

Summer 2015

Effects of Summary Length and Line Spacing on Fixations, Decision Time, Correctness, and Preference of Search Engine Results on a Phablet

Anne Marie McReynolds
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

McReynolds, Anne Marie, "Effects of Summary Length and Line Spacing on Fixations, Decision Time, Correctness, and Preference of Search Engine Results on a Phablet" (2015). *Master's Theses*. 4601.

DOI: <https://doi.org/10.31979/etd.rq8m-35zw>

https://scholarworks.sjsu.edu/etd_theses/4601

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

EFFECTS OF SUMMARY LENGTH AND LINE SPACING ON FIXATIONS,
DECISION TIME, CORRECTNESS, AND PREFERENCE OF
SEARCH ENGINE RESULTS ON A PHABLET

A Thesis

Presented to

The Faculty of the Department of Industrial and Systems Engineering

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Anne-Marie McReynolds

August 2015

© 2015

Anne-Marie McReynolds

ALL RIGHTS RESERVED

The Designated Thesis Committee Approves the Thesis Titled

EFFECTS OF SUMMARY LENGTH AND LINE SPACING ON FIXATIONS,
DECISION TIME, CORRECTNESS, AND PREFERENCE OF
SEARCH ENGINE RESULTS ON A PHABLET

by

Anne-Marie McReynolds

APPROVED FOR THE DEPARTMENT OF INDUSTRIAL AND
SYSTEMS ENGINEERING

SAN JOSÉ STATE UNIVERSITY

August 2015

Dr. Cary Feria Department of Psychology

Dr. Louis Freund Department of Industrial and Systems Engineering

Dr. Emily Wughalter Department of Kinesiology

ABSTRACT

EFFECTS OF SUMMARY LENGTH AND LINE SPACING ON FIXATIONS, DECISION TIME, CORRECTNESS, AND PREFERENCE OF SEARCH ENGINE RESULTS ON A PHABLET

by Anne-Marie McReynolds

In previous studies, a positive relationship has been suggested between the screen size of a mobile device and the preferred summary length of a search result: the bigger the screen, the longer the summary preferred for judging the relevance of a result. While prior research has been focused on three types of devices (cell phones, PDAs, laptops), this study was concentrated on a new class of smartphone called a phablet that could eventually replace all three. In the current research, we investigated how two factors in the design of search result pages—summary length and line spacing—affect performance, behavioral and subjective measures on an information-seeking task executed on a phablet. We examined the effects of summary length (1, 3, 7, 10 lines) and line spacing (single, one and a half, double) on fixations, decision time, correctness, and preference. A direct relationship between summary length, fixations and decision time was found: as summary length increased, fixations and decision time also increased. No relationship between summary length and decision correctness was found. The optimal summary length for effectively judging the relevance of a search result—the one requiring the fewest fixations and shortest decision time—is one line. Because participants did not prefer one-line summaries, it is best to show three lines. As such, three-line summaries suggest a minimal tradeoff between performance and preference.

Table of Contents

List of Tables	viii
List of Figures	ix
1. Introduction.....	1
2. Literature Review.....	2
2.1. Summary Length	2
2.1.1. Measures of accuracy.....	2
2.1.2. Decision time and correctness	4
2.1.3. Preference	7
2.2. Line Spacing	8
2.2.1. Decision time and correctness	8
2.2.2. Fixations.....	9
2.3. Summary.....	10
2.3.1. Summary length	10
2.3.2. Line spacing.....	11
2.4. The Present Study	11
2.4.1. Hypotheses.....	12
2.5. Definition of Terms	14
3. Method	14
3.1. Design	14
3.1.1. Participants.....	16
3.1.2. Apparatus	16
3.1.2.1. Mobile device.....	16

3.1.2.2.	Mobile application	16
3.1.2.3.	Eye-tracking glasses.....	18
3.1.2.4.	Software	19
3.2.	Stimuli.....	19
3.3.	Procedure	23
3.4.	Dependent Measures.....	29
4.	Results.....	29
4.1.	Fixations	29
4.2.	Decision Time.....	31
4.3.	Decision Correctness	34
4.4.	Preference	35
5.	Discussion.....	37
5.1.	Summary Length	37
5.1.1.	Fixations.....	37
5.1.2.	Decision time	38
5.1.3.	Decision correctness	40
5.1.4.	Preference	41
5.2.	Line Spacing	41
5.2.1.	Fixations.....	41
5.2.2.	Decision time	42
5.3.	Limitations.....	42
5.3.1.	Poor track ability.....	43
5.3.2.	Pagination	44

5.3.3. No visible difference between one-line, single-spacing conditions.....	46
5.3.4. Task difficulty.....	48
5.3.5. Forced-choice rating.....	51
5.3.6. Preference for line spacing not calculated.....	52
6. Conclusion.....	52
References.....	53

List of Tables

Table		Page
1.	Study Variables	15
2.	Experimental Conditions.....	15
3.	Descriptive Statistics for Fixations	30
4.	Descriptive Statistics for Decision Time (s)	32
5.	Descriptive Statistics for Decision Correctness	34
6.	Descriptive Statistics for Preference	36

List of Figures

Figure		Page
1.	Native application on Samsung Galaxy Note III	17
2.	Infrared cameras in eye-tracking glasses	18
3.	Correct eye position	19
4.	User interface for the three-line, single-spacing condition	20
5.	Post-task questionnaire for preference selection.....	22
6.	Calibration process for first point	24
7.	Successful calibration point	25
8.	Trial control page	26
9.	Flowchart of experimental procedure for a study session.....	28
10.	Interaction plot of raw data for fixations.....	31
11.	Interaction plot of raw data for decision time	33
12.	Line graph of percent correct for each level of summary length	35
13.	Frequency distribution of raw data for summary length preference	37
14.	Search results for the three-line, one-and-a-half-spacing condition	40
15.	Screen captures from video recorded by an infrared camera in the lens of the eye-tracking glasses.....	43
16.	Ten-line, one-and-a-half-spacing condition	45
17.	One-line, (A) single-, (B) one-and-a-half-, and (C) double-spacing conditions	47
18.	Search results for the one-line, one-and-a-half-spacing condition for the get-wine-recommendations stimulus	48
19.	Four search results for the one-line, single-spacing condition.....	50

1. Introduction

Smartphone screens are getting bigger, turning cell phones into small tablets. Before 2011, when only 35% of all Americans owned a smartphone, screens were no larger than four inches (Taylor, 2014). Since 2013, screen sizes have grown to an average size of five inches (Barredo, 2014). The Samsung Galaxy Note III with a 5.7-inch screen was the third most popular phone sold in February 2014 (Kang, 2014). As of October 2014, 64% of Americans owned a smartphone, and Apple's iPhone 6 with a 4.7-inch screen was the top-selling device (Smith, 2015; Milanesi, 2014). The increasing screen size reflects smartphone user preference and intent, with 34% using their phones to search online rather than on a desktop or laptop computer (Smith, 2015). In 2015, mobile search volume is expected to surpass desktop local search volume for the first time (Pacheco, 2012).

Along these lines, search engines should optimize content intended for mobile devices with smaller screens and constrained interactions (no physical keyboard or mouse) to help searchers quickly judge and use a relevant search result. The attentional cost of clicking on an irrelevant search result is, at best, a noticeable break in the user's task flow and, at worst, task abandonment (Nielsen, 1993). Most search engines present results as a serial list of documents, with each result showing the document title, a summary, and the URL. Summaries help a searcher decide whether to navigate to information sources without having to read the full text of the source documents. In the current research, we investigate how two key factors in the design of search engine

results pages (SERPs)—summary length and line spacing—affect fixations, decision time, correctness, and preference on an information-seeking task.

2. Literature Review

2.1. Summary Length

In previous studies, researchers have attempted to identify an optimal summary length for search results displayed on small screen devices (Sweeney, Crestani, & Tombros, 2002; Sweeney & Crestani, 2006). Sweeney et al. (2002) presented automatically generated summaries of newspaper articles on a mobile phone interface and studied the effects of summary length on relevance judgments. Ten participants were shown a description of a search query and asked to identify as many relevant results as possible in five minutes. Participants could select any result from a ranked list of 50 search results and read its title and three levels of summary (7, 15, and 30% of the original document's length). Participants then reported their relevance judgments and the level of summary used to make their decisions.

2.1.1. Measures of accuracy. Sweeney et al. (2002) hypothesized that longer summaries would improve accuracy, that is, precision and recall. Precision was defined as the fraction of relevant results—the number of correct results divided by the number of retrieved results. Recall was the number of correct results divided by the total number of relevant documents for a topic. The authors of this study figured that longer summaries enabled better decision making by providing participants with more relevant clues. However, summary length did not have a significant effect on measures of accuracy. When the authors analyzed the number of documents at different levels of summary that

participants used to make their judgments, they found that most participants based their relevance decisions on the shortest summaries (7% of the document's length). Because relevance decisions made with long summaries were no more correct than those made with short summaries, the authors concluded that there was no need to produce summaries longer than 7% of the document's length.

However, the relatively small sample size of 10 participants limits the generalizability of Sweeney et al.'s (2002) findings. Another limitation relates to the ecological validity of the study—the extent to which experimental conditions mimic real-life conditions. The mobile browser emulator looked like a smartphone screen on a desktop PC, and the participant used the PC's keyboard to enter text. Using a mouse to click on a simulator differs from the actual experience of using a finger to tap on a device screen. To this end, the authors repeated the experiment with three mobile devices (cell phone, PDA, and laptop).

Like the previous experiment, 30 participants in a between-subjects design (10 per device) were presented with a description of a search query and asked to identify as many relevant results as possible in five minutes. Participants could judge a search result as relevant/not relevant after reading the title or move to the next summary length (7, 15, and 30% of the document). The authors hypothesized a positive relationship between screen size and summary length: the bigger the screen size, the longer the summary used to make relevance judgments. Consistent with their hypothesis, they found that participants used longer summaries on a laptop with a 14.10-inch screen and shorter summaries on a cell phone with a 2.18-inch screen.

Sweeney and Crestani (2006) predicted that longer summaries would be more effective on a larger screen, and shorter summaries would be more effective on a smaller screen. However, the results of their experiments indicated that this was not the case. Participants made precise and correct relevance decisions using small summaries, regardless of the screen size used. The performance difference between short (title and 7%) and long (15 and 30%) summaries suggested a threshold after which summary length degrades effectiveness. In keeping with their previous findings, the authors concluded that it was best to show the same level of summary (title and 7%) on all devices.

2.1.2. Decision time and correctness. Sweeney and Crestani (2006) could not report on the time taken to make relevance decisions at different summary levels, as task duration was a controlled variable (five minutes per query). Cutrell and Guan (2007) found that decision time and correctness improved as summary length increased for “informational” (information-seeking) tasks. In this study, experimenters manipulated task type (informational, navigational) and summary length (one line, two to three lines, six to seven lines) for opposite effects on performance and behavioral measures (click accuracy and total time on task). Informational tasks required the participant to locate specific information (e.g., “Find the address for the Newark Airport”), and navigational tasks required the participant to find a specific Web page (e.g., “Find the homepage of the Pinewood software company”) (p. 4). The experimenters presented 18 participants with a task description and a search results page containing 10 results with one of three summary lengths: one line, two to three lines, and six to seven lines. Participants were

instructed to identify the “best” result. For informational tasks, the “best” result contained the searched terms (i.e., Newark airport address) in the summary. For navigational tasks, the “best” result was often a URL of a website located in a “very obvious place” on the search results page (p. 5).

For informational tasks, task time improved (a reduction of 24 s from 84) with six- to seven-line summaries. For navigational tasks, task time remained the same for summaries less than three lines and increased by 10 s (from 32 s to 42) with the six- to seven-line summaries. If one only focuses on informational tasks, the results indicate that providing more contexts helped users judge whether a given result would likely meet their information need before they navigated to it. To investigate this further, Cutrell and Guan (2007) examined click accuracy, whether a participant clicked on the “best” result—the one that contained the searched terms in the result summary. While the results were not significant, the authors inferred from descriptive statistics that click accuracy improved as summary length increased. In contrast, increasing the summary length had the opposite effect on click accuracy for navigational tasks; click accuracy decreased with six- to seven-line summaries. In sum, participants performed best—made faster and more correct decisions—on navigational tasks with summaries less than three lines and on informational tasks with six to seven lines.

The finding that long summaries improved decision time and correctness on informational tasks is inconsistent with the results of Sweeney and Crestani (2006). The long (six- to seven-line) condition in Cutrell and Guan (2007), for which participants made the fastest and most correct decisions, is more comparable to the 15% (six-line)

summary level in Sweeney and Crestani (2006), for which participants were the least precise. This discrepancy in findings indicates that judging the relevance of news articles may not be similar in nature to an informational search task, where the participant has to find specific information. Still, Cutrell and Guan's (2007) finding that decision time on informational tasks improved with longer summaries seems counterintuitive. Increasing the summary length would reduce the number of search results that fit on a screen (from seven to four), causing the user to scroll more and read less. However, eye-tracking data revealed that the time spent looking at the summary increased with its length, indicating that participants read long summaries regardless of their relevance to the task at hand.

Gonzalez-Caro and Marcos (2011) also found that participants fixated more and longer on the summary for informational tasks (they had to find information about an item) and the URL for navigational tasks (they had to find a specific Web page). The larger size of the summary in relation to the title or URL may have garnered more visual attention in prior studies because of its salience, not because participants were reading the text. Furthermore, prior experience conducting exploratory search tasks online may have influenced gaze behavior. Participants fixated where they expected to find the most information in the search result, which may explain why fixation duration on the summary increased while the overall decision time decreased in Cutrell and Guan (2007).

2.1.3. Preference. In previous studies, clickthroughs have been treated as implicit indicators of user preference. Clarke, Agichtein, Dumais, and White (2007) used clickthrough logs from the Microsoft Windows Live search engine to investigate the

effects of document summary length on user preference. They analyzed clickthroughs during a three-day period and extracted queries that had at least 10 clickthroughs during the sampling period, yielding over 5,000 data points. The data supported that longer summaries (more than two lines) increased the probability of a searcher clicking on a result. The finding that more searchers clicked on search results with longer summaries implied that users preferred summaries with more than two lines of text.

In contrast, the results of a 2,500-person survey by Yahoo indicated that participants had no preference for summary length (Rose, Orr, & Kantamneni, 2007). In this study, Rose et al. (2007) presented each participant with a hypothetical search scenario and a query that might have been used in that situation. The experimenters then showed a search result for that query with a short (two-line), medium (three-line) or long (four-line) summary, after which participants answered questions about its relevance (e.g., “Would you click on this search result?”). They found no effect of summary length on perceptions of relevance when users read search results.

When Kaisser, Hearst, and Lowe (2008) presented participants with more varied summary lengths for queries from log files, they preferred longer summaries—one or more paragraphs or a full document—for general information searches. In the first experiment, 114 participants classified expected search results into one of 17 categories (e.g., person, place, website, general information). Then they estimated the preferred summary length (word or short phrase, sentence, paragraph, an article or full document) for that search result. When the query warranted a search result type of general information, most participants preferred an article or full document for the result. For the

second experiment, participants saw a query/summary-length pair and judged its relevance, rating it on a scale from 0 (very bad response) to 10 (very good response). The results were consistent with estimations of preferred summary length in the first experiment. For general information needs, participants judged longer summaries as more relevant. This finding is consistent with that of Sweeney and Crestani (2006), who reported that participants chose longer, more informative summaries when using a larger screen to make relevance decisions.

2.2. Line Spacing

2.2.1. Decision time and correctness. While the ideal number of lines for a search result summary has been thoroughly investigated, the effects of line spacing on the speed and accuracy of visual search have only been documented in three landmark papers. Ling and Schaik (2007) found that varying line spacing (single, one and a half, double) had a significant effect on task performance: participants performed better—were faster and more accurate at locating a hyperlink in a text-laden webpage—with double-spaced text than with one-and-a-half spacing, and better with one and a half than single spacing. Subjective measures indicated that participants preferred single spacing.

However, Ojanpää, Näsänen, and Kojo (2002) found the opposite result: participants were faster at locating a target word in a dense layout, but they preferred searching the sparse layout. The discrepancy in findings may be due to the greater interlinear distance between words in this study than in other similar studies. List items in Ojanpää et al. (2002) had zero, one, two, three or four empty lines between them. Document text in Ling and Schaik (2007) had one, one-and-a-half or two-line breaks.

Also, the nature of the visual search tasks was fundamentally different in the two studies. In Ling and Schaik (2007), participants performed a feature search, and, in Ojanpää et al. (2002), they performed a conjunction search. Searching for a single feature, like color, resulted in participants performing a parallel search process in Ling and Schaik (2007). The target—a blue hyperlink—“popped out” against a screen of black-and-white text. In contrast, in Ojanpää et al. (2002), participants searched for a target word in a list by examining each word one by one until the target was found, or all items were evaluated. This serial search process took longer. With increasing line spacing, search time significantly increased in Ojanpää et al. (2002). The additional time was attributed to an increase in the number of fixations; more fixations were needed to search the vertical list of eight words spread over a wider area. To this end, search time and the number of fixations were highly correlated.

2.2.2. Fixations. A sample size of three limits the generalizability of Ojanpää et al. (2002) results. Though eye-tracking data in Tarling and Brumby (2010) confirm the finding that search efficiency improves as interlinear distance shrinks. In this study, eye movements were recorded while participants searched for a target word in a list of 36 words for three layout conditions: all dense, all sparse, and mixed. In the mixed condition, 18 items were displayed in two vertical columns—one dense and the other sparse. Eye-tracking data revealed that participants were more efficient when words were densely packed together (all-dense condition); they made fewer eye movements, as more characters could be assessed in a single fixation.

2.3. Summary

2.3.1. Summary length. There are two schools of thought on the ideal summary length for search results—one that will maximize decision correctness and minimize decision time. Sweeney and Crestani's (2006) finding that participants made precise and correct relevance decisions using summaries less than three lines forms the basis of the first school of thought. Since relevance decisions made with longer summaries (greater than or equal to seven lines) were no more correct than those made with short summaries (less than or equal to three lines), the authors concluded that there was no reason to produce summaries longer than three lines.

On the other hand, Cutrell and Guan (2007) found that decision time and correctness improved with six- to seven-line summaries. Their results indicate that providing more contexts helps searchers judge the relevance of a result before they navigate to it. The finding that decision time and correctness improved with longer summaries seems illogical though. Eye-tracking data revealed that time spent looking at the summary increased with its length (Gonzalez-Caro & Marcos, 2011). Longer summaries also meant fewer search results on a page, requiring the searcher to scroll more and read less. The performance difference between three-line summaries and six- to twelve-line summaries in Sweeney and Crestani (2006) hints at a threshold after which summary length degrades effectiveness. Prior research on the effects of summary length on user preference indicates that users preferred two lines to a paragraph (Clarke et al., 2007; Kaiser et al., 2011). The finding that participants judged longer summaries as more relevant is consistent with that of Sweeney and Crestani (2006), who reported that

participants chose longer, more informative summaries when using a desktop PC to make relevance judgments.

2.3.2. Line spacing. The three notable papers on the effects of line spacing on speed and accuracy use different visual search tasks, yielding conflicting results. Ling and Schaik (2007) found that search efficiency improved with interlinear distance. Participants were faster and more accurate at locating a hyperlink in a text-laden webpage with double-spaced text than with one-and-a-half spacing, and better with one-and-a-half spacing than single spacing. However, prior research on word search had the opposite result: participants were faster at locating a target word in a single-spaced list (Ojanpää et al., 2002; Tarling & Brumby, 2010). Ojanpää et al. (2002) attributed slower search times with wider line spacing to an increase in the number fixations; more fixations were required to search a wider area. Eye-tracking data in Tarling and Brumby (2010) revealed that participants made fewer eye movements when words were densely packed together, as more characters could be evaluated in a single fixation.

2.4. The Present Study

The goal of the present study was to investigate how two key factors in the design of search engine results pages—summary length and line spacing—affect fixations, decision time, correctness, and preference on an informational task executed on a phablet. While prior research has been focused on three types of devices (cell phones, PDAs, and laptops), this study was concentrated on a new class of smartphone called a phablet that could eventually replace all three. In fact, phablets outsold laptops and desktop PCs in 2014 and are predicted to surpass tablet sales in 2015 (Chau, 2015).

2.4.1. Hypotheses. In this study, we examined the effects of summary length (one, three, seven, ten lines) and line spacing (single, one and a half, double) on performance and preference. The dependent measures were number of fixations, decision time, correctness, and preference. We hypothesized that:

- Fixation frequency in areas of interest (AOIs) and decision time would decrease as summary length increased up to the seven-line condition, after which performance would deteriorate. According to this logic, participants would adopt a different search strategy for three- and seven-line summaries—one that involved shorter scanning fixations, meaning that irrelevant search results were rejected faster. So, more information—up to seven lines—would improve scanning effectiveness. Cutrell and Guan (2007) found that decision times on informational tasks improved with seven-line summaries. On the other hand, ten lines would be viewed as too much text to scan efficiently, as participants would likely read each word. More fixations and longer decision times for the ten-line condition would indicate an absolute length above which participants' search performance would degrade on an informational task.
- Decision correctness would improve with summary length. Cutrell and Guan (2007) inferred from descriptive statistics that the inclusion of more contextual information improved click accuracy on an informational search task. While their findings were not significant due in part to a small number of observations, this study employed the same experimental task with triple the sample size.

- Preferred summary length would be greater than or equal to three lines of text on a phablet. Sweeney and Crestani (2006) found a direct relationship between summary length and clickthrough rate, with participants having chosen longer, more informative summaries when using a bigger screen. Clarke et al.'s (2007) finding that more users clicked on search results with summaries longer than two lines of text implied that users preferred summaries greater than a line. Rose et al.'s (2007) finding of no significant difference between two-, three-, and four-line summaries indicated that the preferred summary length was likely greater than or equal to three lines in the present study.
- An interaction between summary length and line spacing was predicted. As summary length increased, participants would perform better—be faster and more accurate—with double-spaced summaries. Ling and Schaik (2007) found that wider line spacing improved speed and accuracy when searching for a hyperlink in a screen of text. But for shorter summaries—with fewer words—single spacing would be better. Eye-tracking data in Tarling and Brumby (2010) revealed that word search efficiency improved as interlinear distance decreased; more characters were assessed in a single fixation.

2.5. Definition of Terms

Decision correctness: Percentage of correct decisions.

Fixation: SensoMotoric Instruments (SMI) BeGaze software detected a fixation after a participant steadily gazed into an Area of Interest (AOI) with a radius of 50 pixels for at least 100 ms.

Informational: A type of Web search task in which the goal is to locate specific information (e.g., “Find the address of the Newark Airport”) (Cutrell & Guan, 2007, p. 4).

Navigational: A type of Web search task in which the goal is to find a specific website or homepage (e.g., “Find the homepage of the Pinewood software company”) (Cutrell & Guan, 2007, p. 4).

Phablet: A portmanteau of the words phone and tablet. A phablet is a mobile device larger than a smartphone or PDA with functionality similar to a tablet computer.

Summary: First few sentences of a retrieved document.

3. Method

3.1. Design

This study was a 4 x 3 within-subjects factorial design. As can be seen in Table 1, the factors were summary length (one, three, seven, ten lines) and line spacing (single, one and a half, double).

Table 1.

Study Variables

Factor	Levels of Measurement
IV: Summary Length	1. 1 LN 2. 3 LN 3. 7 LN 4. 10 LN
IV: Line Spacing	1. Single 2. One and a half 3. Double
DV: Fixations	Sum of fixations in AOIs.
DV: Decision Time	Total time taken to select an application.
DV: Decision Correctness	Percentage of correct decisions.
DV: Preference	Summary length preferred the greatest number of times out of six.

Note. LN = line.

Each participant experienced all 12 conditions shown in Table 2 and completed 12 trials—one trial per condition. The order of conditions was randomly assigned across participants.

Table 2.

Experimental Conditions

Summary Length	Line Spacing		
	Single (S)	One and a half (H)	Double (D)
1 LN	1 LN x S	1 LN x H	1 LN x D
3 LN	3 LN x S	3 LN x H	3 LN x D
7 LN	7 LN x S	7 LN x H	7 LN x D
10 LN	10 LN x S	10 LN x H	10 LN x D

Note. LN = line.

3.1.1. Participants. Fifty-six adult college students (37 female, 19 male) from San José State University participated in this study. Each student participated in a 15-minute long session and was compensated with course credit. All observers self-reported normal color vision and normal or corrected-to-normal visual acuity. None of the participants were familiar with the purpose of the experiment.

Twenty participants' eye-tracking data (14 female, 6 male) were excluded from analysis due to poor tracking. More specifically, the eye-tracking glasses indicated poor track ability due to distortion caused by eye make-up (mascara) or drooping eyelids, which may have obscured the pupil.

3.1.2. Apparatus.

3.1.2.1. Mobile device. Participants performed search tasks on a Samsung Galaxy Note III phablet with a 5.7-inch, high-resolution (1080 x 1920 pixels) display. The Samsung Note, which came out in 2011, is credited as the first of its kind (Manjoo, 2014).

3.1.2.2. Mobile application. We developed a native Android application to control the summary length and line spacing of cached search results from Google (see Figure 1). The mobile application was developed in Java (programming language) using the Android Software Development Kit (SDK). Participants were only permitted to view experimental content offline, thus reducing effects of any outside factors that could influence the results. The application tracked the participant's selections, preferences, decision time, and correctness.

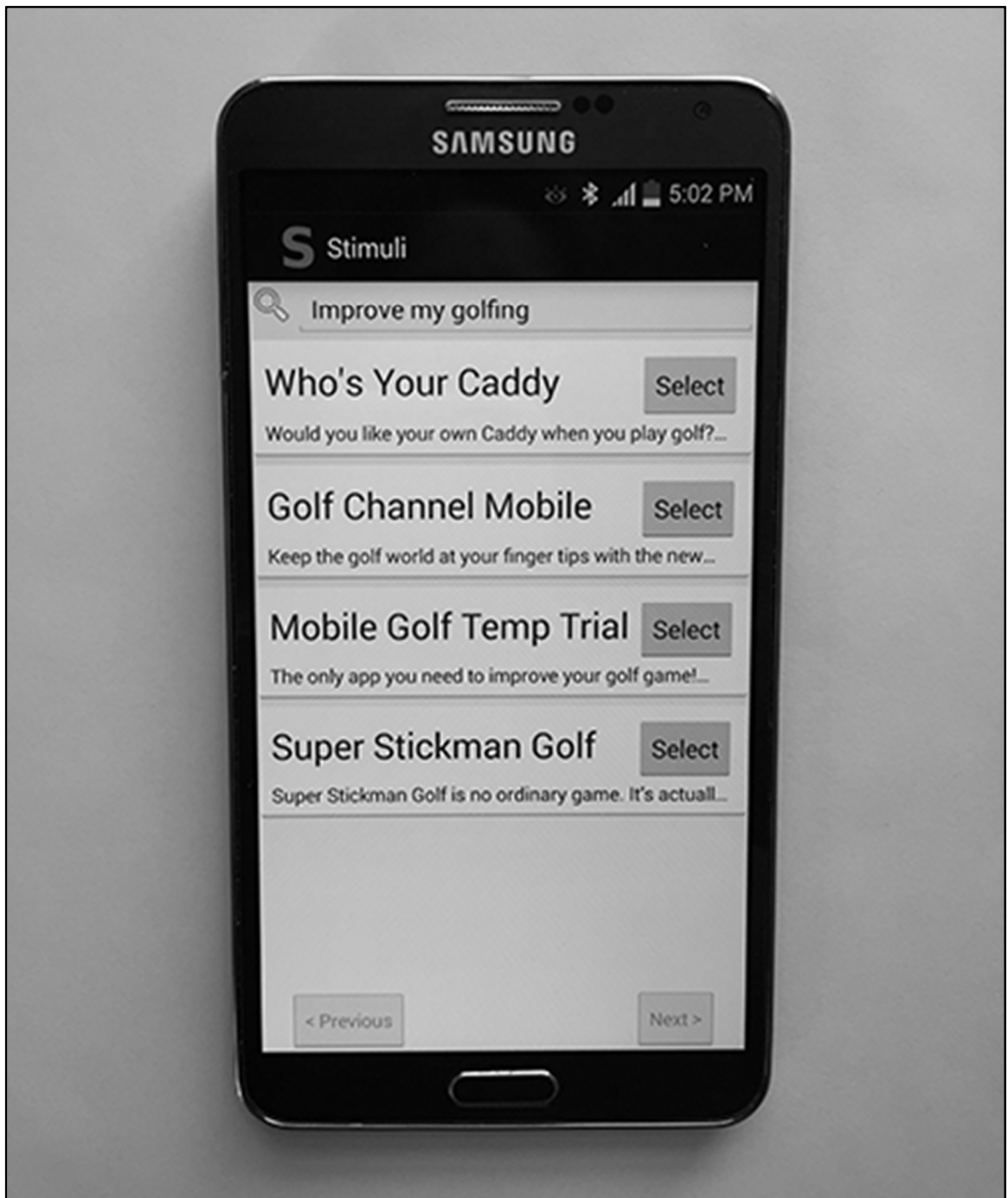


Figure 1. Native application on Samsung Galaxy Note III. This photograph shows the user interface for the one-line, single-spacing condition.

3.1.2.3. Eye-tracking glasses. Two cameras in the rim of the SensoMotoric Instruments' Eye-Tracking Glasses (SMI-ETG) shown in Figure 2 recorded the participant's eye movements. The hardware contained six infrared lights, designed to bounce invisible rays straight into the participant's eyes. An infrared camera in each lens then picked up those signals and used them to coordinate the location of the participant's pupils, tracking where the participant looked. Figure 3 shows a screen capture of a participant's right eye from video recorded by an infrared camera in the rim of the glasses. No other videotapes or audiotapes of participants were created.

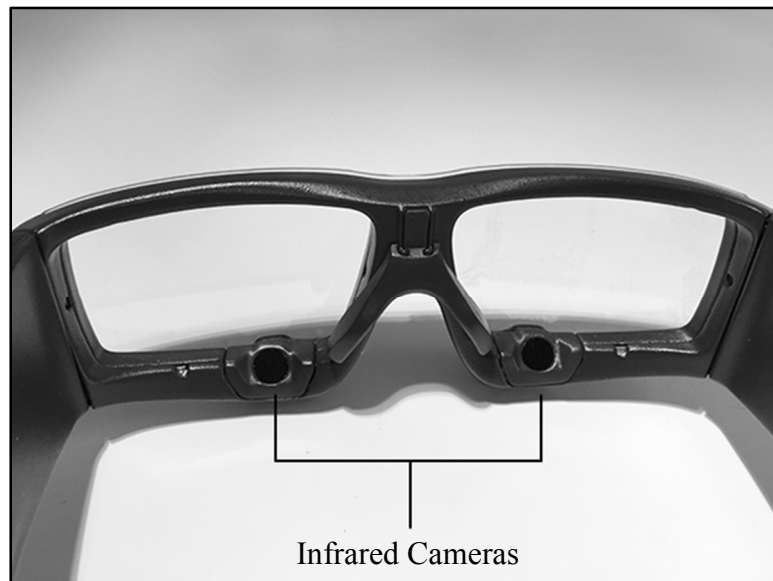


Figure 2. Infrared cameras in eye-tracking glasses. This photograph shows the location of the two infrared cameras in each lens of the glasses.



Figure 3. Correct eye position. This screen capture from video recorded by an infrared camera shows a participant's pupil positioned concentrically within the six infrared light rays.

3.1.2.4. Software. These eye-tracking glasses were attached to a laptop (with Microsoft Windows 7 operating system) via Universal Serial Bus (USB) for use of controller software (BeGaze iViewETG) that recorded eye-tracking data. For each mobile webpage, BeGaze reported fixation coordinates and durations relative to the page coordinates. This software allowed the experimenter to export raw gaze data and perform statistical analysis in a third-party application (IBM SPSS Statistics).

3.2. Stimuli

Participants viewed search result pages with summaries of varying line lengths and spacing in portrait mode. Each of the 12 conditions had a different stimulus with a target search result and three distractors of the same description length. As shown in Figure 4, for one condition, participants saw four search results with three-line summaries.

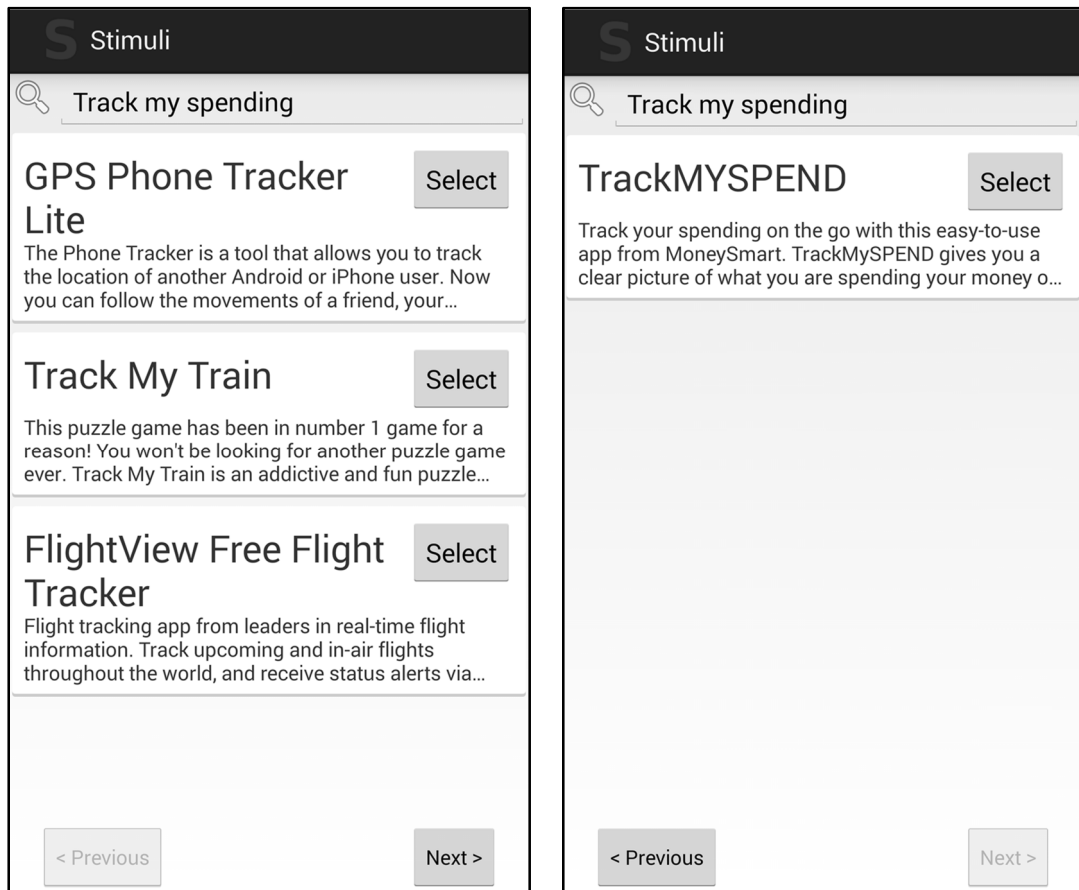


Figure 4. User interface for the three-line, single-spacing condition.

Like in Cutrell and Guan (2007), the target—the “best” result—contained the most searched terms in the summary. For example, for the search query, “track my spending,” the target was an application entitled TrackMYSPEND. Each level of summary length (one, three, seven, ten lines) for this search result contained the phrase, “Track your spending.” The other three applications in this condition were: Track My Train, GPS Phone Tracker Lite, and FlightView Free Flight Tracker. So while every result contained the word “track,” only one result matched the imputed “goal” of the search query string.

All participants experienced all 12 conditions in a single session, completing one trial per condition (for a total of 12 trials). The order of experimental conditions was randomized. No search task (e.g., track my spending) was repeatedly shown for the same condition (e.g., always the seven-line, double-spaced condition). The presentation of applications listed for a given search task and the location of the target result were also randomly assigned.

For six randomly selected trials, a post-task questionnaire prompted participants to select their preferred description (i.e., “Tap the description length you prefer.”) from four summary lengths (one, three, seven, and ten lines) with the same line spacing as that trial (see Figure 5). The summaries were based on the mobile application the participant had previously chosen for the search task. The presentation of summary lengths for preference was also randomized for each trial.

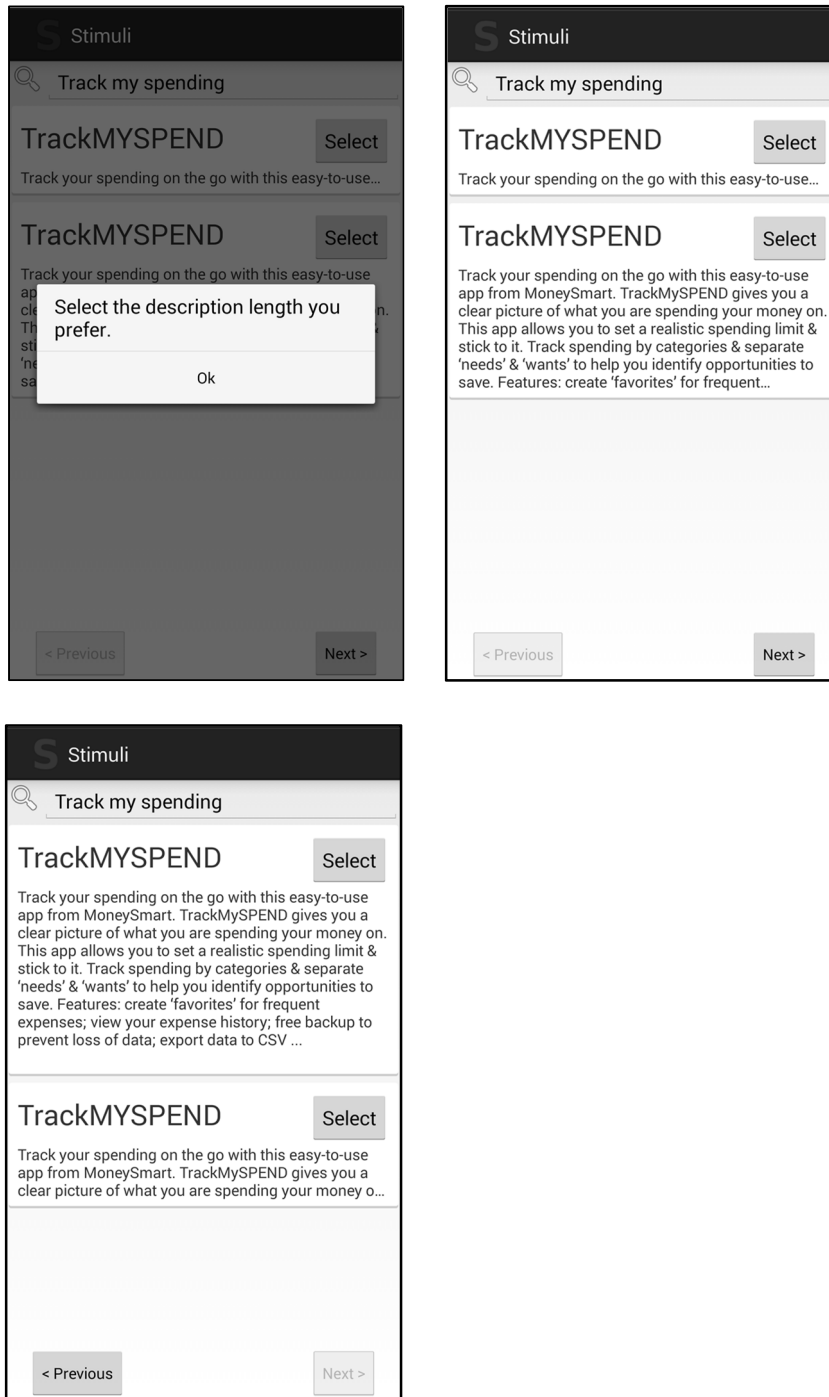


Figure 5. Post-task questionnaire for preference selection. The native mobile application prompted participants to select their preferred description length from four summary lengths (one, three, seven, and ten lines).

3.3. Procedure

1. Prior to beginning, the experimenter presented participants with an Agreement to Participate in Research.
2. At the start of the study, the experimenter fitted the eye-tracking glasses to the participant's head.
 - a. The experimenter tightened the head strap to prevent the glasses from accidentally dropping.
 - b. The experimenter adjusted the glasses until the participant's eyes were positioned concentrically within the frame of the glasses.
 - c. In the event that the warning "insufficient trackability for the subject" appeared, the experimenter changed the nose rest. After changing the nose rest, the experimenter adjusted the eye position by moving the SMI-ETG up or down the bridge of the participant's nose until the participants' pupils were positioned concentrically within the infrared lights (as shown in Figure 3).
 - d. Before starting the calibration process, the experimenter waited 30 s to ensure that the internal eye model had been completely adapted.
3. To improve accuracy, the experimenter performed a three-point calibration prior to the first search task.
 - a. When the experimenter moved the cursor over the scene video view (which showed where the participant was looking), crosshairs appeared next to the gaze cursor (see Figure 6). The numbers beside the crosshairs represented the number of points that still needed to be calibrated out of the total points

necessary for the calibration process. In the case of the first point of a three-point calibration process, the number was 1/3.

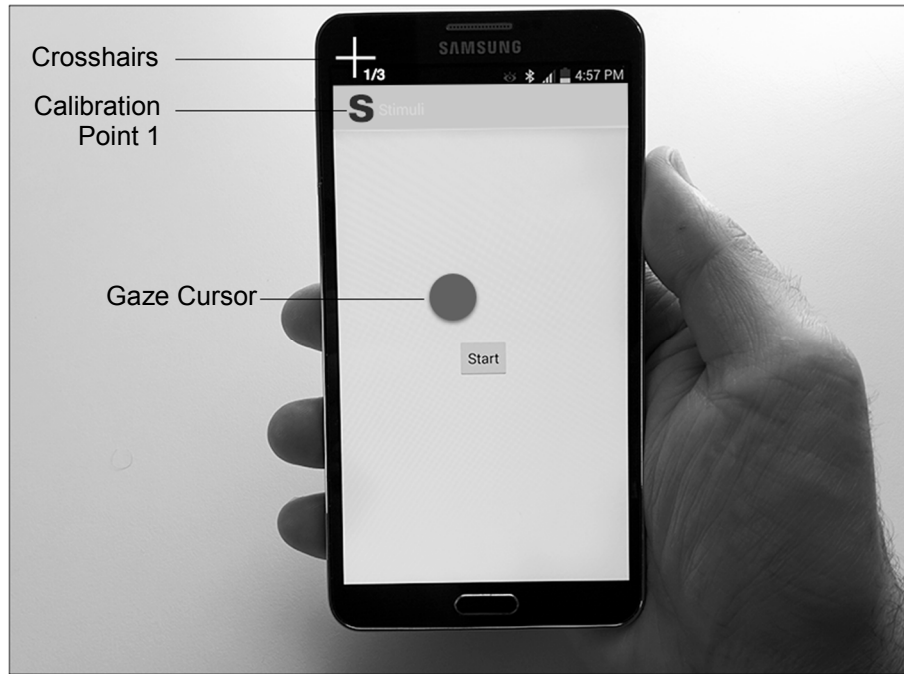


Figure 6. Calibration process for the first point.

- b. The experimenter asked the participant to look at the first calibration point—the big bold “S” in the top left of the screen shown in Figure 6. When the participant fixed her gaze on the target letter, the experimenter clicked with the crosshairs on this point. The calibration process involved correcting the gaze cursor’s offset to the crosshairs position.
- c. The experimenter repeated this process two more times, using the navigation button at the bottom of the device (below the touchscreen) and the start button in the middle of the page as the other two calibration points.
- d. All calibration points were on the same plane, approximately 1.5 m from the participant.

- e. As shown in Figure 7, calibration was considered successful when the crosshairs and gaze cursor aligned, and the calibration status in the video recording display showed “3-Point Calibrated.”

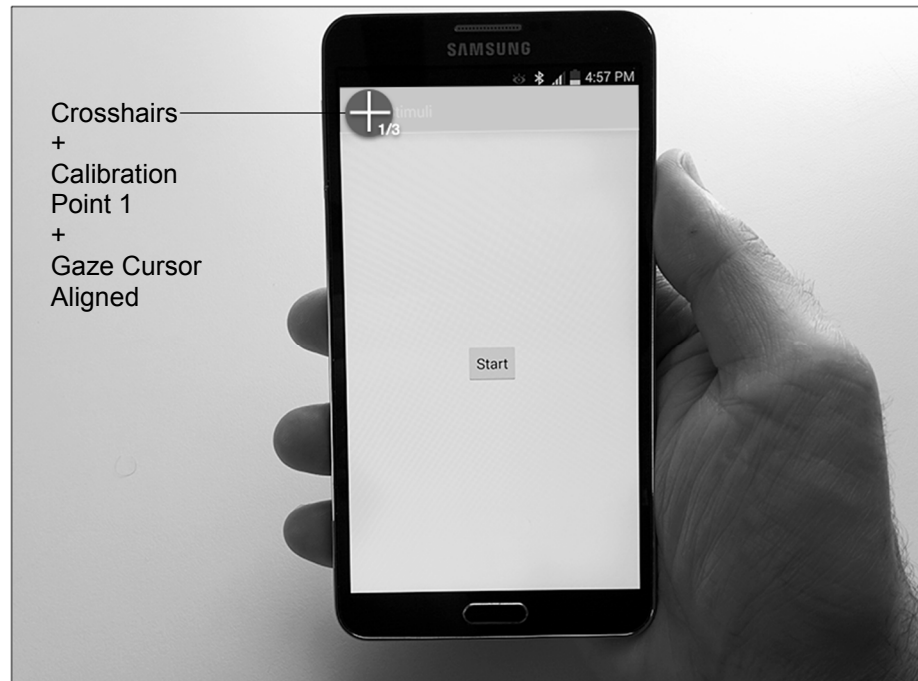


Figure 7. Successful calibration point. This illustration shows the gaze cursor’s offset corrected to the crosshairs position.

- f. The experimenter validated the accuracy of calibration by asking the participant to read out loud the experimental instructions displayed on the following screen. When the gaze cursor matched the instructional text that the participant was reading, the experimenter started the recording session.

4. For each trial, the control page provided participants with a brief search description and motivation for a task (shown in Figure 8).

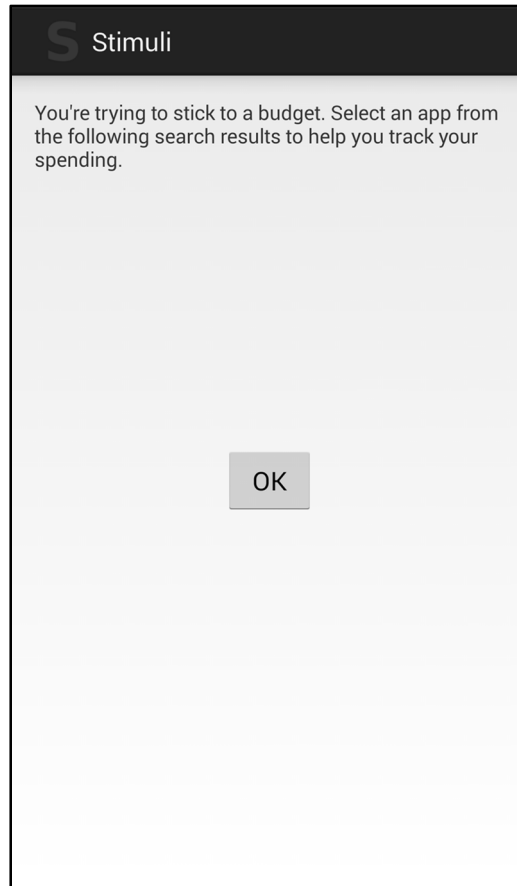


Figure 8. Trial control page.

5. Decision time began when the participant tapped the OK button in the middle of the control page and ended when the participant tapped the Select button on an application summary.
6. Each search task was considered complete when the participant tapped the Select button.

7. For six randomly selected trials of 12, a modal screen cued participants to select their preferred description length (one, three, seven or ten lines) for the application they chose in the search task with the same line spacing as that trial (see Figure 5).
8. Over a 15-minute session, participants performed 12 search tasks on a phablet while wearing eye-tracking glasses. Figure 9 illustrates the experimental procedure outlined above as a flowchart. Other researchers have used similar procedures for eye-tracking studies on information-seeking behavior (Cutrell & Guan, 2007).

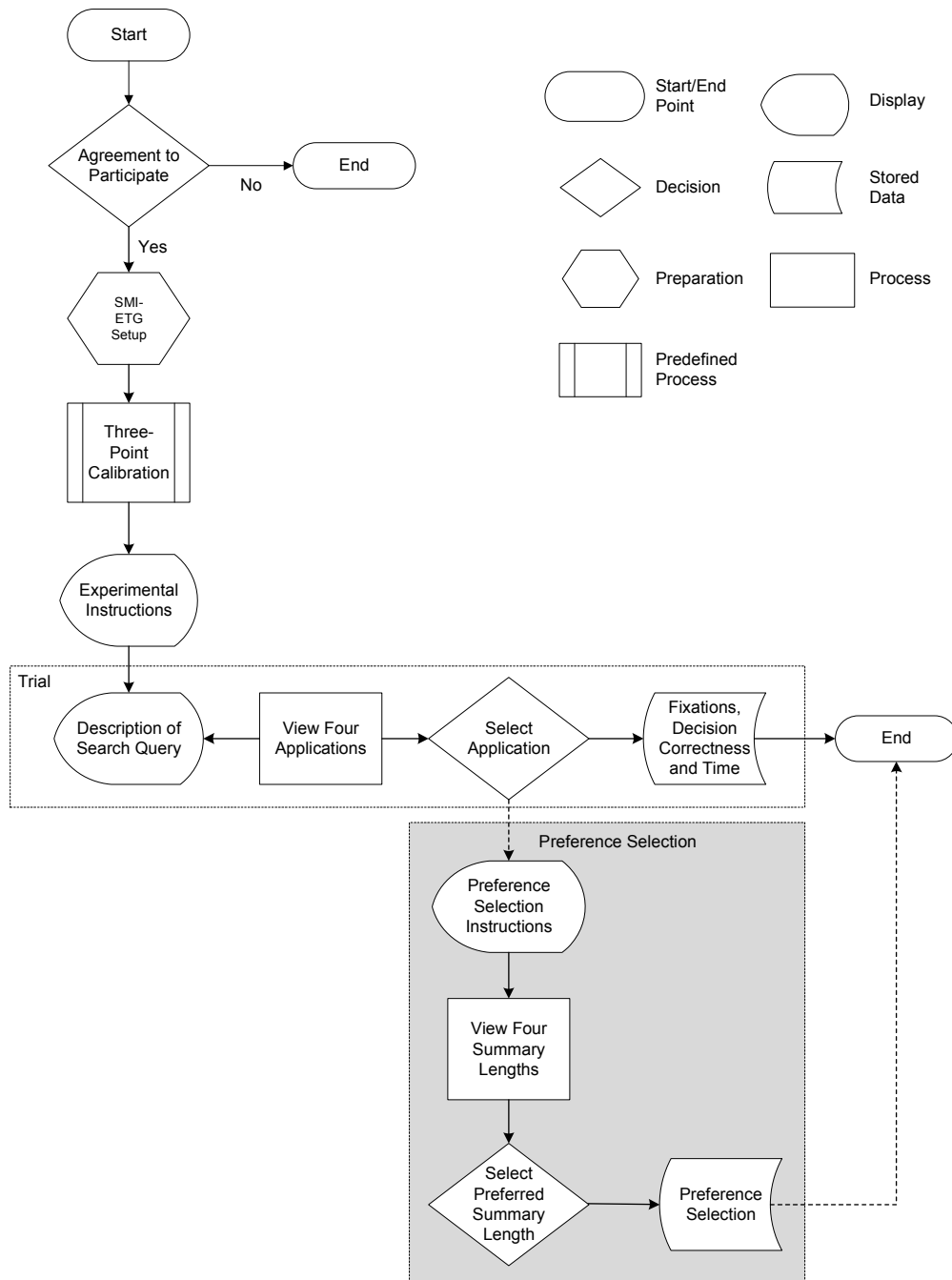


Figure 9. Flowchart of experimental procedure for a study session. For six randomly selected trials of 12, participants were prompted to input their preference selection.

3.4. Dependent Measures

The dependent measures were fixations, decision time, correctness, and preference. When the participant tapped the OK button in the middle of the trial control page in Figure 8, SMI BeGaze software started detecting fixations and the native mobile application began measuring decision time and evaluating decision correctness. SMI BeGaze software detected a fixation after a participant steadily gazed in an AOI with a radius of 50 pixels for at least 100 ms. Figure 9 shows data collection points for these measures, specifically at the end of each trial. The mobile application tallied participant preferences for six randomly selected trials.

4. Results

4.1. Fixations

A two-way (summary length x line spacing) repeated-measures ANOVA was conducted to determine whether there was a statistically significant difference in the number of fixations across the 12 conditions ($N = 36$). The data were not normally distributed in the 12 conditions, as assessed by Normal Q-Q Plots and the Shapiro-Wilk test ($p < .01$). For the summary-length factor, epsilon (ϵ) was 0.86, as calculated according to Huynh-Feldt, and was used to correct the statistical results of the two-way repeated measures ANOVA. Descriptive statistics in Table 3 suggest a direct relationship between fixations and summary length.

Table 3.

<i>Descriptive Statistics for Number of Fixations</i>		
Length ^a	<i>M(SD)</i>	95% CI
Single		
1 Line	39.19(25.12)	[30.69, 47.70]
3 Lines	55.08(29.51)	[45.10, 65.07]
7 Lines	65.67(38.02)	[52.80, 78.53]
10 Lines	86.31(52.81)	[68.44, 104.17]
One and a Half		
1 Line	35.03(15.77)	[29.69, 40.36]
3 Lines	50.92(41.01)	[41.02, 60.82]
7 Lines	85.44(70.69)	[61.52, 109.36]
10 Lines	78.94(59.24)	[58.90, 98.99]
Double		
1 Line	35.36(20.34)	[28.48, 42.24]
3 Lines	50.28(26.68)	[41.25, 59.31]
7 Lines	79.19(55.12)	[60.55, 97.84]
10 Lines	81.19(68.25)	[58.10, 104.29]

^a*n* = 36

A two-way repeated-measures ANOVA yielded a main effect for summary length, $F(2.59, 83.98) = 24.58, p < .01$, partial $h^2 = .41$. Post-hoc analysis with a Bonferroni adjustment revealed statistically significant increases in the number of fixations: from one line to three lines, (15.57 (95% CI, 4.57 to 26.56), $p < .01$), from one line to seven lines (40.24 (95% CI, 23.38 to 57.10), $p < .01$), from one line to ten lines (45.62 (95% CI, 26.80 to 64.44), $p = .01$), and from three lines to ten lines (30.06 (95% CI, 10.78 to 49.33), $p = .01$). For the line-spacing factor, epsilon (ϵ) was 1.00, as calculated according to Huynh-Feldt. The main effect of line spacing was not significant ($F(2.00, 70.00) = .04, p = .97$, partial $h^2 = .00$). For the interaction effect, epsilon (ϵ) was .71, as calculated according to Huynh-Feldt, and was used to correct the statistical results of the two-way repeated measures ANOVA. No interaction effect was found ($F(4.27,$

149.49) = 1.11, $p = .35$, partial $\eta^2 = .03$), indicating that the effect of summary length on fixations was not significantly different across the line spacing levels (see Figure 10).

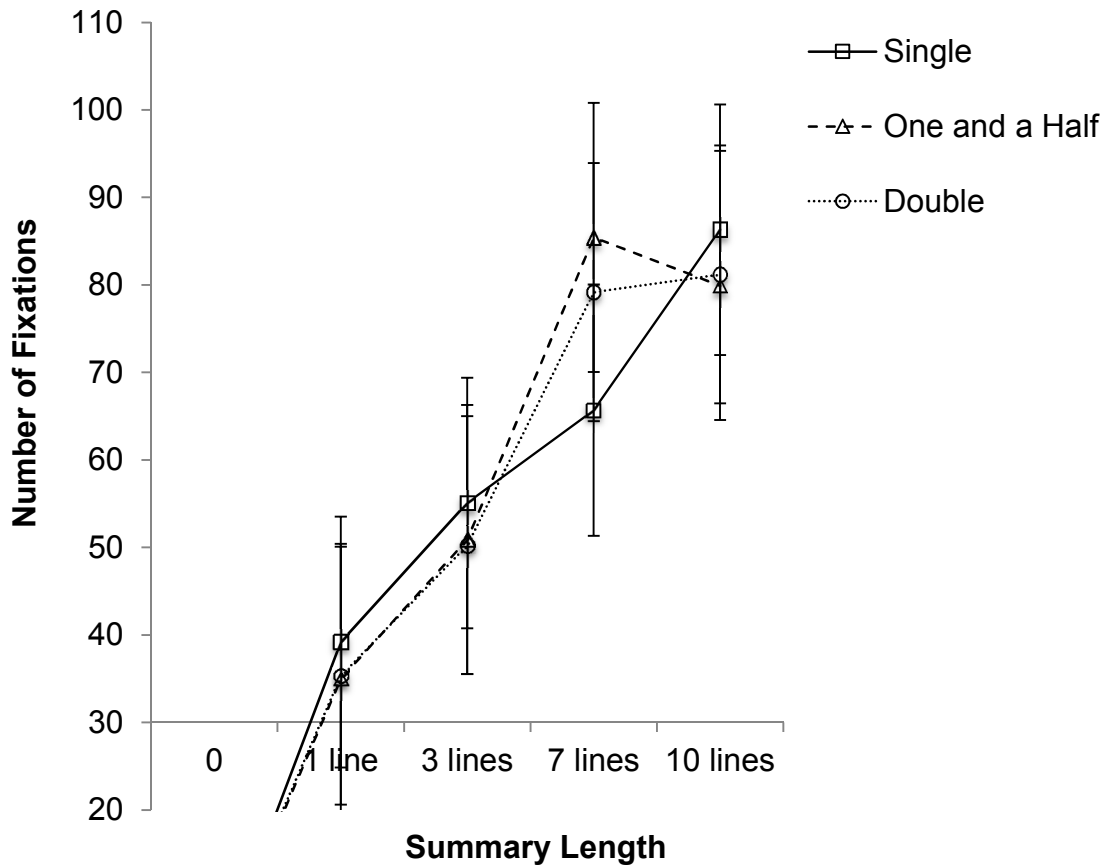


Figure 10. Interaction plot of raw data for number of fixations. The graph shows a main effect of summary length on the number of fixations.

4.2. Decision Time

A two-way (summary length x line spacing) repeated measures ANOVA was conducted to determine whether there was a statistically significant difference in decision time across the 12 conditions ($N = 56$). The data were not normally distributed in nine of the 12 conditions, as assessed by Normal Q-Q Plots and the Shapiro-Wilk test ($p < .05$). For the summary-length factor, epsilon (ϵ) was .87, as calculated according to Huynh-

Feldt, and was used to correct the statistical results of the two-way repeated measures ANOVA, which yielded a main effect, $F(2.61, 143.70) = 35.13, p < .01, \text{partial } h^2 = .39$. Descriptive statistics in Table 4 suggest a direct relationship between summary length and decision time.

Table 4.

Descriptive Statistics for Decision Time (s)

Length ^a	<i>M</i> (<i>SD</i>)	95% CI
Single		
1 Line	23.15(11.65)	[20.02, 26.27]
3 Lines	29.01(13.83)	[25.31, 32.72]
7 Lines	34.02(18.35)	[29.10, 38.93]
10 Lines	37.34(20.28)	[31.91, 42.77]
One and a Half		
1 Line	22.59(8.39)	[20.34, 24.83]
3 Lines	29.04(17.71)	[24.30, 33.78]
7 Lines	37.81(25.53)	[30.98, 44.65]
10 Lines	36.65(22.22)	[30.70, 42.60]
Double		
1 Line	22.20(11.79)	[19.05, 25.36]
3 Lines	26.29(13.41)	[22.70, 29.88]
7 Lines	38.48(19.82)	[33.17, 43.78]
10 Lines	38.66(26.82)	[31.48, 45.84]

^a*n* = 56

Post-hoc analysis with a Bonferroni adjustment of the data revealed statistically significant increases in decision time: from a single line to three lines, (5.47 (95% CI, 1.78 to 9.16), $p < .01$), from a single line to seven lines (14.12 (95% CI, 9.43 to 18.82), $p < .01$), from a single line to ten lines (14.90 (95% CI, 9.10 to 20.70), $p < .01$), from three lines to seven lines (8.65 (95% CI, 4.82 to 12.49), $p < .01$) and from three lines to ten lines (9.43 (95% CI, 4.56 to 14.31). For the line-spacing factor, epsilon (ϵ) was 1.00, as calculated according to Huynh-Feldt. The main effect of line spacing was not significant

($F(2.00, 110.00) = .14, p = .87, \text{partial } h^2 = .00$). For the interaction effect, epsilon (ϵ) was .80, as calculated according to Huynh-Feldt, and was used to correct the statistical results of the two-way repeated measures ANOVA. No significant interaction effect was found, ($F(4.79, 263.19) = .79, p = .55, \text{partial } h^2 = .01$), indicating that the effect of summary length on decision time was not significantly different across the line-spacing levels (see Figure 11).

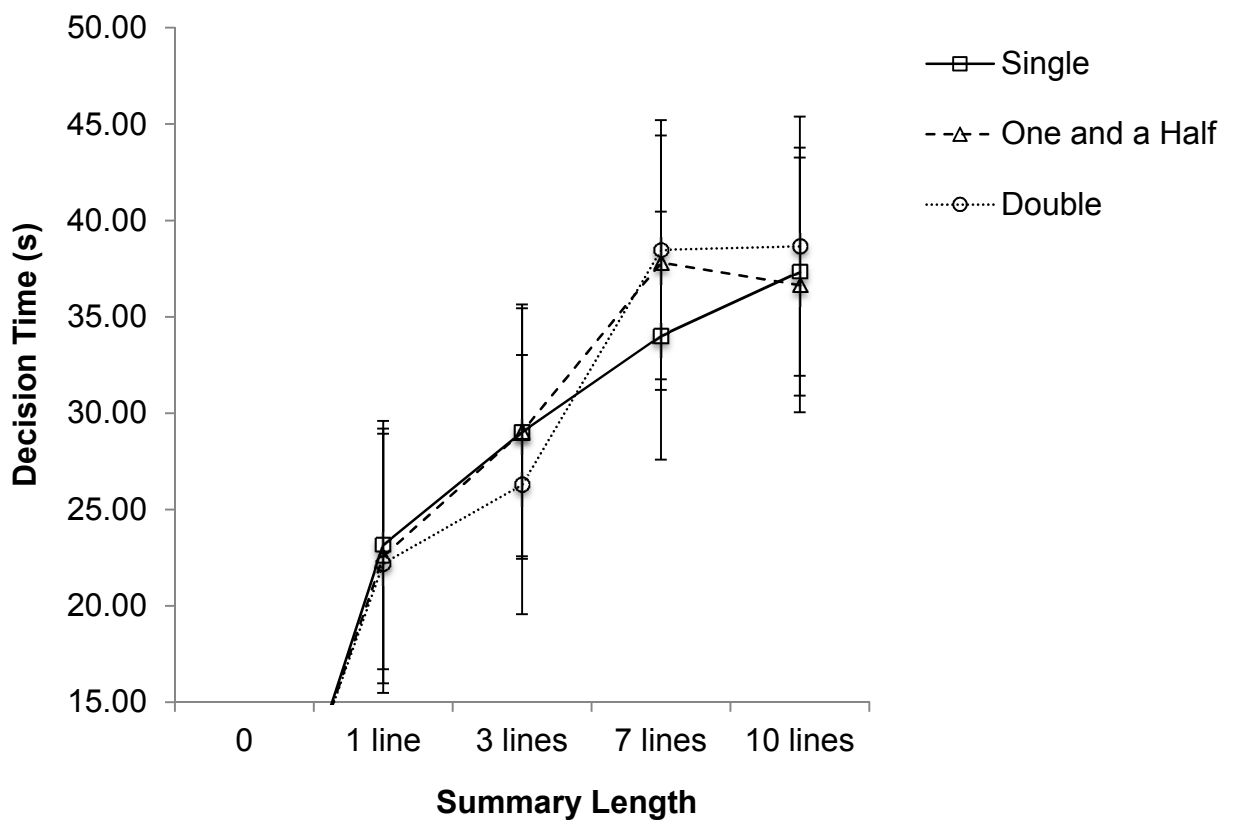


Figure 11. Interaction plot of raw data for decision time (s). The graph shows a main effect of summary length on decision time.

4.3. Decision Correctness

Participants completed three trials for each level of the summary-length factor. The three line-spacing conditions were collapsed together so that the percent correct could be calculated for each level of summary length. A one-way repeated measures ANOVA was conducted to determine whether there was a statistically significant difference in decision correctness across the four summary-length conditions ($N = 56$). The data were not normally distributed across these conditions, as assessed by Normal Q-Q Plots and the Shapiro-Wilk test ($p < .0005$). Epsilon (ϵ) was 1.00, as calculated according to Huynh-Feldt. Descriptive statistics in Table 5 suggest that participants made the most correct decisions when shown ten-line summaries ($M = 84.00$, $SD = 24.67$), with decision correctness decreasing from the ten-line condition to the one-line condition ($M = 81.00$, $SD = 24.67$) to the seven-line condition ($M = 78.00$, $SD = 24.00$) to the three-line condition ($M = 75.67$, $SD = 26.67$). However, there was not a statistically significant difference between the means at the different summary lengths, $F(3, 165) = 1.43$, $p = .24$, $h^2 = .03$ (see Figure 12).

Table 5.

Descriptive Statistics for Decision Correctness

Length ^a	$M(SD)$ %	95% CI
1 Line	81.00(24.67)	[74.33, 87.67]
3 Lines	75.67(26.67)	[68.33, 82.67]
7 Lines	78.00(24.00)	[71.67, 84.33]
10 Lines	84.00(24.67)	[77.33, 90.67]

^a $n = 44$

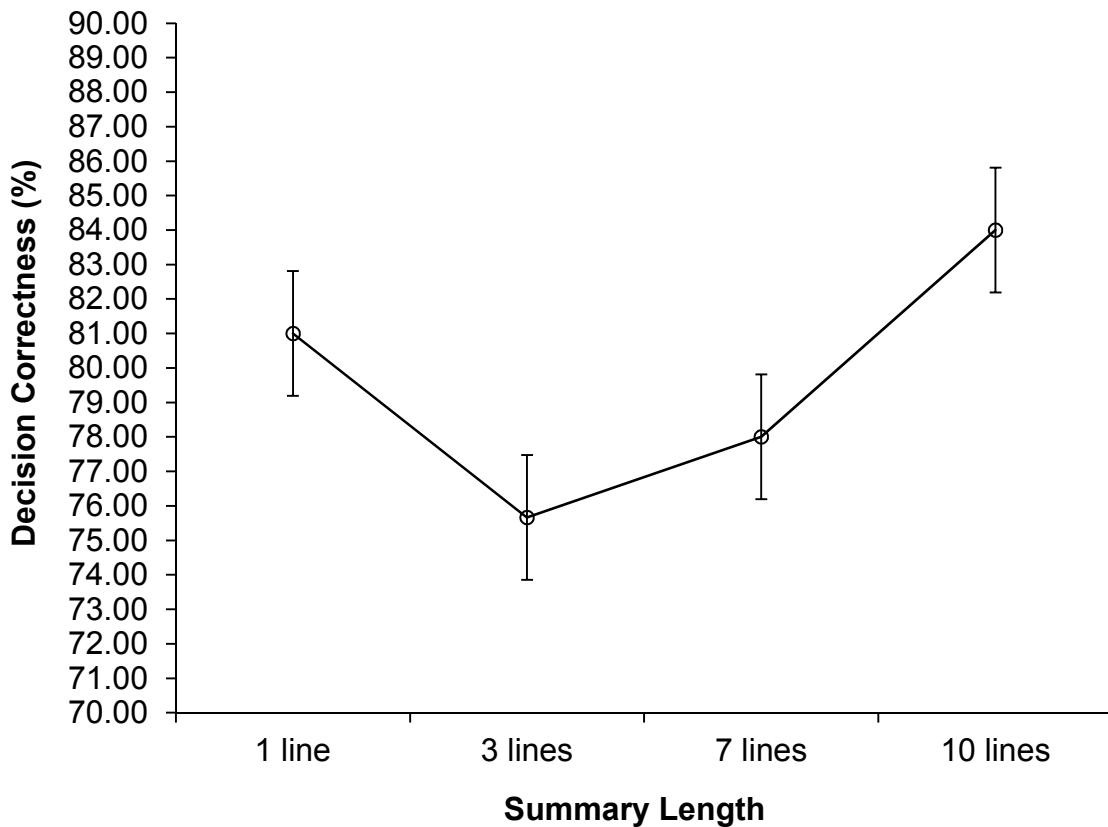


Figure 12. Line graph of percent correct for each level of summary length.

4.4. Preference

For half (six) of the 12 trials, a post-task survey prompted participants to select their preferred description length (i.e., “Tap the description length you prefer.”) from four summary lengths (one, three, seven, or ten lines) with the same line spacing as that trial (see Figure 5). Each participant’s data was then reduced to one data point: his/her most preferred summary length—the one preferred the greatest number of times. Twelve participants’ data were excluded from analysis because they chose equally often two summary lengths ($N = 44$). The histogram in Figure 13 indicates a strong preference for three lines ($n = 20$) followed by seven lines ($n = 14$) and ten lines ($n = 9$). A chi-square

goodness-of-fit test was performed on summary-length tallies in Table 6 to determine whether the four lengths were equally preferred. The minimum expected frequency was 11.00. The chi-square goodness-of-fit test indicated that preference was significantly different from the expected proportions ($\chi^2(3) = 17.64, p < .01$). Post-hoc paired comparisons revealed that the one-line condition differed significantly from: the three-line condition, $\chi^2(1) = 17.19, p < .01$; the seven-line condition, $\chi^2(1) = 11.27, p < .01$; and the ten-line condition, $\chi^2(1) = 6.40, p < .01$. The three-line condition did not differ significantly from the seven-line condition, $\chi^2(1) = 1.06, p = .30$, but did differ significantly from the ten-line condition, $\chi^2(1) = 4.17, p = .04$. The seven-line condition did not differ significantly from the ten-line condition, $\chi^2(1) = 1.09, p = .30$.

Table 6.

Descriptive Statistics for Preference

Length ^a	Observed Frequency
1 Line	1
3 Lines	20
7 Lines	14
10 Lines	9

^a $n = 44$

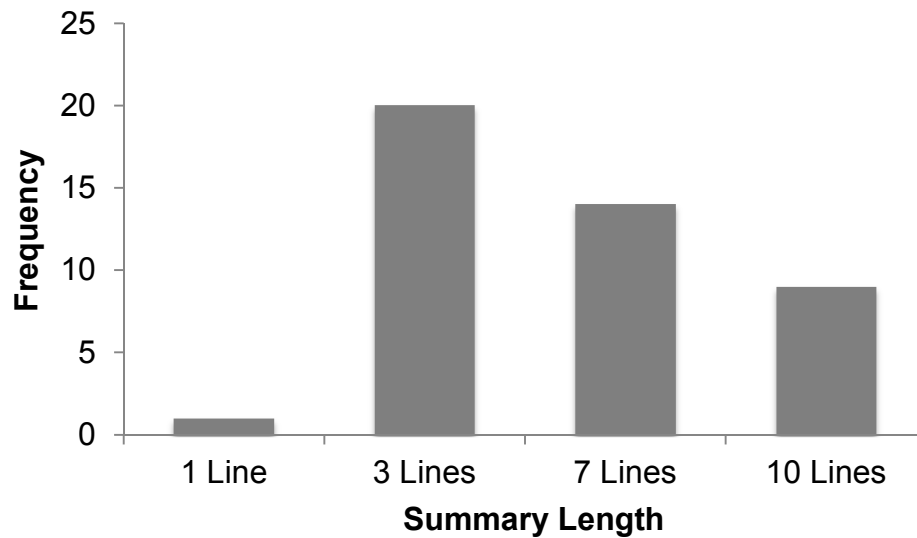


Figure 13. Frequency distribution of raw data for length preference. The graph illustrates participants' first preference choice of three lines, $N = 44$.

5. Discussion

In terms of an optimal summary length for making relevance assessments, Sweeney and Crestani (2006) found no effect of summary length on measures of accuracy for a PDA with a 4.8-inch screen. So the authors suggested adapting the default summary length to user preference: three lines. We also did not find a relationship between summary length and accuracy on a phablet with a 5.7-inch display. In keeping with participant preference, we recommend showing three-line summaries for mobile search results.

5.1. Summary Length

5.1.1. Fixations. We hypothesized that the number of fixations would decrease as summary length increased—up to seven lines—after which performance would deteriorate. With additional contextual information, fewer and longer fixations would be

required to make relevance judgments, as participants could take in more with each fixation. Accordingly, this line of reasoning suggests that more text—up to seven lines—would improve scanning effectiveness. On the other hand, ten lines would require too much cognitive effort, as participants would likely read each word (Cutrell & Guan, 2007).

The results of this study showed a direct relationship between the number of fixations and summary length. Fixation frequency significantly increased from one to three lines, from one to seven lines, from one to ten lines, from three to seven lines and from three to ten lines. There was no significant difference between seven- and ten-line conditions, though. In sum, longer summaries led to more fixations. Since fixation frequency serves as a marker of visual complexity, more fixations suggest that a more careful reading was required for selection confirmation (Goldberg, Stimson, Lewenstein, Scott, & Wichansky, 2002).

5.1.2. Decision time. We hypothesized that decision time would decrease as summary length increased—up to the seven-line condition—after which performance would deteriorate. However, the data indicated a direct relationship between summary length and decision time; as summary length increased, participants took longer to make a decision. Decision time significantly increased from one line to three, from one line to seven, from one line to ten, from three lines to seven and from three lines to ten. Of note, there was no difference in decision time for the seven- and ten-line conditions. These results are inconsistent with the findings of Cutrell and Guan (2007), who found that

decision times on informational tasks decreased with seven-line summaries versus shorter summaries (one line and two to three lines).

One possible explanation for an increased decision time involves the semantic nature of search tasks. Fernandes (2010) found that search tasks with a semantic description of a target word (e.g., type of fruit) required more fixations and time than searching for a specific word (e.g., apple). For the more difficult search tasks in this study, distractors also contained searched terms, so the participant could not identify the target without paying attention to the meanings of words. In Cutrell and Guan (2007), participants may have scanned search results for query terms rather than reading and interpreting longer summaries.

For example, for the plan-a-trip-to-the-movies stimulus, more than one result contained searched terms, requiring the participant to examine each word in context (see Figure 14). Each distractor included a searched word: TripAdvisor Hotels Flights and Field Trip application descriptions contained “trip.” Hong Kong Movie and Fandango had “movie” in their summaries.

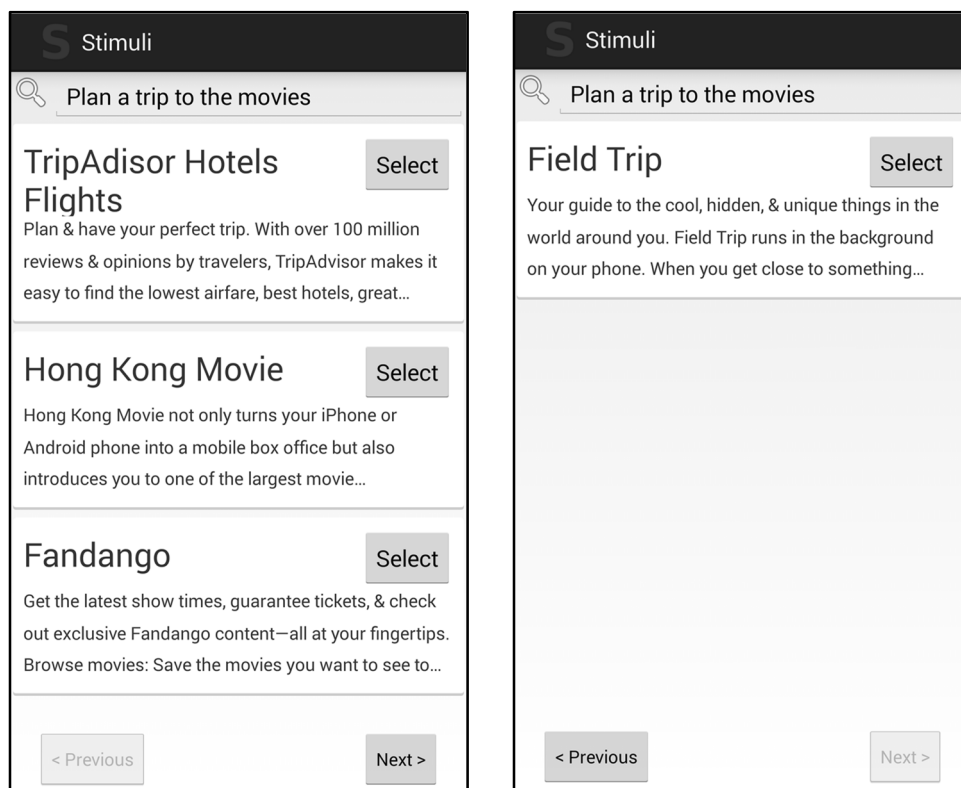


Figure 14. Search results for the three-line, one-and-a-half-spacing condition.

5.1.3. Decision correctness. Despite a harder examination of the summaries, decision correctness did not significantly differ for the four summary lengths. The finding of no relationship between summary length and decision correctness is consistent with the results of Cutrell and Guan (2007) and Sweeney and Crestani (2006). Although Cutrell and Guan's (2007) data suggested that click accuracy improved as summary length increased for informational tasks—from 28 to 39%—statistical tests were not significant due to a small sample size (18) (p. 8). Even with triple the sample size, we did not find evidence corroborating their hypothesis. On a PDA, Sweeney and Crestani (2006) found that varying summary lengths did not significantly affect accuracy.

Performance was consistent across all summary lengths (title, 7, 15, and 30% of the document; up to one, three, six, and twelve lines respectively).

5.1.4. Preference. It was hypothesized that participants would prefer longer summaries (greater than or equal to three lines). According to this logic, longer summaries would allow participants to more effectively scan search results for keywords. While the histogram in Figure 13 indicated a strong preference for the three-line condition, no statistically significant difference was found between the three- and seven-line conditions. However, the finding that participants preferred three-line summaries to one- and ten-line summaries is consistent with the results of a post-experiment questionnaire in Sweeney and Crestani (2002). Participants preferred using the title and short summaries (up to maximum of three lines) for making relevance decisions on micro and small displays (cell phones and PDAs) and longer summaries (six and twelve lines) on normal displays (desktop PCs).

5.2. Line Spacing

5.2.1. Fixations. Prior eye-tracking data in Tarling and Brumby (2010) revealed that participants made fewer eye movements in single-spaced than double-spaced conditions, as more characters could be assessed in a single fixation. However, no significant difference in fixations across line-spacing conditions was found. One possible reason for this discrepancy involves the screen size. Tarling and Brumby (2010) presented stimuli on a 17-inch desktop monitor whereas the present study used a 5.7-inch phablet screen. By virtue of its smaller size, layout density is greater on the phablet across all conditions.

5.2.2. Decision time. Ling and Schaik (2007) reported faster reaction times with double spacing for visual search tasks using desktop monitors. However, we did not find significant differences in decision times across line-spacing conditions in the present phablet study. Again, a narrower screen necessitates an increase in search item density, which may have made text layout changes less noticeable.

Overall, the key findings indicate:

- A direct relationship between summary length, fixations, and decision time; as summary length increases, fixations and decision time increase.
- No relationship between decision correctness and summary length.
- Participants preferred three-line summaries for evaluating the relevance of a search result on a 5.7-inch screen. Although preference for the three-line condition did not differ significantly from the seven-line condition, shorter summaries minimize the attentional costs associated with longer summaries—more eye movements and slower decision times. Fixations significantly increased in all conditions except for three to seven lines. Decision time significantly increased in every condition except for seven to ten lines.

5.3. Limitations

This study is limited in the extent to which it may conclude the optimal amount of information a user wants to see when searching on his/her phablet. We investigated the inconclusive results of prior work by Cutrell and Guan (2007), who suspected a direct relationship between summary length and click accuracy on a desktop PC. Like Sweeney

and Crestani (2006) before them, we found no relationship between summary length and decision correctness.

5.3.1. Poor track ability. Twenty participants' eye-tracking data (14 female, six male) were excluded from analysis due to poor tracking, reducing the sample size from 56 to 36. The SMI-ETG indicated poor track ability, due as eyelashes, drooping upper eyelids or lower eyelid fullness, which can obscure the pupil. For example, picture A (left) in Figure 15 shows a participant's lower eyelid and eyelashes blocking half of the infrared light rays, resulting in poor tracking. We excluded this participant's data from analyses. In contrast, picture B (right) shows another participant's pupil positioned concentrically within the six infrared light rays, allowing for more precise gaze tracking.

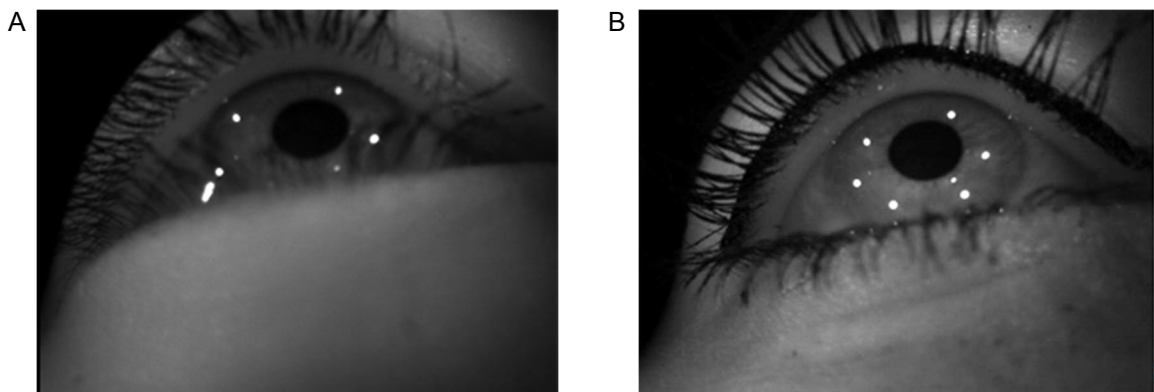


Figure 15. Screen captures from video recorded by an infrared camera in the lens of eye-tracking glasses.

A reduction in sample size may have yielded a loss of statistical power. To this end, an a priori power analysis using the program G*Power indicated that a sample size of 30 would be sufficient to detect a medium effect with a power of .80 and an alpha of .05. With 36 participants, it is unlikely non-significant results can be attributed to sample size.

5.3.2. Pagination. The Stimuli application used pagination instead of scrolling to show longer summaries. In a pilot study, we found that scrolling challenged the accuracy of gaze tracking; the eye tracker failed to report fixation coordinates on a scrolling web page. Pagination better supported the eye-tracking requirement of 100 ms (minimum) in an AOI. However, pagination most likely increased decision time; tapping Previous and Next buttons at the bottom of the page and waiting for a new screen to load required more time than scrolling.

Eight out of 12 conditions had multiple user-interface screens; three-line summaries had two pages, seven-line summaries had three pages, and ten-line summaries had four pages (as shown in Figure 16). Pagination negatively impacted application usability, as paging was inconsistent with user expectations (scrolling) and required learning. Participants may not have considered summaries on different pages in their initial relevance decisions. So participants may have been more accurate when one line was shown (compared to three and seven lines), as descriptive statistics suggested (see Table 5). Decision correctness was not statistically different for lines shown ($p = .24$). Because p-values far exceeded the predetermined level of significance (.05), it is unlikely that non-significant results are due to this possible confound.

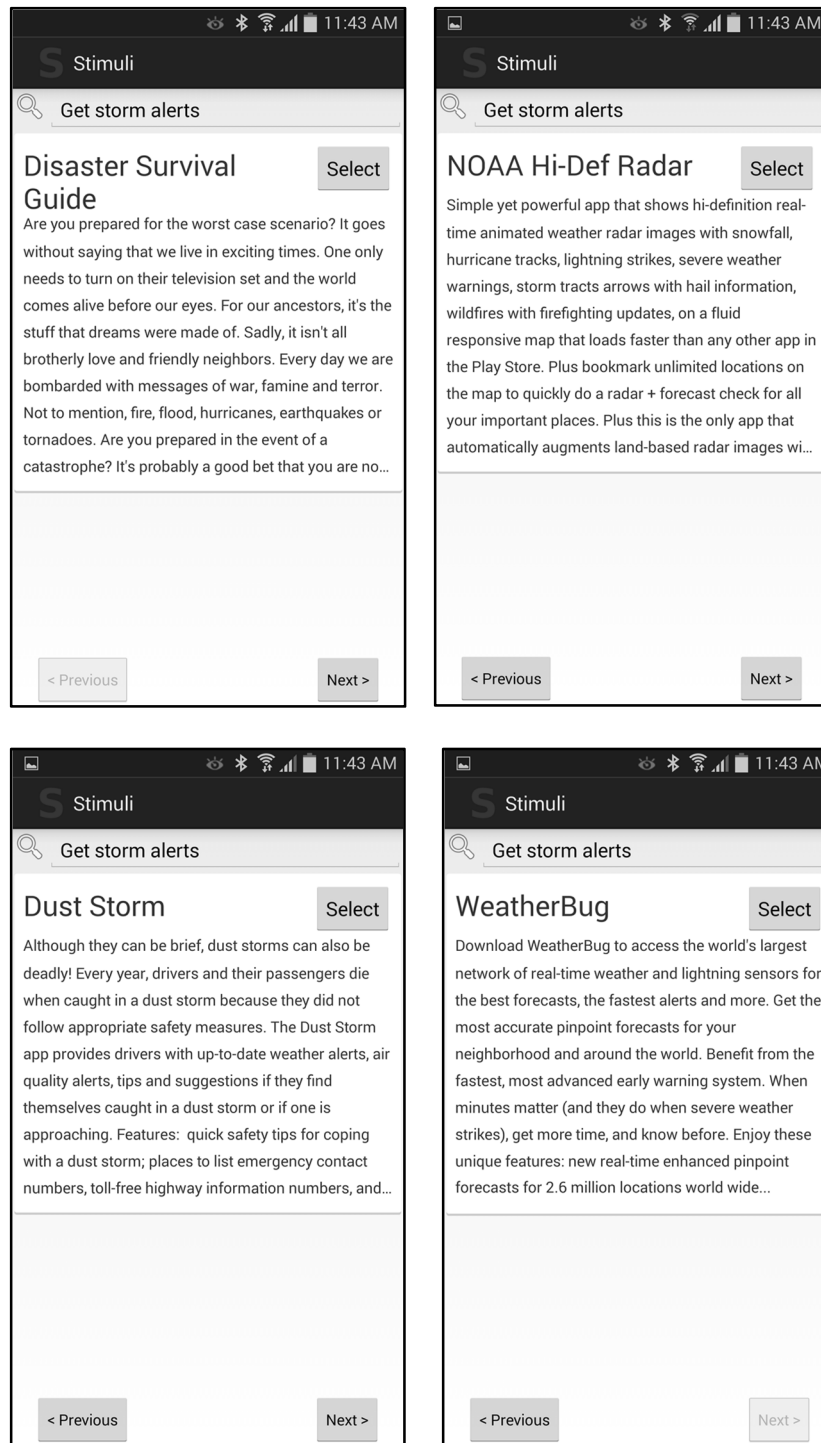


Figure 16. Ten-line, one-and-a-half-spacing condition.

5.3.3. No visible difference between one-line, line-spacing conditions. One-line summaries looked the same across all line-spacing conditions; the one-line, single-spacing condition appeared identical to the one-line, one-and-a-half spacing condition and the one-line, double-spacing condition (see Figure 17). Counting the one-line conditions as three different conditions may explain, in part, non-significant results for line spacing. However, eliminating one-line summary data does not significantly affect the results. We repeated the two-way (summary length x line spacing) repeated-measures ANOVAs without the one-line summary-length condition. The main effect of line spacing was still not significant for fixations and decision time, and no interaction effect was found.

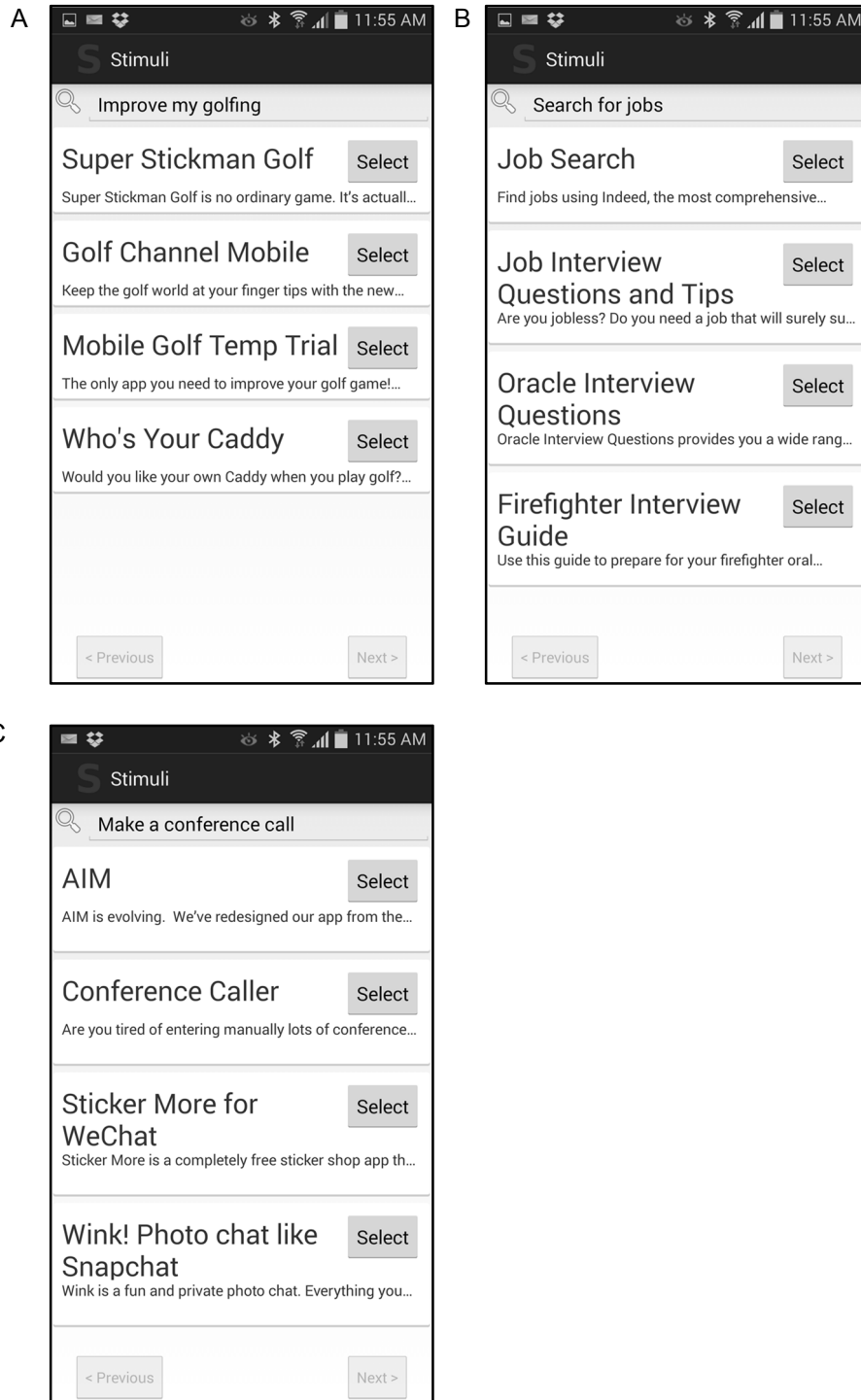


Figure 17. One-line, (A) single-, (B) one-and-a-half-, and (C) double-spacing conditions.

5.3.4. Task difficulty. It is possible that stimuli were not uniform in their level of difficulty; some search results may have been harder than others to judge as relevant, given the hypothetical search scenario. So the “best” answer was not always obvious to the participant. For example, half (28) of all participants (56) misjudged search results for the get-wine-recommendations stimulus:

You’re at a team dinner hosted by your boss. The wine menu arrives, and no one wants the burdensome task of picking a wine—but everyone wants a glass of it. Select an app that suggests food and wine pairings. You’re having pasta for dinner.

Figure 18 shows the four search results for this search task.

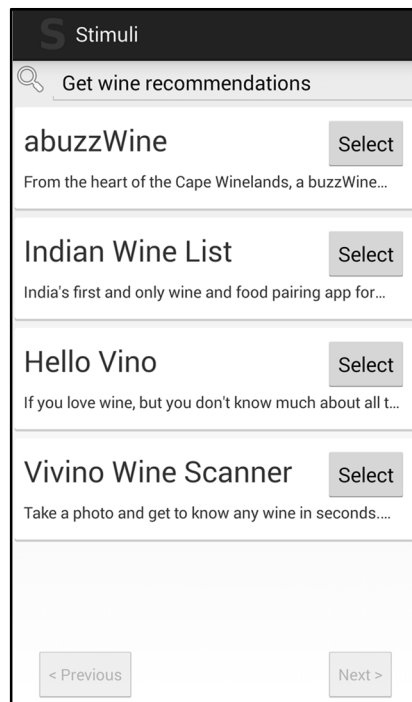


Figure 18. Search results for the one-line, one-and-a-half-spacing condition for the get-wine-recommendations stimulus.

Of the 28 participants who selected the wrong application, nine participants saw these search results as one-line summaries, 13 saw three-line summaries, three saw seven-line summaries and four saw ten-line summaries. This data indicates a direct relationship between summary length and decision correctness, although no statistically significant relationship was found in this study. Participants who saw summaries longer than three lines may have been more accurate in their relevance judgments than those who saw shorter summaries. The one-line summary of the Hello Vino application did not indicate that it would satisfy the user's information need—food and wine pairings. Additional information provided in summaries longer than three lines included:

Hello Vino has the largest food- and wine-pairing database in the App store. Food and wine Pairings: simply choose your dish, then the app will tell you the best wine for the meal. Recommendations: this app will help you before you buy. It's made for the 99% of wine shoppers who need some help in the store or restaurant. Simply tell the app what you're eating, or provide some general taste preferences, and Hello Vino will recommend the perfect bottle for every occasion.

Of the 28 participants who correctly judged search results for this scenario, eight participants saw one-line summaries, one saw three-line summaries, nine saw seven-line summaries and nine saw ten-line summaries. So 64% of participants who correctly selected the Hello Vino application saw summaries greater than three lines compared to 25% of participants who misjudged its relevance. Since the application summary does not specifically refer to food and wine pairings until the sixth line, there was no way for participants who saw shorter summaries to judge confidently this application as relevant.

In keeping with this line of thinking, task difficulty may be inextricably linked to the number of target query terms and their location in summaries. However, when

searched terms were in the application title and the first line of the summary, 32% (18 participants) selected the wrong application for the track-my-bike-ride query.

Figure 19 shows the four search results for the one-line, single-spacing condition.

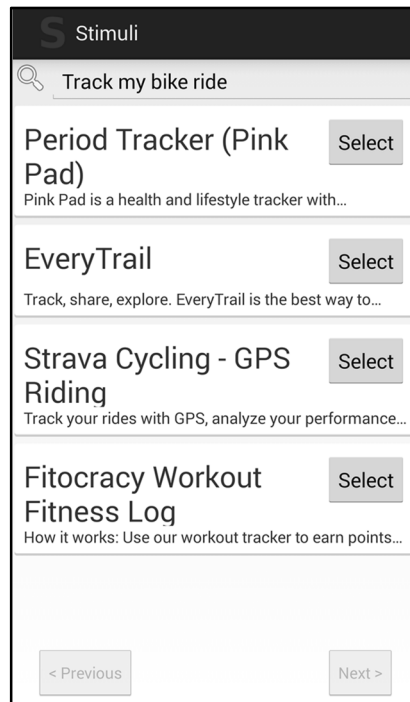


Figure 19. Four search results for the one-line, single-spacing condition.

The application title, Strava Cycling – GPS Riding, and the first four words of the summary, “Track your rides with GPS,” include three query terms: “track,” “ride,” and “GPS.” Still, 89% (16 out of the 18 participants) chose EveryTrail, which included only one query term (“track”) in the one-line summary.

Standardizing task difficulty may have translated to fewer fixations and faster task times as summary length increased. There is sufficient evidence to suggest that providing target query terms rather than a description of an information need would reduce decision

time, as participants would view fewer search results before making a selection (Cutrell & Guan, 2007; Fernandes, 2010).

Establishing inter-rater reliability outside the context of measurement in this study may have helped to standardize task difficulty. For example, several researchers could have rated the questions before the experimenters presented them to participants. The questions with the greatest agreement would comprise the stimuli in the experiment.

5.3.5. Forced-choice rating scale. This study is limited in the extent to which it may confidently conclude the preferred text layout for search results. For six randomly selected trials of a 12, a forced-choice survey question cued participants to select their preferred description (i.e., “Tap the description length you prefer.”) from four summary lengths (one, three, seven, and ten lines) (see Figure 5). The forced-choice rating scale may have biased results by intentionally eliminating those without a definitive preference. To this end, we excluded 12 participants’ data from the analysis because they chose equally often two summary lengths ($N = 44$). Providing an “undecided” option may have significantly improved the accuracy of the survey. We performed a chi-square goodness-of-fit test again, including an “undecided” condition ($N = 56$). The minimum expected frequency was 11.20. The result indicated that preference was significantly different from the expected proportions ($\chi^2(3) = 17.39, p < .01$). Post-hoc paired comparisons revealed that the undecided condition differed significantly from the one-line condition (one line: $\chi^2(1) = 9.31, p < .01$; three lines: $\chi^2(1) = 2.00, p = .16$; seven lines: $\chi^2(1) = .15, p = .70$; ten lines: $\chi^2(1) = .45, p = .51$). This result weakens our

recommendation of three-line summaries, as the three-line condition did not differ significantly from the undecided or the seven-line condition.

5.3.6. Preference for line spacing not calculated. The level of the line-spacing factor was carried over from the search task. So a participant did not choose a summary length independent of line spacing. For example, if a participant experienced the seven-line, single-spacing condition for a search task, she also chose her preferred description length from four single-spaced summary lengths. For this reason, it was not possible to analyze line spacing for preference.

6. Conclusion

In summary, the most compelling finding of this study is the direct relationship between summary length, fixations, and decision time: as summary length increases, fixations and decision time also increase. Given that summary length and line spacing had no effect on decision correctness, the optimal summary length—the one that will impart the least amount of cognitive load on a user’s visual working memory—is one line. However, since participants clearly did not prefer one-line summaries, it is best to show three lines. As such, three-line summaries suggest a minimal tradeoff between performance and preference.

References

- Barredo, A. (2014, May 28). A comprehensive look at smartphone screen size statistics and trends. *Medium*. Retrieved from <https://medium.com/@somospostpc/a-comprehensive-look-at-smartphone-screen-size-statistics-and-trends-e61d77001ebe>
- Chau, M., Mainelli, T., & Shirer, M. (2014, September 3). *A future fueled by phablets – worldwide phablet shipments to surpass portable PCs in 2014 and tablets by 2015* (Research Report). Retrieved from International Data Corporation website. <http://www.idc.com/getdoc.jsp?containerId=prUS25077914>
- Clarke, C. L., Agichtein, E., Dumais, S., & White, R. W. (2007). The influence of caption features on clickthrough patterns in Web search. *Proceedings from SIGIR '07: The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 135-142. doi:978-1-59593-597-7/07/0007
- Cutrell, E., & Guan, Z. (2007). What are you looking for: An eye-tracking study of information usage in Web search. *Proceedings of the CHI '07: The SIGCHI Conference on Human Factors in Computing Systems*, 407-416. doi:978-1-59593-593-9/07/0004
- Fernandes, B. (2010). *Knowing what you're looking for changes how you search: an eye-tracking study into the effect of task description and item density on visual search performance and strategy* (Unpublished master's thesis, University College London, London, United Kingdom). Retrieved from https://www.ucl.ac.uk/ucllc/studying/taught-courses/distinction-projects/2009_theses/FernandesB.pdf
- Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N. S., & Wichansky, A. M. (2002). Eye tracking in Web search tasks: Design implications. *Proceedings of the 2002 Symposium on Eye-Tracking Research and Applications Symposium*, 51-48.
- González-Caro, C., & Marcos, M. C. (2011). Different users and intents: An eye-tracking analysis of Web search. *Proceedings of Web Search and Data Mining*, 9-12. doi:978-1-4503-1/11/02
- Kaiser, M., Hearst, M. A., & Lowe, J. B. (2008). Improving search results quality by customizing summary lengths. *Proceedings of Association of Computational Linguistics-08*, 701-709.

- Kang, T. (2014, November 19). Top 10 smartphones sold globally during Sep 2014 [Blog post]. Retrieved from <http://www.counterpointresearch.com/top10sep2014>
- Ling, J., & van Schaik, P. (2007). The influence of line spacing and text alignment on visual search of Web pages. *Displays*, 28(2), 60-67. doi:10.1016/j.displa.2007.04.003
- Manjoo, F. (2014, September 17). With iPhone 6, Apple aims for versatility. *The New York Times*. Retrieved from <http://www.nytimes.com/2014/09/18/technology/personaltech/review-apple-iphone-6-plus-phablet.html?ref=technology>
- Milanesi, C. (2015). *Apple market share grows across Europe, U.S. and China* (Research Report). Retrieved from Kantar Worldpanel ComTech website: <http://www.kantarworldpanel.com/global/News/Apple-market-share-grows-across-Europe-US-and-China>
- Nielsen, J. (1993). Usability Heuristics. In *Usability Engineering* (2nd ed., pp. 115-155). San Francisco, CA: Morgan Kaufmann.
- Ojanpää, H., Näsänen R., & Kojo, I. (2002). Eye movements in the visual search of word lists. *Vision Research*, 42(12), 1499-1512. doi:10.1016/S0042-6989(02)00077-9
- Pacheco, E. (2012). *Mobile local search volume will surpass desktop local search in 2015* (Research Report). Retrieved from BIA Kelsey website: <http://www.biakelsey.com/Company/Press-Releases/120418-Mobile-Local-Search-Volume-Will-Surpass-Desktop-Local-Search-in-2015.asp>
- Rose, D. E., Orr, D., & Kantamneni, R. G. P. (2007). Summary attributes and perceived search quality. *Proceedings of the 16th International Conference on World Wide Web*, 1201-1202. doi:10.1145/1242572.1242765
- Smith, A. (2015). *U.S. smartphone use in 2015* (Research Report). Retrieved from Pew Research Center website: <http://pewinternet.org/Reports/2013/Smartphone-Ownership-2013/Findings.aspx>
- Sweeney, S. O., Crestani, F., & Tombros, A. (2002). Mobile delivery of news using hierarchical query-biased summaries. *Proceedings of the 2002 ACM Symposium on Applied Computing*, 634-639.
- Sweeney, S., & Crestani, F. (2006). Effective search results summary size and device screen size: Is there a relationship? *Information Processing & Management*, 42(4), 1056-1074.

- Tarling, K. A., & Brumby, D. P. (2010). Density guides visual search: sparse groups are first even when slower. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 54(18), 1311-1315.
- Taylor, B. (2014, July 21). Why smartphone screens are getting bigger: specs reveal a surprising story. *PC World*. Retrieved from <http://www.pcworld.com/article/2455169/why-smartphone-screens-are-getting-bigger-specs-reveal-a-surprising-story.html>