of the task.

4.2 RQ2: Passage Length

Table 4 presents results related to **RQ2**. Like modality of presentation, passage length or its interaction with presentation modality did not have a significant effect on the relevance judgement accuracy (comparing rows **Ia** and **Ib**, Table 4). The TOST procedure revealed that for **XS** ($p = 0.01$) and **L** ($p = 0.001$) passages, judgement accuracy was *equivalent* across both conditions. Aggregating judgements via majority voting increased relevance judgement accuracy across all passage lengths for both text and voice conditions

Table 4: RQ2: Effects of passage length and presentation modality on accuracy of relevance judgements (with Krippendorff's α, Cohen's κ) and time taken. A bold number indicates that the metric for the corresponding presentation modality is significantly more than that for the other modality for the particular passage length. ^{xs,s,m,l,xl} indicates significant difference (within the same experimental condition) compared to XS, S, M, L, XL passage lengths. ⋆ indicates equivalence between the two conditions.

	Metrics	Mode	Passage Length				
			XS	S	M	L	XL
I	Accuracy (%)	text	66.7(±19.0)⋆	74.5(±14.7)	66.5(±17.5)	61.0(±18.0)⋆	74.0(±19.0)
		α, κ	0.37, 0.57	0.51, 0.67	0.43, 0.55	0.29, 0.50	0.44, 0.68
		voice	67.7(±21.9)⋆	64.06(±15.0)	72.4(±19.4)	61.5(±13.9)⋆	64.0(±16.7)
		α, κ	0.39, 0.56	0.44, 0.49	0.49, 0.63	0.27, 0.51	0.35, 0.48
II	Maj. Voting Acc. (%)	text	75	83	79	75	92
		κ	0.73	0.81	0.74	0.73	0.91
		voice	79	79	79	67	79
		κ	0.78	0.78	0.73	0.62	0.76
III	Time Taken (sec.)	text	14.11(±6.3)	15.25(±7.7)	15.15(±6.8)	21.39(±12.41)	21.86(±12.7)
		voice	17.3(±5.0)^{m,l,xl}	**25.47**(±11.8)^{xl}	**28.45**(±14.8)^{xs,xl}	**31.04**(±6.15)^{xs,xl}	**45.39**(±9.6)^{xs,s,m,l}

(comparing rows **Ia**-**IIa** and **Ib**-**IIb**, Table 4). However, for **XL** passages (**IIa**-**IIb**, Table 4), the difference in accuracy after majority voting was more than 10% (with text being more accurate). We also observed a higher difference in Cohen's κ and Krippendorff's α for **XL** passages between the text and voice conditions. These results indicated a higher inter-annotator agreement and reliability of judgements for text compared to participants in voice with regards to **XL** passages.

With respect to the time taken for judging, we have already seen (Section 4.1) that presentation modality significantly affected the time to judge. Mixed factorial ANOVA showed that passage length had a significant main effect ($F = 21.6$, $p = 3.3e^{-15}$) on the time taken to assess. A post-hoc test revealed a significant difference in the time taken to judge of the following pairs of passage lengths (with the latter passage length category taking more time): **XS**–**M** ($p = 0.02$), **XS**–**L** ($p < 0.001$), **XS**–**XL** ($p < 0.001$), **S**–**XL**($p < 0.001$) and **M**–**XL**($p = 0.001$). There was also a significant interaction effect between passage length and presentation modality on the amount of time taken. Pairwise Tukey test revealed that except for **XS** passages, judging relevance in voice took significantly longer for participants as compared to doing the same in text (**bold numbers**, row **III**, Table 5). In voice (**IIIb**, Table 5), it took participants significantly longer to judge relevance, as passages (audio clips) increased in length. Superscripts (in Table 4) indicate which pairs of passage length were significantly different in voice in terms of time taken per judgement.

In summary, we did not observe a significant difference in relevance judgement accuracy across different passage lengths in both conditions. We observed judging relevance of **XS** passages was *equivalent* in terms of accuracy and time taken across both text and voice. However, for **XL** passages, relevance judgements in text were more reliable (indicated by majority voting accuracy, α and κ when compared to that in voice). There was no clear trend between passage length and assessor agreement observed in contrast to findings from [12], possibly due to differences in the type of documents assessed. Although it took longer on average to judge a lengthier passage in text, there was no significant difference in terms of the time taken to judge relevance of different passage

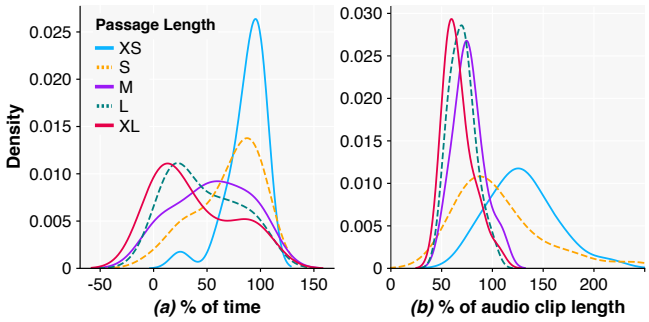

Figure 4: The trend of voice participants judging relevance w.r.t. time taken for passages of various length: *(a)* % of time participants listened to the entire audio clip; and *(b)* at what point was relevance judged (as a % of audio clip length).

lengths (a similar trend as observed in [37, 65]). For longer passages, participants in voice took significantly longer to judge relevance than in text. For **XL** passages, we found that participants were taking twice as long in voice when compared to text.

Why does it take longer for participants to judge longer passages in the voice condition? In order to control for confounding variables, we did not let participants speed up the audio clips, nor did we provide them with a seeker bar to skip ahead. We found evidence that participants moved on to the next Q/P pairing as soon as they were satisfied with their assessment. Indeed, they did not wait for the audio clip to finish playing before moving on to the next Q/P pair for longer passages (Figure 4 (a)). We also let participants mark the relevance of a passage in voice only after 50% of the audio clip had been played (Section 3.1). However, as seen from Figure 4 (b), participants took longer to judge relevance (rather than right at the 50% mark). For **XL** passages, it was at the 66% of the audio clip on average. This suggests that it indeed took more time for participants in voice compared to text to assimilate the information and come to a judgement decision for longer passages.

Table 5: RQ3: Summary of main effects of *Presentation Modality (PM), Working Memory (WM), Inhibition (IN),* and effects of the interaction of WM and IN with PM on accuracy of relevance judgement, time taken, and perceived workload. A ✓ indicates significant effect of a 3-way ANOVA test ($p < 0.05$) on the particular dependent variables and ✗ indicates no significant effect.

	PM *(Presentation)*	**WM** *(Working Memory)*	**IN** *(Inhibition)*	**WMxPM**	**INxPM**
I Accuracy	✗	✗	✗	✗	✓ ($F = 4.89, p = 0.03$)
II Time Taken (sec.)	✓ ($F = 22.17, p < 0.001$)	✗	✗	✗	✗
III Mental Demand	✗	✗	✗	✗	✗
IV Effort	✗	✗	✗	✓ ($F = 5.1, p = 0.03$)	✗
V Temporal Demand	✗	✓ ($F = 7.88, p = 0.01$)	✓ ($F = 7.39, p = 0.01$)	✗	✗
VI Frustration	✓ ($F = 8.36, p = 0.008$)	✗	✗	✗	✗
VII Performance	✓ ($F = 5.83, p = 0.02$)	✗	✗	✗	✗

4.3 RQ3: Assessor Cognitive Abilities

Table 5 contains the results for our third research question. Here, ✓ indicates a significant effect ($p < 0.05$) on the particular dependent variable, and ✗ indicates no significant effect.

None of the independent variables—modality of passage presentation (**PM**), working memory (**WM**), and inhibition (**IN**)—had a significant main effect on judgement accuracy. The interaction between the **IN** of participants and presentation modality (**IN** x **PM**) had a significant effect on the accuracy ($F = 4.89, p = 0.03$). Pairwise Tukey test revealed that in voice participants with higher **IN** performed significantly better than those with lower **IN** ($70.5 \pm 7.2\%$ vs. 59.5 ± 4.8 %). The post-hoc test ($p = 0.01$) also revealed participants with low **IN** performed significantly better in text than those in voice (70.0 ± 9.5 % vs. 59.5 ± 4.8 %). We found significant main effects of **PM** on the time taken to judge relevance ($F = 22.17, p < 0.001$), reaffirming findings from Section 4.1 and Section 4.2.

With respect to the perceived workload, working memory had significant main effects on perceived temporal demand ($F = 7.88, p = 0.01$). A post-hoc test ($p < 0.001$) revealed that participants with high **WM** reported significantly less temporal demand as compared to those with low **WM** (2.5 ± 1.3 vs. 4.6 ± 1.7 respectively). **IN** also had significant main effects on perceived temporal demand ($F = 7.4, p = 0.01$). A post-hoc test ($p < 0.001$) revealed that participants with high **IN** reported significantly less temporal demand as compared to those with low **IN** (2.74 ± 1.4 vs. 4.59 ± 1.9, respectively). Presentation modality had significant main effects on perceived frustration ($F = 8.36, p = 0.008$) and performance ($F = 5.83, p = 0.02$)—confirming observations from Section 4.1—with participants in voice reporting a lower workload. Lastly, the interaction between **WM** and presentation modality (**WM** x **PM**) had a significant effect on perceived effort for the task ($F = 5.1, p = 0.03$). Post-hoc tests revealed that participants with high **WM** felt that judging using text required significantly more effort when compared to those in voice ($p = 0.001$).

In summary, we found that **IN** is a more important trait than **WM**, specifically for relevance judgement accuracy in the voice modality. Low **IN** participants in the voice condition were less accurate—since we *did not control for the audio device of the participants*, and consequently not for the background noise they were subjected to, low **IN** participants in voice were less effective in focusing on the passages while judging relevance [71, 72]. We leave exploring the effect of background noise as future work. In our study, the interplay between cognitive abilities and modality of presentation on perceived workload had different effects. High **IN** and **WM** participants felt less temporal demand. High **WM** in text felt more perceived effort compared to those in voice. Our results imply that we should design tasks for collecting relevance assessments to match the preference and abilities of crowdworkers [5, 52].

5 CONCLUSIONS

We explored the feasibility of using voice as a modality to collect relevance judgements of query-passage pairs. We investigated the effect of passage length and the cognitive abilities of participants on judgement accuracy, the time taken, and perceived workload.

RQ1 On average, the relevance judgement accuracy was equivalent across both text and voice. Participants also perceived equal or less workload in voice when compared to text.

RQ2 For **XS** passages, the performance and time taken for relevance judgements was *equivalent* between both voice and text. As passages increased in length, it took participants significantly longer to make relevance judgements in the voice condition; for **XL** passages voice, participants took twice as much time and the judgements were less reliable compared to text.

RQ3 Inhibition impacted the relevance judgement accuracy in the voice condition—participants with higher inhibition were significantly more accurate than those with lower inhibition.

Our results from **RQ1** suggest that we can leverage the voice modality for this task. **RQ2** points to the possibility of designing hybrid tasks, where we can use the voice modality for judging shorter passages and text for longer passages. The results of **RQ3** showed that selecting the right participants for the relevance judgement task is important. We should be mindful to personalise the task to match the preference and abilities of crowdworkers [5, 52].

There are several open questions for future work. We did not provide participants with the option to speed-up voice passages—*does letting them speed-up or skip passage parts reduce time for longer passages without reducing accuracy?* We also did not test the limit of length—*how long can documents be for equal accuracy in the text and voice modality?* Future work should also explore mobile devices for playing voice passages—*can we collect relevance judgements by offering more flexibility to crowdworkers?* Lastly, since asking to provide rationales for judgements has been shown to improve relevance judgement accuracy of crowdworkers in the text modality [39], exploring the effects of rationale in voice-based relevance judgements should be a worthwhile endeavour.

REFERENCES

[1] A.L. Al-Harbi and M. Smucker. 2014. A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th information interaction in context symposium*. 195–204.

[2] A. Al-Maskari and M. Sanderson. 2011. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management* 47, 5 (2011), 719–729.

[3] M. Almeida, M. Bilal, A. Finamore, I. Leontiadis, Y. Grunenberger, M. Varvello, and J. Blackburn. 2018. Chimp: Crowdsourcing human inputs for mobile phones. In *Proceedings of the 2018 World Wide Web Conference*. 45–54.

[4] O. Alonso and S. Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Vol. 15. 16.

[5] M. Alsayasneh, S. Amer-Yahia, E. Gaussier, V. Leroy, J. Pilourdault, R.M. Borromeo, M. Toyama, and J.M. Renders. 2017. Personalized and diverse task composition in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* 30, 1 (2017), 128–141.

[6] J. Anderton, M. Bashir, V. Pavlu, and J.A. Aslam. 2013. An analysis of crowd workers mistakes for specific and complex relevance assessment task. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1873–1876.

[7] J. Arguello and B. Choi. 2019. The effects of working memory, perceptual speed, and inhibition in aggregated search. *ACM Transactions on Information Systems (TOIS)* 37, 3 (2019), 1–34.

[8] J. Arguello, W.C. Wu, D. Kelly, and A. Edwards. 2012. Task complexity, vertical display and user interaction in aggregated search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 435–444.

[9] R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics* 34, 4 (2008), 555–596.

[10] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 667–674.

[11] J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *arXiv preprint cmp-lg/9602004* (1996).

[12] P. Chandar, W. Webber, and B. Carterette. 2013. Document features predicting assessor disagreement. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 745–748.

[13] B. Choi, R. Capra, and J. Arguello. 2019. The effects of working memory during search tasks of varying complexity. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 261–265.

[14] A. Chuklin, A. Severyn, J.R. Trippas, E. Alfonseca, H. Silen, and D. Spina. 2019. Using audio transformations to improve comprehension in voice question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 164–170.

[15] P. Clough and M. Sanderson. 2013. Evaluating the performance of information retrieval systems using test collections. (2013).

[16] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos. 2021. Overview of the TREC 2020 deep learning track. *arXiv preprint arXiv:2102.07662* (2021).

[17] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E.M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[18] T.T. Damessie, F. Scholer, and J.S. Culpepper. 2016. The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proceedings of the 21st Australasian Document Computing Symposium*. 41–48.

[19] R. Datta, D. Joshi, J. Li, and J.Z. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)* 40, 2 (2008), 1–60.

[20] D. Davidson. 1977. The effect of individual differences of cognitive style on judgments of document relevance. *Journal of the American Society for information Science* 28, 5 (1977), 273–284.

[21] A. Diamond. 2013. Executive functions. *Annual review of psychology* 64 (2013), 135.

[22] F. Faul, E. Erdfelder, A.G. Lang, and A. Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[23] X. Fu, E. Yilmaz, and A. Lipani. 2022. Evaluating the Cranfield Paradigm for Conversational Search Systems. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. 275–280.

[24] S.M. Gass, J.N. Behney, and B. Uzum. 2013. Inhibitory control, working memory and L2 interaction. In *Psycholinguistic and sociolinguistic perspectives on second language learning and teaching*. Springer, 91–114.

[25] R. Gligorov, M. Hildebrand, J. Van Ossenbruggen, L. Aroyo, and G. Schreiber. 2013. An evaluation of labelling-game data for video retrieval. In *European Conference on Information Retrieval*. Springer, 50–61.

[26] S. Gordon-Salant and S.S. Cole. 2016. Effects of age and working memory capacity on speech recognition performance in noise among listeners with normal hearing. *Ear and hearing* 37, 5 (2016), 593–602.

[27] K. Hagerty. 1967. Abstracts as a Basis for Relevance Judgment. (1967).

[28] D.K. Harman. 1993. Overview of the first TREC conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. 36–47.

[29] S.G. Hart and L.E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[30] M. Hertzum. 2021. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. *Ergonomics* 64, 7 (2021), 869–878.

[31] D. Hettiachchi, Z. Sarsenbayeva, F. Allison, N. van Berkel, T. Dingler, G. Marini, V. Kostakos, and J. Goncalves. 2020. " Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[32] M. Hosseini, I.J. Cox, N. Milić-Frayling, G. Kazai, and V. Vinay. 2012. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings 34*. Springer, 182–194.

[33] K.S. Jones and C. J. van Rijsbergen. 1976. Information retrieval test collections. *Journal of documentation* (1976).

[34] H.J. Jung and M. Lease. 2011. Improving consensus accuracy via z-score and weighted voting. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[35] M.J. Kane and R.W. Engle. 2003. Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of experimental psychology: General* 132, 1 (2003), 47.

[36] G. Kazai, J. Kamps, and N. Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval* 16 (2013), 138–178.

[37] J.A. Konstan, B. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl. 1997. Grouplens: Applying collaborative filtering to usenet news. *Commun. ACM* 40, 3 (1997), 77–87.

[38] B. Koopman and G. Zuccon. 2014. Relevation! An open source system for information retrieval relevance assessment. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 1243–1244.

[39] M. Kutlu, T. McDonnell, T. Elsayed, and M. Lease. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research* 69 (2020), 143–189.

[40] J. Lai and N. Yankelovich. 2006. Speech interface design. (2006).

[41] J. Lee, S.S. Rodriguez, R. Natarrajan, J. Chen, H. Deep, and A. Kirlik. 2021. What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 512–520.

[42] H. Luan, Y.T. Zheng, M. Wang, and T.S. Chua. 2011. VisionGo: towards video retrieval with joint exploration of human and computer. *Information Sciences* 181, 19 (2011), 4197–4213.

[43] C. Lustig, C.P. May, and L. Hasher. 2001. Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General* 130, 2 (2001), 199.

[44] A. MacFarlane, A. Albrair, C.R. Marshall, and G. Buchanan. 2012. Phonological working memory impacts on information searching: An investigation of dyslexia. In *Proceedings of the 4th Information Interaction in Context Symposium*. 27–34.

[45] I. Mackie, J. Dalton, and A. Yates. 2021. How deep is your learning: the DL-HARD annotated deep learning dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2335–2341.

[46] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 1–32.

[47] D. Maxwell and C. Hauff. 2021. LogUI: Contemporary Logging Infrastructure for Web-Based Experiments. In *Proceedings of the 43th ECIR*. (In press).

[48] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C.E. Kahn, and W. Hersh. 2010. Overview of the CLEF 2009 medical image retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 72–84.

[49] C. Murad and C. Munteanu. 2020. Designing voice interfaces: Back to the (curriculum) basics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[50] C. Murad, C. Munteanu, L. Clark, and B.R. Cowan. 2018. Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 269–276.

[51] C. Nowacki, A. Gordeeva, and A.H. Lizé. 2020. Improving the usability of voice user interfaces: a new set of ergonomic criteria. In *International Conference on Human-Computer Interaction*. Springer, 117–133.

[52] P. Organisciak, J. Teevan, S. Dumais, R. Miller, and A. Kalai. 2014. A crowd of your own: Crowdsourcing for on-demand personalization. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 2. 192–200.

[53] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

[54] N. Rajput and A.A. Nanavati. 2012. Evaluation of mobile and pervasive speech applications. (2012).

[55] S.M. Randhawa, T. Ahmad, J. Chen, and A.A. Raza. 2021. Karamad: A Voice-based Crowdsourcing Platform for Underserved Populations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[56] K. Roitero, E. Maddalena, S. Mizzaro, and F. Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Information Processing & Management* 58, 6 (2021), 102688.

[57] N. Roy, D. Maxwell, and C. Hauff. 2022. Users and Contemporary SERPs: A (Re-) Investigation. In *45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM DL, 2765–2775.

[58] M. Rudner, T. Lunner, T. Behrens, E.S. Thorén, and J. Rönnberg. 2012. Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology* 23, 08 (2012), 577–589.

[59] T. Saracevic. 1969. Comparative effects of titles, abstracts and full texts on relevance judgments. *Proceedings of the American Society for Information Science* 6, 1 (1969), 293–299.

[60] T. Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American society for information science and technology* 58, 13 (2007), 1915–1933.

[61] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. 2010. "Your word is my command": Google search by voice: A case study. In *Advances in speech recognition*. Springer, 61–90.

[62] F. Scholer, D. Kelly, W.C. Wu, H.S. Lee, and W. Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 623–632.

[63] F. Scholer, A. Turpin, and M. Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 1063–1072.

[64] J. Sherwani, D. Yu, T. Paek, M. Czerwinski, Y.C. Ju, and A. Acero. 2007. Voicepedia: Towards speech-based access to unstructured information. In *Eighth Annual Conference of the International Speech Communication Association*.

[65] M. Moritaand Y. Shinoda. 2012. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval*. 272–281.

[66] B. Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (2000), 63–65.

[67] A. Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.

[68] A. Singhal, G. Salton, M. Mitra, and C. Buckley. 1996. Document length normalization. *Information Processing & Management* 32, 5 (1996), 619–633.

[69] M.D. Smucker and C.L. Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 95–104.

[70] M.D. Smucker, G. Kazai, and M. Lease. 2012. *Overview of the trec 2012 crowdsourcing track*. Technical Report. Sch. of Info., Uiv. Texas Austin.

[71] V. Stenbäck. 2016. *Speech masking speech in everyday communication: The role of inhibitory control and working memory capacity*. Vol. 1559. Linköping University Electronic Press.

[72] V. Stenbäck, E. Marsja, M. Hällgren, B. Lyxell, and B. Larsby. 2021. The contribution of age, working memory capacity, and inhibitory control on speech recognition in noise in young and older adult listeners. *Journal of Speech, Language, and Hearing Research* 64, 11 (2021), 4513–4523.

[73] J.R. Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology* 18, 6 (1935), 643.

[74] T. Tombros and F. Crestani. 1999. A study of users' perception of relevance of spoken documents. *Rapport technique TR-99-013, Berkeley, CA* (1999).

[75] J.R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. 2015. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*. 991–994.

[76] M.L. Turner and R.W. Engle. 1989. Is working memory capacity task dependent? *Journal of memory and language* 28, 2 (1989), 127–154.

[77] R. Vaish, K. Wyngarden, J. Chen, B. Cheung, and M.S. Bernstein. 2014. Twitch crowdsourcing: crowd contributions in short bursts of time. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3645–3654.

[78] A. Vashistha, P. Sethi, and R. Anderson. 2017. Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 1855–1866.

[79] A. Vashistha, P. Sethi, and R. Anderson. 2018. BSpeak: An accessible voice-based crowdsourcing marketplace for low-income blind people. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[80] A. Vtyurina, C. Clarke, E. Law, J. R Trippas, and H. Bota. 2020. A mixed-method analysis of text and audio search interfaces with varying task complexity. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 61–68.

[81] J. Wiley and A.F. Jarosz. 2012. How working memory capacity affects problem solving. In *Psychology of learning and motivation*. Vol. 56. Elsevier, 185–227.

[82] C. Xu, Z. Li, H. Zhang, A. Rathore, H. Li, C. Song, K. Wang, and W. Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 14–26.

[83] J. Xu, C. Chen, G. Xu, H. Li, and E.R.T Abib. 2010. Improving quality of training data for learning to rank using click-through data. In *Proceedings of the third ACM international conference on Web search and data mining*. 171–180.

[84] N. Yankelovich and J. Lai. 1998. Designing speech user interfaces. In *CHI 98 Conference Summary on Human Factors in Computing Systems*. 131–132.

[85] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J.M. Jose, and L. Azzopardi. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval* 16 (2013), 267–305.

[86] K. Zyskowski, M.R Morris, J.P. Bigham, M.L. Gray, and S.K. Kane. 2015. Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1682–1693.