Information Processing and Management 59 (2022) 102985

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

E-government information search by English-as-a Second Language speakers: The effects of language proficiency and document reading level

Morgan Harvey [a,*], David Brazier [b]

[a] *Information School, The University of Sheffield, UK*
[b] *School of Computing, Edinburgh Napier University, UK*

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Information behaviour
Language proficiency
Reading level
E-government services

ABSTRACT

A rapid increase in the use of web-based technologies – and corresponding changes in government and local council policies – in recent years, means that many vital services are now provided solely online. While this has many potential benefits, it can place additional burdens on certain demographic groups, some of whom may become considerably disadvantaged or even disenfranchised. This is particularly problematic for English-as-a Second Language (ESL) speakers, who are often immigrants or refugees and thus have a greater need to access these e-government services, and who may struggle to understand and assess the relevance of complex documents. In this work we investigate the search behaviours and performance of native English speakers and two different groups of ESL speakers when completing e-government tasks, and the effect of document readability/complexity. In contrast with previous work, our results show significant differences between groups of varying language proficiency in terms of objective search performance, time on task, and self-perceived performance and confidence. We also demonstrate that document reading level moderates the effect of language proficiency on objective search performance. The findings contribute to our existing understanding of how English language proficiency affects search for e-government topics, and have important implications for the future development of e-government services to ensure more equitable access and use.

1. Introduction

With Internet technologies now assimilated into virtually every facet of our daily lives, near unlimited access to unprecedented amounts of information allows for many tasks to be completed through digital means. This is assuming, of course, that one has the necessary access, experience, proclivity and language skills. As companies and governments move services from “traditional” information access paradigms (e.g., face to face, telephone, written application or letters) to digital platforms, such as web portals or mobile applications, users are being forced to comply with the digital requirements placed upon them, regardless of their circumstances or their risk of being segregated (Vinson, 2009). This can place a huge burden on users with regards to their access to such, often vital, digital resources. This is particularly the case for certain groups of users (Savolainen, 2016) - older people, those from disadvantageous socio-economic backgrounds and people interacting with services and resources that are not written in their mother tongues (Brazier & Harvey, 2017a).

[*] Corresponding author.
E-mail address: m.harvey@sheffield.ac.uk (M. Harvey).

https://doi.org/10.1016/j.ipm.2022.102985
Received 2 June 2021; Received in revised form 12 April 2022; Accepted 16 May 2022
Available online 31 May 2022

With web content written in the English language making up 63.7% of all content (W3Techs, 2022), non-native speakers of English (English as a Second Language speakers or ESL) face considerable challenges when attempting to search for, assess and understand information online. This is particularly so for government and local council documents, where a lack of resources for translation means that these are frequently not offered in the searcher's first language (Alam & Imran, 2015; Harvey, Hastings, & Chowdhury, 2021). Work has shown that, despite often being very confident in their English abilities, ESL speakers perform poorly on migration-related search tasks (Brazier & Harvey, 2017a). ESL speakers have been shown to behave differently from native speakers when searching, although results in terms of performance have often been inconclusive (Bogers, Gäde, Hall, & Skov, 2016; Brazier & Harvey, 2018), perhaps because the ESL speakers studied were nevertheless proficient in the second ('L2') language.

Reliance on users' information or digital literacy has a direct impact on the veracity of the information to which they are exposed through their ability to assess a given document's "currency, relevance, authority, accuracy and purpose" (Parsazadeh, Ali, & Rezaei, 2018, pp.76). Although useful and trustworthy documents may exist to solve a user's information need, they may lack the skills necessary to separate veracious and useful documents from fallacious and unhelpful ones (Helbig, Gil-García, & Ferro, 2009). The accuracy of a document's content may be particularly crucial if it is being used to solve problems related to digital services – such as housing, council tax or visas – or health-related issues (Swire-Thompson & Lazer, 2020). Furthermore, a user's general literacy – their ability to understand the content they are presented with – may have a considerable effect on the documents they choose when attempting to resolve a given information need (Hahnel, Goldhammer, Kröhne, & Naumann, 2018).

In this work we build and expand upon previous work in the literature to specifically investigate the search behaviours and performance of both native and non-native speakers on contextually-relevant tasks, and the effect of document readability/complexity on this, in a UK context. We compare the behaviour and performance of three different groups of users: native English speakers, ESL speakers with high levels of English proficiency (as in previous work), and, crucially, less proficient ESL speakers, who were at the time attending a mandatory University course to improve their English skills.

Our results provide deeper insights into the differences (and similarities) between how native and non-natives use English-language search engines and the effect of document complexity (as measured using the new Dale–Chall reading level) on this behaviour and performance. Our results highlight that ESL speakers are certainly not a homogeneous group in terms of their search behaviour and performance and that different potential issues and biases may arise depending on their (perceived) level of ability. We also identify the importance of document reading level and show that this acts as a moderator between user language proficiency and performance in terms of successfully selecting relevant documents. These findings add to the existing literature on how language proficiency affects search for e-government topics and have important implications for the future development of more equitable and useable e-government services.

2. Related work

This research considers the need for people with varying levels of English language ability to search for and understand documents in the context of e-government tasks. We investigate the effect of document complexity (i.e., reading level) on the performance of users and how this interacts with their English proficiency. First, we review literature on the use of e-government services; the information behaviours of ESL speakers; how language proficiency affects search and reading behaviour; and measuring document reading levels. Note that in the below we refer to the 'L1' language, which is the reader's native language, and 'L2' languages, which is any second languages they may speak.

2.1. Use of e-government services

Numerous works consider e-government services, the public's engagement with such services and barriers to their use (Aham-Anyanwu & Li, 2017; Burroughs, 2009; Komba & Lwoga, 2015; Lambert, 2013), as well as e-government use within the field of information retrieval (Freund, 2013). The UK government's drive towards digital-only services in recent years, which culminated in the "Digital by Default" campaign (Al-Muwil, Weerakkody, El-Haddadeh, & Dwivedi, 2019; Yates, Kirby, & Lockley, 2015), has seen many services moved solely online, yet there remains a limited amount of research exploring the topic – particularly in a UK context – since this development.

Inequalities relating to the accessibility of digital technologies have led to the development of the term "digital divide" (Selwyn & Facer, 2007). Initially referring to the distinction between those who did or did not have access to the Internet, over time this changed to mean the skills gap between capable users and novice or inexperienced users (Van Dijk, 2006). Many scholars now suggest that it is the significant variation in literacy skills that is contributing most to this digital divide (Clinton, 2019; Cohron, 2015; Macevičiūtė & Manžuch, 2018). While governmental websites do contain relevant information, it may be difficult to access for some users due to a lack of technological access, knowledge, English language ability or limited interpretative or analytical skills (Alam & Imran, 2015; Choudrie, Ghinea, & Songonuga, 2013; Oduntan & Ruthven, 2019; Ruokolainen & Widén, 2020). There are issues related to a lack of translation for non-English speaking people, implying that local authorities may be oblivious to this growing population or simply do not have the resources to cater to their needs (Brazier & Harvey, 2017a; Harvey et al., 2021).

Research into user engagement with e-government services suggests that many factors can impact how people assess relevance. One of the most influential factors is the content and, more specifically, how long documents are and the complexity of the language used (Aham-Anyanwu & Li, 2017). In the use of Canadian e-government documents Freund (2013) establishes that a document's genre also plays an important role in users establishing relevance, but did not evaluate the effect of reading level or users' language ability. Other work has shown that having an awareness of the impact users' native languages have on accessibility and use of such services is imperative (Burroughs, 2009). An important, but as yet uninvestigated, question is the extent to which a user's English language proficiency affects the selection and relevance judgements of online English documents in e-government contextual search tasks.

2.2. Information behaviour of ESL speakers

Studies between native and non-native speakers' information behaviours reveal contradictory evidence. Some reveal significant differences between these groups (Brazier & Harvey, 2018; Steichen & Lowe, 2020), while others have found that there may not be such stark differences (Bogers et al., 2016; Haley & Clough, 2017; Steichen & Lowe, 2020). Chu and colleagues (Chu, Jozsa, Komlodi, & Hercegfi, 2012; Chu & Komlodi, 2017) suggest that users who search using a second language require significantly more time, submit more query reformulations and view/assess a greater number of websites and that those with only an intermediate grasp of the English language struggle with query reformulation when searching for English-language documents. Bogers et al. (2016) considered the problem of searching for books and found, somewhat in contrast, that English non-natives spend more time on task than English native speakers, but that there is otherwise very little difference between natives and non-natives in relation to the number of queries, query length, or depth of result inspection. They surmised this could be as a result of their users' experience in searching for books in English and having, albeit not native-level, at least proficient L2 language skills.

In their study of multilingual users, Rózsa, Komlodi, and Chu (2015) identified a propensity for short specific, often one-word queries to be submitted, which led to a large proportion of vague results and overwhelmed users. This in turn led to users spending more time reading documents or limiting their selection to just one specific document and exhausting the content there, rather than exploring the results list more widely. Brazier and Harvey (2017a, 2017b) studied the search behaviours and performance of ESL speakers when given search tasks that new immigrants to a country might need to perform. They found that, while most users were very confident in their English language searching abilities, they did not tend to perform very well. Lack of confidence and proficiency in English results in a tendency to rely on assistive functionality, such as autofill or recommended links (Rózsa et al., 2015; Steichen & Lowe, 2020). While this is sufficient for platforms that offer such functionality (i.e., modern search engines), in-page content search and a lack of assistive functionality in web documents have seen an over-reliance on other search engines (e.g., using Google) or selection of documents based purely on URL (host details) rather than the document's subject matter (Brazier & Harvey, 2018).

In a study between native language and foreign language information seeking, Józsa, Köles, Komlódi, Hercegfi, and Chu (2012) identified two distinct search strategies; superficial or cursory and in-depth, with little differences in performance when applying an in-depth strategy in both languages. Alternatively, it was found the superficial strategy in a foreign language performed much worse than in the native language. One explanation being that foreign language users, who may not be as familiar with nuances in the language, may miss signs of such subtle markers when not thoroughly analysing a document and thus may gather a lower quality result set. These missed signs can be linked to information scent, as detailed in the Information Foraging Theory (Pirolli, 2009), which is the perception of the value and cost of following a trail of information, based on its adjacent cues, such as hyper-links within a web document. If these cues are missed or misinterpreted it is understandable that users may judge the time and effort of continuing to exceed the benefits of proceeding further (Kralisch & Berendt, 2005).

2.3. Language as a determinant

Typically, lower levels of proficiency result in non-native speakers having less developed terminological knowledge than that of native speakers (Kralisch & Berendt, 2005). Even in the event of attaining fluency in both a native and secondary language, second language reading is inherently more complex than L1 reading (Grabe, 2009; Koda, 2007, pg.129). From a web search perspective, language proficiency can have a significant impact on the search experience and outcomes (Chu et al., 2012; Hahnel et al., 2018; Józsa et al., 2012; Kang, 2014).

Grabe (2009, pg.6) identified that electronic communication methods amplify the requirement for skilled reading rather than compensating for weak literacy skills, which more recent research shows continues to be of concern (Clinton, 2019). Even with the advent of assistive functionalities (Clough & Eleta, 2010), which are utilised by a large proportion of online search engines and websites, having the skills to identify, interpret and evaluate information pertinent to the task or goal is fundamental to a user's full inclusion in both digital and non-digital communities (Clinton, 2019; Józsa et al., 2012). Education level and domain knowledge have been found to contribute (Weber, Becker, & Hillmert, 2018), with a segregation between second language users with high and low educational levels (Kang, 2014; Kralisch & Berendt, 2005). Domain or topic knowledge has been shown to affect search behaviours (Savolainen & Kari, 2006; Tamine & Chouquet, 2017), accounting for differences in the sites that were visited, the vocabulary used for querying, search behaviour patterns and the overall success of the search (Arguello, Choi, & Capra, 2018; White, Dumais, & Teevan, 2009).

2.4. Content and the role of language

The language of content is especially important when users are trying to find information. Whether the language is intended to cater towards the lay public or a specific target audience dictates the ease with which it can be read, interpreted, processed and used. There has been an extensive amount of research into multilingual search and information seeking, which has explored the content language and participants code-switching across query formulation and results evaluation (Aula & Kellar, 2009; Steichen & Freund, 2015; Steichen & Lowe, 2020; Wang & Komlodi, 2018). These studies all confirming what Berendt and Kralisch (2009) proposed: There is often a reliance on English content in the resolution of tasks for certain topics (Aula & Kellar, 2009) (e.g., science and technology, entertainment, and nature and environment) but a tendency to choose native languages (where English is the second language) for topics related to health, home and family, and shopping. Even in the cases where a non-English native language was the preference, there was still a large proportion of second language content selected for trustworthiness and relevance (Steichen & Lowe, 2020). This may have particular relevance in the context of e-government tasks, which predominantly fall in the latter group of topics.

2.4.1. Reading

A user's ability to read content is paramount to the information addressing a particular need, or in relaying often vital details about a topic, which the author believes to be of importance (Coiro, 2011; Hahnel et al., 2018; Kang, 2014). The ability to critically evaluate information ensures a user can determine that information, and its source, are both reliable and accurate, and to recognise bias (Kang, 2014; Leu, Kinzer, Coiro, Castek, & Henry, 2017). This is an important aspect of online reading and places a greater emphasis on being able to cross-examine content across multiple sources to ensure both the content and the source retain their validity (Clinton, 2019; Park, Yang, & Hsieh, 2014).

Leu et al. (2017) highlighted the difficulties in differentiating promotional and advertising efforts on the Internet and the increased challenges that unedited information and the merger of advertising and educational content have caused. Judgement of the credibility of online information is a factor that has been explored quite extensively in information seeking studies, both in terms of web documents and search engine results (Kattenbeck & Elsweiler, 2019; Schwarz & Morris, 2011). This places further emphasis on the development of composite skills to ensure the identification, selection and judgement of relevant web documents (Hahnel et al., 2018; Kattenbeck & Elsweiler, 2019).

However, as highlighted by Clinton (2019), this does little for those who are poor readers. It therefore becomes essential that authors of online content do not assume user ability. This causes concerns when considering the rise of misinformation (Ruokolainen & Widén, 2020), especially when users are using low-quality online resources for (what could be) vital tasks in e-governmental contexts (Brazier & Harvey, 2018). When such services are being moved solely online this puts additional import on these skills, especially for ESL speakers (amongst numerous other groups at risk of the digital divide (Lloyd, 2020; Oduntan & Ruthven, 2019; Selwyn & Facer, 2007)). The current COVID-19 pandemic, and the attendant need for health information and the risks of poor access, understanding or overload of information, is a perfect example (Khan, Asif, & Jaffery, 2020; Soroya, Farooq, Mahmood, Isoaho, & Zara, 2021).

2.4.2. Measuring document reading level

The UK Government Digital Service team, which manages and maintains the UK Government's online presence, has documented the importance of reading level complexity as a means to ensure accessibility of information for the general public, where the average reading age is reported as 9 years old (Cawthorne & Barnes, 2016). It is highlighted that using reading level formulae is imperative to gauge whether content is suitable for consumption before staff can release or post documents. While this applies to national government content, the same cannot be said of non-governmental content, including that produced by local councils (Leu et al., 2017).

The importance of the readability of online documents has been shown in the broad works of authors such as Collins-Thompson and colleagues in the field of information science (Collins-Thompson, Bennett, White, De La Chica, & Sontag, 2011; Collins-Thompson & Callan, 2005) and the plethora of reading level formulae that have been developed since the turn of the 20th Century (DuBay, 2004). While the use of these formulae for web documents has been called into question (Benjamin, 2012; Collins-Thompson & Callan, 2005), they are still heavily used, can act as effective proxies to "true" reading levels (Ojha, Ismail, & Kuppusamy, 2018), and can be used effectively to re-rank documents for different language abilities (Collins-Thompson et al., 2011).

Such formulae typically consist of two key components: a measure of syntactic complexity and a measure of word-level complexity. Syntactic complexity is often determined by measuring the average sentence length, the idea being that shorter sentences are typically easier to interpret. Many approaches, including Flesch Reading Ease, Flesch Kincaid (Flesch, 1948), LIX (Björnsson, 1983), and the Gunning Fog index (Gunning et al., 1968), use average word length or number of syllables to approximate word-level complexity. However, this may not be very effective as shorter words are not necessarily easier than longer ones (e.g., consider "uninteresting" and "insipid"). More successful approaches rely on lists of words determined to be "difficult" and consider the ratio of words within a document that feature in such a list to be a good measure of word-level complexity.

A frequently-used formula that takes such an approach is the new Dale–Chall reading formula (Chall & Dale, 1995), which has been previously shown to be effective for short web documents (e.g., Collins-Thompson et al. (2011), Pancer, Chandler, Poole, and Noseworthy (2019)). Many of such measures were developed primarily to assess reading levels for school-age children and work has questioned their suitability in the context of adult users or non-native language learners (e.g., Uitdenbogerd (2005)). However, analysis by Chall and Dale (1995) of the updated Dale–Chall measure, which we use here, has shown its utility in evaluating documents intended for college and graduate students and for air force academy trainees. The analysis by Chall and Dale (1995) also demonstrates a high level of intra-measure correlations, suggesting that they are all generally measuring similar elements of readability.

Furthermore, work by Greenfield (2004) has demonstrated the suitability of non ESL-specific readability formulas (including the new Dale–Chall) for assessing readability for language learners. The authors conclude that "[The] findings support the conclusion that the classic formulas are indeed fundamentally valid for a broad spectrum of English readers that includes non-native as well as native readers"(Greenfield, 2004, pg.11). The suitability of the new Dale–Chall formula, and others that use high-frequency (or "easily acquired") word lists, is also supported by work that has shown that word frequency is a strong predictor of word-level reading complexity for L2 learners (Chen & Truscott, 2010; Koirala, 2015). This same work also demonstrates that simpler measures of word-level complexity, such as the word character length, do not reliably predict difficulty for L2 learners.

2.5. Research questions

Following the above review of relevant literature, and our overall aim as outlined in the introduction (i.e., to investigate the search behaviours and performance of both native and non-native speakers on e-government-related tasks, and the effect of document readability/complexity), our Research Questions (RQs) are as follows:

1. What is the effect of English language proficiency (between three groups of varying levels) on perceptions of task relevance and clarity and self-reported performance and task engagement, and on objective search performance?
2. What is the effect of English language proficiency (between three groups of varying levels) on time on task?
3. What is the effect of English language proficiency (between three groups of varying levels) on the reading level of bookmarked documents?
4. Does the reading level of bookmarked documents have any effect on the relevance of the same?
5. How does English language proficiency (between three groups of varying levels) affect the reading levels of bookmarked documents and relevance of the same?

3. Method

3.1. Prior work

This study builds and significantly expands upon prior work by the authors (Brazier & Harvey, 2017b, 2018) through expansion of data collection tools, metrics and participant population to widen the scope for comparison and generalisability of the work this study complements and extends. It uses a web browser-based user interaction logging tool to capture participants' behaviours when searching for documents relevant to four context-relevant search tasks and includes an extra group to represent a different population (i.e., less proficient ESL speakers). The logs captured the documents users viewed and, ultimately, bookmarked allowing us to download them and computationally analyse their reading complexity.

3.2. Data collection tool

Attempts were made to utilise pre-existing digital tools to log user interactions, however, such systems were either not available for use (Vuong, Jacucci, & Ruotsalo, 2017), no longer maintained (Weth & Hauswirth, 2013), or not entirely fit for purpose. Prior work had recorded such interactions manually by recording studies using Morae Manager and tagging interactions to calculate metrics. Concerns around the accuracy due to human error, despite attempts to mitigate for these, meant an alternative was required. Previous work had identified Chrome as a well known and well used web browser among both ESL and English native speakers. As such, a bespoke (custom) Google Chrome extension with a web interface was developed to record user interactions with the browser and online documents, storing the data in local storage. Upon completion of the study a logfile was generated locally, and downloaded as a .csv file.

The web interface (Fig. 1) provided not only logging functionality but also presented to participants the study information sheet, consent form, questionnaires, the means to check current task descriptions, task time and the ability to end the task (Fig. 2) when desired.

The extension recorded all consent form and questionnaire data; search engine query terms typed and submitted; the Search Engine results page (SERP) links including their ranks, snippets and whether there were any adverts present. It also recorded all open browser tabs and windows, and whether they were active. Bookmark interactions, including whether they were saved or deleted were also included. Each interaction was recorded along with a date and timestamp. The web interface provided a timer for users, so they were able to identify remaining time for each task. Tasks were a maximum of 10 min and hardcoded into the web interface, although participants were provided the opportunity to end the task early if they felt they had a sufficient number of documents to complete the task. Participants were given up to 5 additional minutes to read the task description and complete pre- and post-task questionnaires, resulting in the experiment taking no more than one hour in total.

Tasks were distributed to participants via the web interface and extension using a Latin square design to mitigate against ordering and potential learning effects (Kelly et al., 2009; Peters, Braschler, & Clough, 2012).

3.3. Textual analysis

We wrote bespoke Java code to systematically download the content of each of the bookmarked URLs, parse these to extract the raw text content, and calculate reading levels scores. We used Apache Tika[1] to parse PDF documents into raw text and Java Boilerpipe[2] to strip HTML and any extraneous "boilerplate" clutter (e.g., ads, hyperlink lists, navigation, etc.) from the downloaded files. We then processed the resulting raw content text using the Java Fathom library[3] to calculate the reading level scores (see Section 4.5 for more details).

[1] https://tika.apache.org/.

[2] https://mvnrepository.com/artifact/de.l3s.boilerpipe/boilerpipe.

[3] http://freshmeat.sourceforge.net/projects/java-fathom.

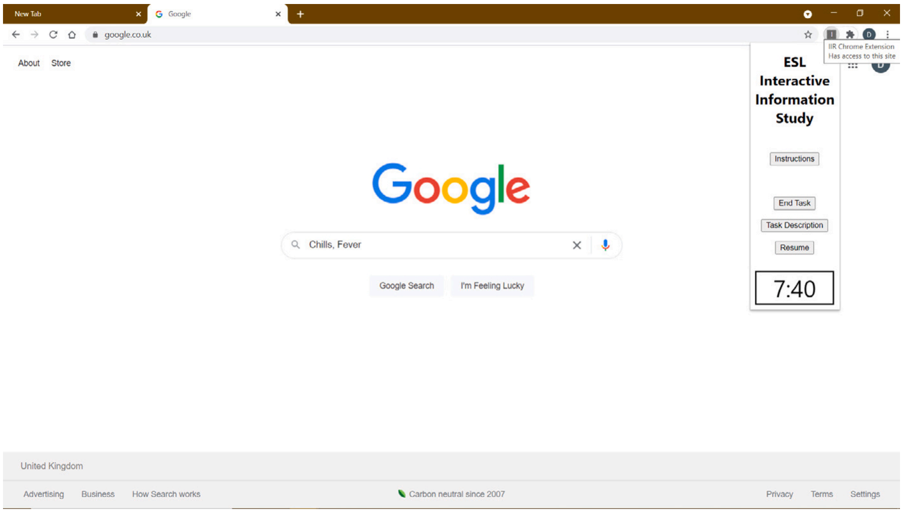
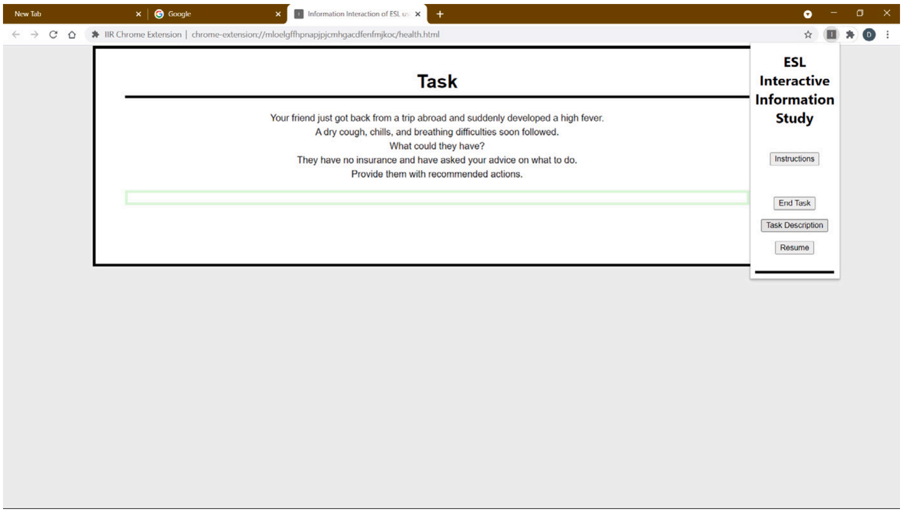


Fig. 1. Chrome extension in-study menu view.


New Tab
Google
Information Interaction of ESL u
IIR Chrome Extension | chrome-extension://mloelgffhpnapjpjcmhgacdfenfmjkoc/health.html
ESL Interactive Information Study
Instructions
End Task
Task Description
Resume

Task

Your friend just got back from a trip abroad and suddenly developed a high fever.
A dry cough, chills, and breathing difficulties soon followed.
What could they have?
They have no insurance and have asked your advice on what to do.
Provide them with recommended actions.

Fig. 2. Chrome extension in-study task description.

Out of 931 unique bookmarked URLs, we were able to download and parse the content from 879, meaning that we achieved 94.4% coverage. Failure to retrieve and/or parse the original URL content was either due to the link being broken – presumably the content had been re/moved in the three months since the study had been conducted – or because the file was a PDF and the Tika parser was unable to extract any textual content.

3.4. Process

The study was conducted in a lab at a large UK University, with all participants using a desktop computer. All instructions and documentation were in English and had been written in collaboration with an English language teacher in the University to ensure they would be readily understandable to those possessing proficiency at the lower limits of the University's requirements. Upon arrival participants were asked to carefully read the study instructions and for use of the extension, and were asked to verbally confirm that they had understood these instructions. They were then required to complete the consent form and complete a pre-study demographic questionnaire. Following this they were allocated a task scenario, for which they could read the description and had to fill in a pre-task questionnaire to gauge their domain knowledge, interest in the topic and the perceived difficulty of the task using five-point Likert scales (see Table 1).

Participants were asked to start at *google.co.uk* to begin their search for relevant documents/sources, and to bookmark within the browser any document they deemed relevant as they went. When participants felt they had sufficient relevant documents, or

Table 1
Pre-task questions.

Pre-Q1	I have searched about this topic before.
Pre-Q2	I know about this topic.
Pre-Q3	I am interested in this topic.
Pre-Q4	It will be difficult to find information about this topic.

Table 2
Post-task questions.

Q1	I was given enough information to complete the task.
Q2	It was clear what was being asked.
Q3	The task was relevant to me.
Q4	The task was easy to understand.
Q5	I was engaged in the task
Q6	I performed the task to the best of my ability
Q7	I found the task difficult
Q8	I am confident the content I found satisfied the task
Q9	I am confident about the search query terms I used.
Q10	I am confident I identified relevant websites
Q11	I am confident in my ability to read the website content
Q12	I am confident in my ability to understand the content of the websites I visited
Q13	I am confident the search task was completed.

the timer reached the 10 min limit, the task ended and participants completed a post-task questionnaire (again using 5-point Likert scales), as seen in Table 2. This process was repeated until all four tasks had been completed, at which time the system thanked the participant and the final interaction log file could be downloaded. Tasks were presented to participants in a pseudo-random order to mitigate against any potential order effects.

3.5. Metrics

To determine relevance judgements, all bookmarks were assessed by the authors (two native English-speaking IR researchers) using a voting strategy – any bookmarks not given the same score were discussed and a single score was agreed – and given scores on a 4-point scale, where 1 is not relevant, 2 is tangentially relevant, 3 is partially relevant and 4 is relevant. Query classification were determined in line with Chu et al. (2012), with the added category of ‘repeat’ as some queries were resubmitted without amendment and determined by the same researchers. Inter-rater agreement was high and is discussed further in the analysis section below. In order to avoid any potential for bias, all bookmarks made by all of the participants were aggregated together on a per-topic basis, any duplicates were removed, and then each unique bookmark was assessed for relevance. In this manner there was no way for the assessors to tell who had bookmarked each document or what group they belonged to.

3.6. Task formulation

Although some existing research has considered information seeking behaviours of non-native English speakers for e-government tasks, these either pre-date the digital by default initiative, are purely qualitative in nature, are literature reviews or are not based on UK e-government services (Dwivedi & Williams, 2008; Freund, 2013; Kolsaker & Lee-Kelley, 2008). Although the findings of these studies are likely to be of some relevance in the UK context, it is difficult to know how cultural influences affect people’s behaviours and interaction, and none have directly considered the effect of document reading level. Some research utilises the TREC GOV2 collection (Clarke, Craswell, & Soboroff, 2004); however, the documents within this collection are from a US government perspective, considerably pre-date the digital by default initiative and also pose problems with task engagement and performance, in terms of interest and relevance to participants (Borlund, 2013). To address concerns about task context, and to enable users to engage in system development through recommendations, we adopted a participatory design approach by building on the work of Borlund (2003) and eliciting needs from the participants themselves. To permit generalisability and quantitative approaches to data collection and analysis, this project used information needs previously elicited from a study with a group of 7 international Ph.D. students from a UK university, adapted into search tasks. The full methodology and findings from this synthesis of tasks is currently being drafted for future publication.

The search tasks were defined based on the results of the previously mentioned short study and were designed to reflect actual information seeking situations in an attempt to be relevant, UK specific, and a more interesting search experience for the participants (Edwards & Kelly, 2016). We note that the work of Freund (2013) also establishes suitable task types in the e-government context, and while the way tasks are formulated in this work differ, the general task types do align. The tasks developed for and used in this work are:

1. *Task 1. Your friend from Peru and their family (2 members) are coming to visit you for 6 months while you are in the UK. Develop a list of instructions to help them apply for the necessary visas.*

2. *Task 2. A family member is coming to the UK to live and wants information on housing. They have heard there are a number of options and have asked you for advice. Identify the options available to them and recommend which they should choose. Give reasons to support your recommendation.*
3. *Task 3. Your friend just got back from a trip abroad and suddenly developed a high fever. A dry cough, chills, and breathing difficulties soon followed. What could they have? They have no insurance and have asked your advice on what to do. Provide them with recommended actions.*
4. *Task 4. Your elderly neighbours have heard about the UK government's 'digital by default' initiative and are concerned about whether this will affect them and their friends at the local community centre. They have asked you to find out more about it. Use your best judgement to highlight what would impact them with reasons for your choices.*

3.7. Participants

A combination of techniques were utilised to recruit participants including posters (also posted online[4]), which could be accessed by Quick Response (QR) code or directly from each study's unique URL. Email and face-to-face recruitment was also employed to maximise exposure of the study, including invited talks during a summer school language programme at the same UK University. It is difficult to determine the success of this method in terms of conversion rate as recipients of the emails were sent the adverts directing them to the online sign up form. All three adverts were viewed over 1200 times[5] with some 20 participants registered through this platform. The remainder of the participants signed up either in person or via a direct email to the researchers.

The total sample population included 42 participants across 23 separate programmes, which is larger than or comparable to similar studies (e.g., Chu et al. (2012), Freund (2013) and Liu et al. (2019)). These were composed of three different groups: 12 native speakers of English (*nat*); 17 proficient ESL speakers (*esl-pg*); and 13 less proficient ESL speakers (*esl-ug*). Of the 42, 18 participants identified as female, 7 (n = 17) *esl-pg*, 6 (n = 13) *esl-ug* and 5 (n = 12) *esl-pg*. The overall mean average age was 28.119 (SD 6.634), with a range of 38 (19–57 years). The *esl-ug* group were younger on average at 22.615 years (SD 2.292), with the *esl-pg* and *nat* closely aligned at 30.392 (SD 6.450) and 31.818 (SD 9.185) respectively. There were a total of 19 different nationalities across Asia, Africa, Europe and North America, and between them participants reported speaking 21 languages other than English to some degree. Mandarin Chinese was the most prominent non-English spoken language (n = 9), followed by Arabic and Spanish (n = 5), Thai and German (n = 3), and French and Italian (n = 2). Other languages with a single speaker each included: Hausa, Hindi, Kurdish, Igbo, Italian, Malay, Portuguese, Turkish, Urdu and Vietnamese. We note that with the exception of Spanish, German, French and Italian, all other languages are either unrelated or only very distantly related to English.

The *esl-ug* group were all enrolled in English remedial classes provided by the University's summer language school to prepare them for academic writing – and life – in the UK. Although they were required to complete the course to improve their English proficiency before starting their University degree programmes, the University sets a minimum requirement of 4.5 on the IELTS[6] scale in order to be admitted. Those scoring between 4.5 and 6 must then take the remedial classes. The first two groups were composed of Ph.D. students, while the third group were all undergraduate students. Previous work has found the performance of fluent ESL speakers to be similar to native speakers (e.g., Brazier and Harvey (2017a, 2018)); however, these studies did not include less fluent users of English. The *esl-ug* group represents such users, i.e., those who have relatively little fluency in the language but who need to use it in their everyday lives.

Prior to completing the tasks, we asked participants to rate their English proficiency from 1 (native) to 5 (beginner). All members of the *nat* group stated they were native; all but one of those in the *esl-pg* group indicated they had good proficiency (2); and the median response for the *esl-ug* group was 4, indicating limited proficiency. We did not ask participants for their individual IELTS scores, although lower bounds for the two non-native groups can be deduced from the University's minimum entry requirements. The *esl-ug* participants will have a minimum score of 4.5, while the *esl-pg* group will have a minimum score of 6.5. All three groups indicated similar high frequency of use of information technology and search engines in their everyday lives. There were, however, significant differences reported in terms of experience in using English-language search engines (F(2,39) = 169.21, $p \ll 0.01$, $\eta^2 = 0.9$): the *esl-ug* reported significantly less experience than both the *esl-pg* and *nat* groups, while there was no significant difference between the *esl-pg* and *nat* groups.

4. Analysis

This section outlines the findings of the research including participant data, task time, reading level, and anecdotal researcher response on participant behaviours from the data collection sessions.

[4] www.callforparticipants.com.

[5] Registered members of the callforparticipants website also had access to the advert.

[6] International English Language Testing System (IELTS) is an international standardised test of English proficiency for non-native English language speakers.

Table 3
Pre-task questionnaire responses.

Group	Pre-Q1	Pre-Q2	Pre-Q3	Pre-Q4
nat	1.71	1.93	2.77	2.36
esl-pg	2.02	2.38	3.07	2.37
esl-ug	1.72	2.04	2.81	2.34

4.1. Pre-task perceptions

When questioned pre-task about their prior experiences searching about (Pre-Q1), knowledge of (Pre-Q2), interest in (Pre-Q3), and perception of task difficulty (Pre-Q4) the *esl-ug* group were approximately identical to the native group in their average results (See Table 3). It is interesting to note that the *esl-pg* group had more prior experience searching about the topics (and, consequently, more knowledge about them). This is perhaps unsurprising given their relatively recent arrival in the country and the requisite need to complete government forms and processes around immigration. Interest was generally quite high, although this did vary somewhat across the individual topics. Overall, the *esl-ug* group were most interested in the housing task—likely due to nature of the task and that these participants had only been in the UK for a short period prior to the study taking place, and likely going through similar processes or having completed such tasks themselves.

The *esl-ug* group's perceptions of difficulty in finding information (Pre-Q4) is especially of interest as this group's experience in using search engines in English was significantly lower than the other groups as was their English proficiency. Despite this, all 3 groups had similar expectations of how difficult the tasks would be (F(2875) = 0.067, $p = 0.936$), indicating a discord between perceived experience/ability and perceived difficulty in performing tasks requiring ability in the same.

4.2. Post-task perceptions

We asked participants a number of questions after completing each task regarding perceived task clarity, relevance and how easy it was to understand what was expected of them. Encouragingly, there were no differences between the groups regarding clarity (Q2 "It was clear what was being asked"; F(2875) = 1.16, $p = 0.314$), and although all groups generally found the tasks easy to understand, those in the *esl-ug* group did find the tasks somewhat harder to understand (Q4 "The task was easy to understand"; $\mu_{nat} = 4.17$, $\mu_{esl-pg} = 4.18$, $\mu_{esl-ug} = 3.61$; F(2875) = 30.04, $p \ll 0.01$, $\eta^2 = 0.06$). Curiously, there was significant variation between the *esl-pg* group and the other two groups regarding having enough information to complete the tasks (Q1 "I was given enough information to complete the task"; F(2875) = 25.29, $p \ll 0.01$, $\eta^2 = 0.05$), although again all groups generally felt they had sufficient information. Interestingly, a few of the *esl-ug* participants (but not those from the other groups) were anecdotally observed using their mobile devices to translate terms they either did not recognise or did not understand well. These were either terms used in the task descriptions or terms for keyword searching but not for other activities, such as search results or document reading purposes.

Overall perception of task relevance was positive, with all groups judging tasks 1, 2 and 3 partially relevant or higher. Task 4 was perceived as slightly less relevant, particularly among the *esl-ug* participants. This is, again, likely due to the nature of the task and the fact that these participants had only spent a few weeks in the UK at the time of the study taking place—and unlikely to be integrated with their local community. An ANOVA between post task difficulty and participant group identified a significant difference (Q7 "I found the task difficult"; F(2875) = 8.997, $p \ll 0.01$, $\eta^2 = 0.02$), with the *esl-ug* group finding the tasks more difficult than the other two groups, although we note that the effect size here is small. This was mirrored by the results on questions regarding task engagement (Q5 "I was engaged in the task"; F(2875) = 171.29, $p \ll 0.01$, $\eta^2 = 0.28$) and the extent to which participants felt they had performed to the best of their abilities (Q6 "I performed the task to the best of my ability"; F(2875) = 43.88, $p \ll 0.01$, $\eta^2 = 0.09$).

The *esl-ug* participants were also significantly less confident in their post-task performance when reflecting whether the content they found satisfied the task (F(2875) = 63.52, $p \ll 0.01$, $\eta^2 = 0.13$), the search queries used were good (F(2875) = 65.85, $p \ll 0.01$, $\eta^2 = 0.13$), identifying relevant websites from the SERP (F(2875) = 76.86, $p \ll 0.01$, $\eta^2 = 0.15$), ability to read the website content (F(2875) = 211.95, $p \ll 0.01$, $\eta^2 = 0.33$), ability to understand the content (F(2875) = 227.96, $p \ll 0.01$, $\eta^2 = 0.34$) and that the task was complete (F(2875) = 31.43, $p \ll 0.01$, $\eta^2 = 0.07$). For all of these questions, post-hoc pairwise t-tests with Bonferroni correction showed significant differences between *esl-ug* and the other two groups but not between the other two groups.

4.3. Time on task

We calculated the participants' task times (for each of the four tasks) based on the log data (see Table 4). Interestingly, as with the perception-based measures analysed above, there was very little difference between the *esl-pg* and *nat* groups' task times, whereas some noticeable differences can be seen between the *esl-ug* and other groups. An ANOVA between task time and participant group identified a significant difference (F(2164) = 22.325, $p \ll 0.01$, $\eta^2 = 0.21$). Post-hoc pairwise comparisons with Bonferroni correction showed significant differences between the *nat* and *esl-ug* ($\mu_{nat} - \mu_{esl-ug} = 173.84s; p \ll 0.01$) groups and between the *esl-pg* and *esl-ug* groups ($\mu_{esl-pg} - \mu_{esl-ug} = 164s; p \ll 0.01$): participants in the *esl-ug* group spent significantly less time on the tasks than those in the other two groups. We note also that, although the differences are not significant, the *nat* group tended to spend more time on tasks than the *esl-pg* group. We also note that we did not immediately force a participant's browser to go to the post-task questionnaire

Table 4
Median task times by topic and group.

Topic	esl-ug	esl-pg	nat
1	447	545	601
2	356	583	606
3	464	536	606.5
4	339	605	604

Table 5
Document relevance scores (i.e., search performance) by group.

Measure	esl-ug	esl-pg	nat
Ratio relevant (binary)	0.443	0.661	0.738
Mean score (4-item)	2.630	3.040	3.158

Table 6
Descriptive statistics of interaction between group and relevance on Dale–Chall score.

Group	Relevance	Mean	Std. Deviation	N
nat	rel	7.28	2.66	168
	non-rel	6.52	2.64	66
esl-ug	rel	6.57	1.49	100
	non-rel	7.40	2.15	116
esl-pg	rel	6.69	2.36	272
	non-rel	6.56	2.69	156

when the time limit was reached. Instead the system spawned a modal pop-up dialog with a button to open the questionnaire form. The small number of seconds elapsed after the 600 s limit in some cells of Table 4 is merely the reaction time of the participant to the dialog and so these indicate that these participants used all of the allotted time.

4.4. Document relevance

As mentioned earlier, relevance judgements were conducted independently by two experienced IR researchers, who had a high level of initial agreement ($\alpha = 0.98$ for binary relevance; $\alpha = 0.83$ for 4-item relevance). The small number of documents for which there were disagreements were discussed before coming up with a single judgement. Using the 4-item relevance scale, 25 bookmarked documents were judged to be completely non-relevant, 313 were tangentially relevant, 203 partially relevant and 337 were judged to be relevant to the information need. Converting these into binary relevance scores results in 338 non-relevant documents and 540 relevant ones. Average relevance of documents bookmarked for each task could be used as a measure of task difficulty. Making this assumption, we find significant differences in the difficulty of tasks ($\chi^2(9) = 169.2$, $p \ll 0.01$): documents bookmarked for task 1 had a mean relevance score of 3.48, while those for task 4 obtained an average relevance score of only 2.53.

To obtain a score for each user in terms of their task performance, we considered all of the documents they bookmarked over the four tasks and calculated the ratio of those documents that were deemed to be relevant. The resulting scores were normally distributed with a mean score of 0.614 (i.e. 61% of documents bookmarked by an "average" participant were judged to be relevant) and a standard deviation of 0.18. The minimum score attained by a single user was 0.267 and one user's 8 bookmarks were all judged to be relevant. An ANOVA showed statistically significant differences in score by participant group (F(2,38) = 15.52; $p \ll 0.01$, $\eta^2 = 0.45$). Post hoc pairwise t-tests with Bonferroni correction showed that there was a significant difference between the *nat* and *esl-ug* groups (t = 5.21; $p \ll 0.01$; $\mu_{nat} - \mu_{esl-ug} = 0.295$) and also between the *esl-pg* and *esl-ug* groups (t = 4.23; $p \ll 0.01$; $\mu_{esl-pg} - \mu_{esl-ug} = 0.218$). This demonstrates that the *esl-ug* group did indeed perform significantly worse than the other two groups and less than half of the documents the group bookmarked as being relevant actually were ($\mu_{esl-ug} = 0.44$). Although the difference was not significant (t = 1.44; $p \ll 0.16$;), the *nat* group did perform better than the *esl-pg* group on average ($\mu_{esl-pg} = 0.66$, $\mu_{nat} = 0.74$, $d = 0.382$).

We also considered the 4-item relevance scores over the different groups, which displayed a similar pattern: there were significant differences between the *esl-ug* group and the other two groups and the *nat* group performed best overall (see Table 5).

4.5. The effect of reading level

In order to estimate the reading level (complexity/difficulty) of bookmarked documents, we used the Dale–Chall readability score (Chall & Dale, 1995), which is based on the average length of sentences and ratio of "difficult words" within a given document. Difficult words are defined to be any words not contained within a list of 3000 words that groups of fourth-grade American students could reliably understand. We downloaded the web pages or PDF documents bookmarked by participants, some of which we could not obtain (see Method section above for details) and, after stripping out any HTML or PDF metadata tags, calculated the Dale–Chall score for each. Out of a total of 931 bookmarked documents, we were able to retrieve and parse 878 documents (i.e. 94.3%). The

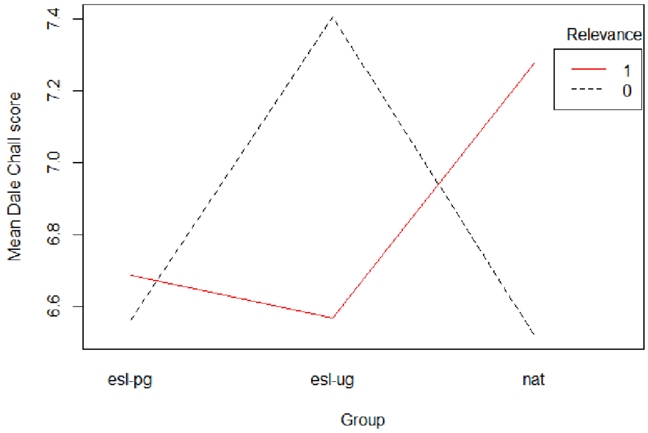


Fig. 3. Interaction plot of mean Dale–Chall score by participant group and document relevance.

Table 7
Results of two-way ANOVA test.

Source	DF	Sum Sq.	Mean Sq.	F	Pr. (>F)
Group	2	35.6	17.797	3.084	0.046
relevance	1	0.1	0.055	0.01	0.922
group*binary	2	66.8	33.391	5.784	0.003
Residuals	872	5032.8	5.772		

Dale–Chall scores were approximately normally distributed with a mean of 6.85 – indicating a reading level appropriate for an 8th grader – and a standard distribution of 2.42.

An ANOVA identified significant differences between the groups in terms of the Dale–Chall scores of the documents they bookmarked (F(2875) = 3.054, p = 0.048, $\eta^2 = 0.007$), although the effect size is small and post-hoc analysis did not identify any significant pairwise differences between the groups. However, when considering the relevance of the documents as an additional factor, there were clearer differences. Further investigation identified an interesting, and significant, interaction effect on the Dale–Chall scores by participant group and relevance between the *nat* and *esl-ug* groups (F(1872) = 5.79, p = 0.003, $\eta^2 = 0.01$). As shown in Fig. 3, Tables 6 and 7, the documents bookmarked by participants in the *esl-ug* group that were deemed to be relevant had significantly lower Dale–Chall scores than the ones that were non-relevant. Conversely, for the *nat* group, their relevant bookmarked documents had significantly higher Dale–Chall scores than the non-relevant ones they selected.

Note that automated reading formulae are not necessarily well suited to evaluating web documents, which are often written in a succinct manner using non-standard formatting. A histogram analysis of the number of words per document revealed a number of outlier documents that either contained 0 words or in excess of 30,000 words. These documents were found to be non-html documents, such as PDF manuals, javascript-based dynamic webpages or heavily image-based pages. When removing documents of less than 10 words, as a proxy for more standard formatted web documents, it was clear that the previously-identified interaction effect between reading level and relevance among the groups continued, albeit again with a small effect size (F(1872) = 5.04, p = 0.007, $\eta^2 = 0.01$).

4.6. Predicting success using reading level and group

To further verify the relationship identified between reading level, participant group (as a proxy of English language proficiency) and success in terms of bookmarking relevant documents, we attempted to predict (binary) success using the other two variables as IVs. We split the data into training and testing sets with a ratio of 75:25 using stratified sampling to ensure consistent distributions across the DV. We scaled all Dale–Chall values and trained both Logistic Regression and Support Vector Machine models and evaluated performance on the testing set. We also evaluated the performance of a "baseline" model, which always predicts the majority class (i.e., '1' in this case).

Results of these experiments are shown in Table 8 and suggest that, although success is very difficult to predict, both the group and Dale–Chall features do help to increase prediction accuracy compared to the majority class baseline. It is surprising that these two features alone are able to provide up to a 6% improvement in success prediction, despite knowing nothing else about the search session or about the searcher. We do, however note the generally low specificity scores, which can only be partially explained by the class imbalance in the testing data, and which is only partially mitigated by using higher levels of smoothing (i.e., higher γ values in the RBF kernel).

Table 8
Results of classification experiments. Acc. = accuracy. All SVM models used the Radial Basis Function (RBF) kernel.

Model	Acc.	Sensitivity	Specificity	Balanced Acc.
Baseline	0.616	**1**	0	0.5
LR	0.635	0.742	**0.375**	0.564
SVM ($\gamma = 0.01$)	0.626	0.963	0.083	0.523
SVM ($\gamma = 0.1$)	**0.653**	0.919	0.226	0.572
SVM ($\gamma = 1$)	0.6393	0.867	0.274	**0.583**
SVM ($\gamma = 10$)	0.621	0.859	0.238	0.547

5. Discussion & conclusions

5.1. Key contributions

This study provides several interesting insights into the information behaviours of ESL speakers when conducting e-governmental topic search tasks and explores the influence of document reading level. Our work builds on a series of existing studies by the likes of Bogers et al. (2016), Brazier and Harvey (2017a, 2017b, 2018) and Chu et al. (2012) by considering how language proficiency affects search behaviour and performance. However, we expand upon these works in 3 significant ways: by considering 3 different groups of users (i.e., natives, ESL speakers with high proficiency, and ESL speakers with low proficiency); by investigating search through the important – and for ESL speakers often highly relevant – lens of e-government services; and by considering the effect of document reading level on the observed behaviour and performance. Although not wholly generalisable, our findings can support future research in this area and the development of online content and e-government systems.

5.2. English language proficiency (RQ 1)

By including ESL speakers with poorer English language skills, we have obtained quite different results to those found of Bogers et al. (2016) and Brazier and Harvey (2018). They found relatively few differences between their ESL and native groups, which was also broadly the case in our data when comparing the more proficient ESL group (i.e., *esl-pg*) with the native speakers. However, we found many significant differences between the less proficient ESL group (*esl-ug*) and the other two groups, particularly in terms of search performance/literacies (i.e., identifying relevant documents), time on task, as well as self-perceived performance and confidence.

This suggests that, once one has attained a certain level of proficiency in a second language, there is little to distinguish one from a native speaker in terms of search behaviour and performance but that having a lower level of proficiency can result in very large differences. This may be complementary to, but is not directly comparable with, the work of Kang (2014), as they considered very different education levels; language ability aside, we would not expect the differences in general knowledge between undergraduate and postgraduate students to account for the differences we observed. Also, given the nature of the search tasks used, which participants generally knew little about prior to the experiments, it is unlikely that group-level differences in domain or topic knowledge were a confounding effect (Arguello et al., 2018; White et al., 2009).

These findings have implications for both interface design and the writing of documents for public consumption, particularly in the e-government context, where ESLs are frequent users and where finding, identifying (and understanding) relevant documents can be vitally important (Józsa et al., 2012). It would, for example, be beneficial to offer multiple versions of documents written to aim at different comprehension levels and an interface that helps users to identify the versions most beneficial to them (Collins-Thompson et al., 2011). This would certainly be less costly for governments and local authorities than offering documents in multiple different languages to cater for various immigrant groups as translators would not be required (Alam & Imran, 2015).

5.3. Time on task (RQ 2)

Regarding time on task, the results directly contradict the findings of Bogers et al. (2016) - and to a certain extent, those of Rózsa et al. (2015): the *esl-ug* group spent significantly less time on all 4 tasks than either of the other two groups. This contradiction may again be due to the earlier works studying exclusively ESL speakers with high levels of English proficiency. It may be that our *esl-ug* users' lower levels of familiarity with the language meant that they missed or misinterpreted vital signs of information scent, causing them to more quickly give up on tasks rather than making the effort to investigate further (Józsa et al., 2012; Kralisch & Berendt, 2005). This suggests the need to assist such users further in exploring and understanding document and SERP content, perhaps by integrating dictionaries and glossaries for less common terms. Our earlier suggestion of having multiple versions of documents aimed at varying comprehension levels would also help users to more easily identify relevant documents and make them less likely to abandon their searches early.

5.4. The impact of reading level (RQs 3, 4 & 5)

The findings empirically support the supposition of Aham-Anyanwu and Li (2017) that document complexity may have a significant impact on user engagement with and use of e-government resources. They also serve to corroborate the findings of Hahnel et al. (2018) that reading skills support information processing strategies and the ability of a user to identify relevant documents from a SERP. Our results demonstrate that this finding applies also to adult searchers and that automated document reading level metrics can successfully be applied in this context. The key insight was that the *esl-ug* group were much less likely to identify relevant documents if those documents had a high reading level, while native speakers, conversely, were more likely to bookmark (identify as relevant) more complex documents. This significant interaction effect also suggests that less proficient speakers of English can identify relevant documents, but only when they are written in more understandable language and that native speakers tend to display the opposite behaviour—they are more likely to correctly identify documents as relevant if they are written at a higher level of complexity.

This has obvious implications for ESL speakers but also for the native speakers. It is possible that the native speakers tended to ignore documents that were actually relevant if they were written at a lower complexity level, perhaps indicating a lack of trust and assuming that “simpler” content is less likely to be accurate or useful. It may also be possible to develop interfaces that initially provide only basic summaries of document content to less proficient users but surface more detailed and “complex” document versions for proficient or experienced users.

Furthermore, our findings when using reading level and searcher language proficiency to predict search success could be used to augment existing approaches, which typically rely on much more fine-grained session-level data (e.g., Guo, Lagun, and Agichtein (2012)).

5.5. Limitations

While this work was designed to mitigate a number of risks to internal and external validity, it is not without its limitations. The sample sizes of the study groups, while reasonable, are lower than would be ideal to be able to offer generalisable outcomes. This may mean that some of our statistical analyses are underpowered—there may be additional effects between the groups that would be significant given larger samples sizes. We also note that our research was conducted in a purely UK context and, although we would expect many of the effects to be universal, it is possible that some are due to UK-specific factors.

Future research would greatly benefit from a wider pool of participants, both the native English speakers and ESL speakers of varying proficiency and origin, as well as considering whether the results translate to non-UK contexts. The impact of the lab-based environment on participants and the subsequent impact on internal validity should not be understated. Lab type conditions, artificial task allocations and time limits have anecdotally (and empirically Liu et al., 2019) been highlighted as contributing factors to some behaviours, such as using mobile phone devices to translate task descriptions or keyword terms before submission. Efforts were made to support participants if they did not understand any tasks they were given, and it must be acknowledged that this may well have influenced results.

Although we did try to reduce some of the effects caused by the predefined and prescriptive nature of the tasks (e.g., by co-designing tasks with ESL speakers, and so they were deemed relevant or partially relevant by all participant groups), the contrived nature of such an experiment cannot be entirely obviated. These issues were also partially mitigated by, for example: using Google as the online search platform (selected as the most frequently used online search tool in the pre-study questionnaire); and including time constraints as a comfort factor to avoid fatigue-related effects. Future work will factor these issues and the use of the data collection platform (from this work) will be adapted to focus on remote studies of ESL information search behaviours.

As outlined in Sections 3.3 and 2.4.2, the use of certain reading formulae to analyse web documents is not ideal in certain cases, and while efforts were made to mitigate for these, there are other valid and reliable methods for eliciting reading level of web documents (Collins-Thompson & Callan, 2004). We also recognise that a measure designed with native speakers of English in mind may not be optimal in the case of non-natives or language learners. However, as we are attempting to compare performance of both L1 and L2 speakers, yet need to choose a single readability measure, there will always be a tension as to whether to choose one specifically developed for L1 or L2 speakers. We also note that attempts to produce measures better suited for learners (e.g., Uitdenbogerd (2005)) introduce additional complexity whilst only achieving marginally better prediction performance. Uitdenbogerd (2005)’s results suggest, in fact, that average sentence length alone performed better than standard readability measures in their context of interest (i.e., for L2 reading of French documents).

We used the Dale–Chall score because of its prior frequent use in other similar studies, its relative simplicity, its proven reliability for assessing a wide range of document collections for readability (Chall & Dale, 1995), and its use of a high-frequency (or “easily-acquired”) word list to estimate word-level complexity. Although other measures may provide marginally better accuracy, our results do suggest that the Dale–Chall scores for the documents in our study serve at the very least as good proxies to true document complexity and provide at least a reasonable approximation to ESL readability.

5.6. Conclusions and future work

While our work serves to increase knowledge of the effect of language proficiency in searching for and assessing e-governmental documents, there are a number of suggestions for future work in this area of study. These would not only support the internal and external validity of the research but would further explore the role and impact of reading level on both ESL and native English-speaking users of e-government platforms.

Firstly, expanding the participant pool to incorporate a wider range of native and ESL speakers – which our Chrome plugin would permit in remote settings – would serve to increase the generalisability of the results and may result in additional significant findings, some of which may have been underpowered in the present study. Removal of time constraints and allowing users to follow a more realistic search experience without the mediating effects of lab conditions would also provide for more authentic results and account for the highlighted concerns around internal validity, although this may impact external validity. It may be insightful to allow users to provide their own more granular relevance judgements for a deeper understanding into the role document reading level and user comprehension play in the selection of documents for e-government topics.

Consideration of other readability measures designed specifically for language learners (and/or those who have English as a second language) would be interesting to determine whether this has any demonstrable effect on the key findings and outcomes. We note that, despite years of research on this topic, there is still considerable contradiction in the literature as to which measures are most appropriate for assessing readability for L2 readers, e.g., the results of Chen and Truscott (2010) and Koirala (2015) as opposed to those of Björnsson (1983). Further studies in this area are clearly warranted.

Finally, we intend to conduct more detailed and specific analyses of the wealth of data collected by our bespoke logging tool to identify differences between the groups in terms of how they interacted with the search system when completing the tasks. This will also include an expansion of the initial classification work conducted for this paper to understand in more detail the utility of English language proficiency and document reading level in predicting search success. Such research may lead to further insights into how systems could be adapted or developed to better assist non-native speakers in searching for and understanding documents in their L2 language(s).

CRediT authorship contribution statement

Morgan Harvey: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **David Brazier:** Conceptualization, Methodology, Software, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization.

References

Aham-Anyanwu, N., & Li, H. (2017). E-public engagement: Formulating a citizen content engagement model. In *25th European conference on information systems (ECIS)* (pp. 753–770).

Al-Muwil, A., Weerakkody, V., El-Haddadeh, R., & Dwivedi, Y. (2019). Balancing digital-by-default with inclusion: A study of the factors influencing e-inclusion in the UK. *Information Systems Frontiers*, *21*(3), 635–659.

Alam, K., & Imran, S. (2015). The digital divide and social inclusion among refugee migrants: A case in regional Australia. *Information Technology & People*, *28*(2), 344–365.

Arguello, J., Choi, B., & Capra, R. (2018). Factors influencing users' information requests: Medium, target, and extra-topical dimension. *ACM Transactions on Information Systems (TOIS)*, *36*(4), 41.

Aula, A., & Kellar, M. (2009). Multilingual search strategies. In *CHI'09 extended abstracts on human factors in computing systems* (pp. 3865–3870). ACM.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63–88.

Berendt, B., & Kralisch, A. (2009). A user-centric approach to identifying best deployment strategies for language tools: the impact of content and access language on web user behaviour and attitudes. *Information Retrieval*, *12*(3), 380–399.

Björnsson, C.-H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, 480–497.

Bogers, T., Gäde, M., Hall, M., & Skov, M. (2016). Analyzing the influence of language proficiency on interactive book search behavior. In *Proceedings of iconference 2016*. iSchools.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3).

Borlund, P. (2013). Interactive information retrieval: An introduction. *Journal of Information Science Theory and Practice*, *1*(3), 12–32.

Brazier, D., & Harvey, M. (2017a). E-government and the digital divide: A study of english-as-a-second-language users' information behaviour. In *European conference on information retrieval* (pp. 266–277). Springer.

Brazier, D., & Harvey, M. (2017b). Strangers in a strange land: A study of second language speakers searching for e-services. In *Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 281–284).

Brazier, D., & Harvey, M. (2018). A comparative study of native and non-native information seeking behaviours. In *European conference on information retrieval* (pp. 237–248). Springer.

Burroughs, J. M. (2009). What users want: Assessing government information preferences to drive information services. *Government Information Quarterly*, *26*(1), 203–218.

Cawthorne, C., & Barnes, E. (2016). Looking at the different ways to test content.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*(5), 693–713.

Choudrie, J., Ghinea, G., & Songonuga, V. N. (2013). Silver surfers, e-government and the digital divide: An exploratory study of UK local authority websites and older citizens. *Interactive Computing*, *25*(6), 417–442.

Chu, P., Jozsa, E., Komlodi, A., & Hercegfi, K. (2012). An exploratory study on search behavior in different languages. In *Proceedings of the 4th information interaction in context symposium* (pp. 318–321). ACM.

Chu, P., & Komlodi, A. (2017). TranSearch: A multilingual search user interface accommodating user interaction and preference. In *2017 CHI conference extended abstracts* (pp. 2466–2472). ACM.

Clarke, C. L., Craswell, N., & Soboroff, I. (2004). Overview of the TREC 2004 terabyte track. In *TREC, Vol. 4* (p. 74).

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, *42*(2), 288–325.

Clough, P., & Eleta, I. (2010). Investigating language skills and field of knowledge on multilingual information access in digital libraries. *International Journal of Digital Library Systems (IJDLS)*, *1*(1), 89–103.

Cohron, M. (2015). The continuing digital divide in the United States. *The Serials Librarian*, *69*(1), 77–86.

Coiro, J. (2011). Predicting reading comprehension on the internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research*, *43*(4), 352–392.

Collins-Thompson, K., Bennett, P. N., White, R. W., De La Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 403–412).

Collins-Thompson, K., & Callan, J. P. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference of the north American chapter of the association for computational linguistics: HLT-NAACL 2004* (pp. 193–200).

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*(13), 1448–1462.

DuBay, W. H. (2004). The principles of readability. Online Submission.

Dwivedi, Y. K., & Williams, M. D. (2008). Demographic influence on UK citizens'e-government adoption. *Electronic Government, An International Journal*, *5*(3), 261–274.

Edwards, A., & Kelly, D. (2016). How does interest in a work task impact search behavior and engagement? In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval* (pp. 249–252). ACM.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221.

Freund, L. (2013). A cross-domain analysis of task and genre effects on perceptions of usefulness. *Information Processing & Management*, *49*(5), 1108–1121.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Ernst Klett Sprachen.

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal*, *26*(1), 5–24.

Gunning, R., et al. (1968). *Technique of clear writing*. McGraw Hill Higher Education.

Guo, Q., Lagun, D., & Agichtein, E. (2012). Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 2050–2054).

Hahnel, C., Goldhammer, F., Kröhne, U., & Naumann, J. (2018). The role of reading skills in the evaluation of online information gathered from search engine environments. *Computers in Human Behavior*, *78*, 223–234.

Haley, A. N., & Clough, P. (2017). Affective experiences of international and home students during the information search process. *New Review of Academic Librarianship*, *23*(4), 396–420.

Harvey, M., Hastings, D. P., & Chowdhury, G. (2021). Understanding the costs and challenges of the digital divide through UK council services. *Journal of Information Science*, Article 01655515211040664.

Helbig, N., Gil-García, J. R., & Ferro, E. (2009). Understanding the complexity of electronic government: Implications from the digital divide literature. *Government Information Quarterly*, *26*(1), 89–97, From Implementation to Adoption: Challenges to Successful E-government Diffusion.

Józsa, E., Köles, M., Komlódi, A., Hercegfi, K., & Chu, P. (2012). Evaluation of search quality differences and the impact of personality styles in native and foreign language searching tasks. In *Proceedings of the 4th information interaction in context symposium* (pp. 310–313). ACM.

Kang, H. (2014). Understanding online reading through the eyes of first and second language readers: An exploratory study. *Computers & Education*, *73*, 1–8.

Kattenbeck, M., & Elsweiler, D. (2019). Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management*, *71*(3), 368–391.

Kelly, D., et al. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, *3*(1–2), 1–224.

Khan, S., Asif, A., & Jaffery, A. E. (2020). Language in a time of COVID-19: Literacy bias ethnic minorities face during COVID-19 from online information in the UK. *Journal of Racial and Ethnic Health Disparities*, 1–7.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, *57*, 1–44.

Koirala, C. (2015). The word frequency effect on second language vocabulary learning. In *Critical CALL–proceedings of the 2015 EUROCALL conference, Padova, Italy* (pp. 318–323). Research-publishing. net.

Kolsaker, A., & Lee-Kelley, L. (2008). Citizens' attitudes towards e-government and e-governance: a UK study. *International Journal of Public Sector Management*, *21*(7), 723–738.

Komba, M. M., & Lwoga, E. T. (2015). Government information seeking behaviour of citizens in selected districts of tanzania. *International Research: Journal of Library and Information Science*, *5*(4).

Kralisch, A., & Berendt, B. (2005). Language-sensitive search behaviour and the role of domain knowledge. *New Review of Hypermedia and Multimedia*, *11*(2), 221–246.

Lambert, F. (2013). Seeking electronic information from government resources: A comparative analysis of two communities' web searching of municipal government websites. *Government Information Quarterly*, *30*(1), 99–109.

Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2017). New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. *Journal of Education*, *197*(2), 1–18.

Liu, C., Liu, Y.-H., Gedeon, T., Zhao, Y., Wei, Y., & Yang, F. (2019). The effects of perceived chronic pressure and time constraint on information search behaviors and experience. *Information Processing & Management*, *56*(5), 1667–1679.

Lloyd, A. (2020). Shaping the contours of fractured landscapes: Extending the layering of an information perspective on refugee resettlement. *Information Processing & Management*, *57*(3), Article 102062.

Macevičiūtė, E., & Manžuch, Z. (2018). Conceptualising the role of digital reading in social and digital inclusion. In *Proceedings of ISIC, the information behaviour conference, Krakow, Poland, 9-11 October, 2018: Part 1*. Information Research.

Oduntan, O., & Ruthven, I. (2019). The information needs matrix: A navigational guide for refugee integration. *Information Processing & Management*, *56*(3), 791–808.

Ojha, P. K., Ismail, A., & Kuppusamy, K. (2018). Perusal of readability with focus on web content understandability. *Journal of King Saud University-Computer and Information Sciences*, *33*, 1–10.

Pancer, E., Chandler, V., Poole, M., & Noseworthy, T. J. (2019). How readability shapes social media engagement. *Journal of Consumer Psychology*, *29*(2), 262–270.

Park, J., Yang, J., & Hsieh, Y. C. (2014). University level second language readers' online reading and comprehension strategies. *Language Learning & Technology*, *18*(3), 148–172.

Parsazadeh, N., Ali, R., & Rezaei, M. (2018). A framework for cooperative and interactive mobile learning to improve online information evaluation skills. *Computers & Education*, *120*, 75–89.

Peters, C., Braschler, M., & Clough, P. (2012). *Multilingual information retrieval: From research to practice*. Springer Science & Business Media.

Pirolli, P. (2009). *Information foraging theory: Adaptive interaction with information*. Oxford University Press.

Rózsa, G., Komlodi, A., & Chu, P. (2015). Online searching in english as a foreign language. In *Proceedings of the 24th international conference on world wide web companion* (pp. 875–880). International World Wide Web Conferences Steering Committee.

Ruokolainen, H., & Widén, G. (2020). Conceptualising misinformation in the context of asylum seekers. *Information Processing & Management*, *57*(3), Article 102127.

Savolainen, R. (2016). Approaches to socio-cultural barriers to information seeking. *Library & Information Science Research*, *38*(1), 52–59.

Savolainen, R., & Kari, J. (2006). User-defined relevance criteria in web searching. *Journal of Documentation*, *62*(6), 685–707.

Schwarz, J., & Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1245–1254).

Selwyn, N., & Facer, K. (2007). Beyond the digital divide. In *Opening education reports* (p. 24). Bristol: Futurelab.

Soroya, S. H., Farooq, A., Mahmood, K., Isoaho, J., & Zara, S.-e. (2021). From information seeking to information avoidance: Understanding the health information behavior during a global health crisis. *Information Processing & Management*, *58*(2), Article 102440.

Steichen, B., & Freund, L. (2015). Supporting the modern polyglot: A comparison of multilingual search interfaces. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3483–3492). ACM.

Steichen, B., & Lowe, R. (2020). How do multilingual users search? An investigation of query and result list language choices. *Journal of the Association for Information Science and Technology*, *n/a*(n/a), 1–18.

Swire-Thompson, B., & Lazer, D. (2020). Public health and online misinformation: challenges and recommendations. *Annual Review of Public Health*, *41*, 433–451.

Tamine, L., & Chouquet, C. (2017). On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing & Management*, *53*(2), 332–350.

Uitdenbogerd, S. (2005). Readability of french as a foreign language and its uses. In *Proceedings of the Australian document computing symposium* (pp. 19–25). Citeseer.

Van Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. *Poetics*, *34*(4–5), 221–235.

Vinson, T. (2009). *The origins, meaning, definition and economic implications of the concept social inclusion/exclusion: incorporating the core indicators developed by the European union and other illustrative indicators that could identify and monitor social exclusion in Australia*. Canberra: Department of Education, Employment and Workplace Relations.

Vuong, T., Jacucci, G., & Ruotsalo, T. (2017). Watching inside the screen: Digital activity monitoring for task recognition and proactive information retrieval. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 109.

W3Techs (2022). Usage of content languages for websites. https://w3techs.com/technologies/overview/content_language Last accessed: 15th February 2022.

Wang, J., & Komlodi, A. (2018). Switching languages in online searching: A qualitative study of web users' code-switching search behaviors. In *Proceedings of the 2018 conference on human information interaction & retrieval* (pp. 201–210).

Weber, H., Becker, D., & Hillmert, S. (2018). Information-seeking behaviour and academic success in higher education: Which search strategies matter for grade differences among university students and how does this relevance differ by field of study? *Higher Education*, 1–22.

Weth, C. v. d., & Hauswirth, M. (2013). Dobbs: Towards a comprehensive dataset to study the browsing behavior of online users. In *Proceedings of the 2013 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT)-Vol. 01* (pp. 51–56). IEEE Computer Society.

White, R. W., Dumais, S. T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining* (pp. 132–141). ACM.

Yates, S. J., Kirby, J., & Lockley, E. (2015). 'Digital-by-default': reinforcing exclusion through technology. *Defence of Welfare*, *2*, 158–161.