

# Randomised Controlled Trial of the Usability of Major Search Engines (Google, Yahoo! and Bing) When using Ambiguous Search Queries

Wynand Nel<sup>1</sup><sup>a</sup>, Lizette de Wet<sup>1</sup><sup>b</sup> and Robert Schall<sup>2</sup><sup>c</sup>

<sup>1</sup>Department of Computer Science and Informatics, University of the Free State, Nelson Mandela Drive, Bloemfontein, South Africa

<sup>2</sup>Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa

**Keywords:** Ambiguous Search Query, Bing, Google, Graeco-Latin Square Design, Randomised Controlled Trial, System Usability Scale, Usability, User Experience, World Wide Web, Yahoo!.

**Abstract:** Performing digital searches, like searching the World Wide Web (WWW), is part of everyday life with the Internet being the primary source of information. The enormous size of the WWW led to the development of search engines, and many researchers use search engines to find specific information. Users generally prefer short queries, potentially causing ambiguity so that the search engine returns a surfeit of results. In this study, the current usability state of the major search engines (*Google*, *Yahoo!* and *Bing*) when ambiguous search queries are used was investigated. Participants completed pre and post-test questionnaires, including a System Usability Scale (SUS) questionnaire during a usability test. Each participant also performed different searches on three different occasions using three ambiguous search terms (*shoot*, *divide* and *seal*) in a randomised order according to a Graeco-Latin Square design. The study results suggest that the participants perceived the usability of *Google* to be the highest, followed by *Yahoo!* and *Bing*. Tasks that involved navigating more web pages in search of an answer were more difficult, and the order in which the tasks was completed did not have an impact on the results.

## 1 INTRODUCTION

In today's information age, people search daily for information. These searches/information retrieval (IR) can be non-electronic (using printed material), or electronic (digital searches). IR can be defined as the process of searching, retrieving and understanding of information stored in a computer system, catalogue or file (Dictionary.com, 2018; Merriam-Webster, 2018; TheFreeDictionary by Farlex, 2015).


For many people, the Internet is the primary source of information. Thus, digital searches using the World Wide Web (WWW) are part of daily life. In the current study, the usability of major search engines (*Google*, *Yahoo!* and *Bing*), when ambiguous search queries (*shoot*, *divide* and *seal*) are used, was evaluated.


## 2 BACKGROUND


The Internet links millions of computers worldwide through telephone cables, local and wide area networks, undersea cables and satellites. Through this interconnectivity, computers communicate and exchange information (Mouton, 2001).

"How big is the WWW?" is a question asked by many; however, there is no single answer. The size of the WWW can be measured in various ways, like counting the number of registered domains, counting the number of websites being used or trying to estimate the number of individual web pages (Fowler, 2018).

The first Google size index in 1999, only counting unique Uniform Resource Locators (URLs), consisted of 26 million pages but has reached the 1 trillion mark already in 2008 (Alpert & Hajaj, 2008). The index

<sup>a</sup> <https://orcid.org/0000-0001-5579-6411>

<sup>b</sup> <https://orcid.org/0000-0001-6819-6984>

<sup>c</sup> <https://orcid.org/0000-0002-4145-3685>

grew to 30 trillion pages in 2013 (Koetsier, 2013) and further increased to 130 trillion in 2016 (Schwartz, 2016a,b). Currently, the estimated minimum size of the indexed WWW is 4.42 billion pages, measured across the number of non-overlapping unique pages as indexed by *Google*, *Yahoo!* and *Bing* (de Kunder, 2018).

Google has the highest market share (Figure 1), and by the second quarter of 2020, it had a global market share of 92.12%, followed by Bing (2.56%) and Yahoo! (1.70%) (“Search Engine Market Share Worldwide”, 2020).

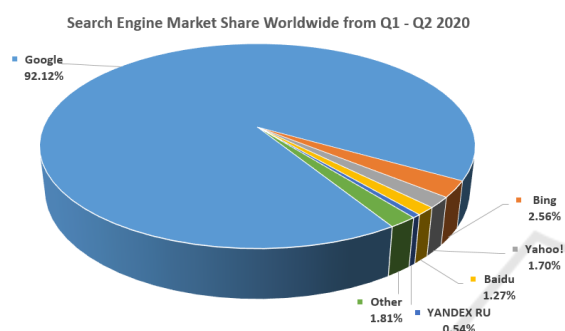


Figure 1: Search Engine Market Share Worldwide: Q1-Q2 2020 (“Search Engine Market Share Worldwide”, 2020).

Furthermore, Google was also ranked the top search engine in South Africa (Figure 2) with a score of 96.17%, followed by Bing and Yahoo! scoring 2.89% and 0.66%, respectively (“Search Engine Market Share South Africa”, 2020).

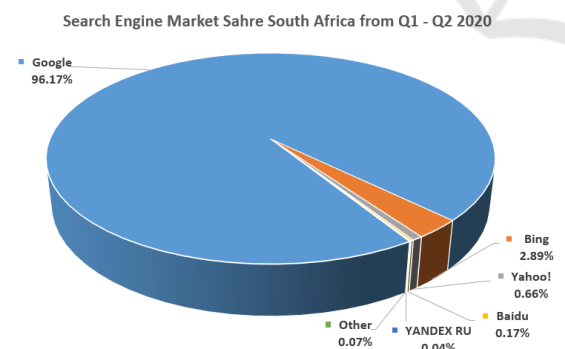


Figure 2: Search Engine Market Share South Africa: Q1-Q2 2020 (“Search Engine Market Share South Africa”, 2020).

Edosomwan and Edosomwan (2010) state that the ultimate goal of a website is to share information, but the sheer size of the WWW poses a challenge for IR. A user, searching for specific information on the WWW, needs to know the relevant address (URL) of that information on the WWW. The problem originates

from the fact that the user usually does not know this web address beforehand. Thus, using search engines has become the only practical way to search the WWW (Goodman & Cramer, 2010).

Search engines are among the most accessed web sites (Edosomwan & Edosomwan, 2010; Oberoi & Chopra, 2010); according to a study by Breytenbach and McDonald (2010), 84% of Internet users use search engines to find information from various spheres of life such as travel, literature, food, science and music. The user supplies the search engine with either a single or multiple words (often referred to as *search terms*, *keywords*, *search queries* or *search strings*), and the search engine then returns a list of documents or web pages which match the word(s).

The words *search terms*, *keywords*, *search queries* and *search strings* are often used synonymously, but for search professionals, it is important to distinguish between them. The Oxford English Dictionary (2010) defines a *query* as “to put a question or questions to someone”. A *search term*, on the other hand, is defined as a *specific word, phrase or term that is used electronically to retrieve information from files or databases* (Gabbert, 2017; “Search term”, n.d.; “Search term”, n.d.).

Gabbert (2017) and Patel (2015) explain that the term *keywords* is the exact term that a person or company target in a paid search or organic search campaign. The *Cambridge Dictionary* (“Meaning of ‘search query’ in the English Dictionary”, 2018) defines a *search query* as the “words that are typed into a search engine in order to get information from the Internet” and is usually what users enter into a search engine (Patel, 2015). The crucial difference between *search queries* (what users type) and *keywords* (what companies are targeting) is depicted in Figure 3. Note that the spelling errors in Figure 3 were made deliberately. Different search queries, entered by users, can all point to the same keyword.

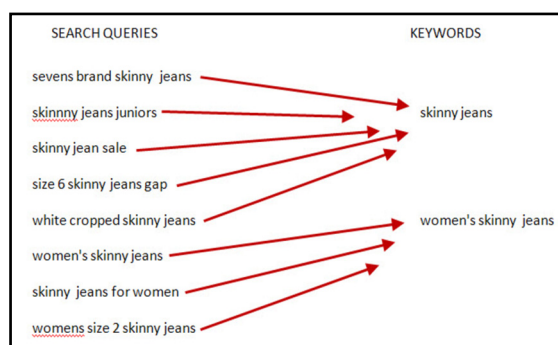


Figure 3: Difference between Search Queries and Keywords (Gabbert, 2017).

A *search string* is defined similarly to a search query, namely as the combination of characters, numbers and symbols (words) entered into a search engine that makes up the search being conducted (Rouse, 2016; “Search string”, 2017). This paper will use the notions of *search term* (a specific word or phrase entered into a search engine) and *search query* (all the words entered into a search engine when a search is performed).

Often search engines return thousands of results. The user then must inspect these results until the required information is found. A potential cause of the excessive amount of search results is the type of search query that might have been used. According to Teevan, Dumais and Horvitz (2007), users favour short queries, possibly ambiguous, so that the search engine returns more results than needed. The reason for the excess results is the search engine’s inability to determine exactly what the user is searching for. For example, if the user enters “bat” as a search query, this could refer to an interest in animals, but the search engine might return results on the device used for playing sports. Large numbers of irrelevant results can cause user frustration as many irrelevant results need to be looked at. However, search engines are constantly developing new methods to interpret and respond to user intent (Goodman & Cramer, 2010), trying to make the search results more useful for finding what the user is looking for, faster (Linden, 2007).

### 3 SEARCH QUERIES

WWW search queries can be categorised into informational, transactional and navigational queries (Breytenbach & McDonald, 2010; Krug, 2006). With *informational queries*, the user tries to obtain information about a specific topic, whereas, with a *transactional query*, a user would shop for a product or download information from a website or interact with the result of a website. With *navigational queries*, the goal is to navigate a user to a specific URL, like the homepage of an organisation.

Search queries can also be described as *ambiguous* (it has more than one meaning, for example, *bat*); *broad* (it has many sub-categories, for example, *religion*); *proper nouns* (it can be locations or names, for example, *Abrahams*); or *clear* (it is specific with a narrow topic, for example, *University of Oslo*) (Azzopardi, 2007; Dou, Song & Wen, 2007; Elbassuoni, Luxenburger & Weikum, 2007; Sanderson, 2008; Song, Luo, Wen, Yu & Hon, 2007). The focus of this paper is on ambiguous search queries.

#### 3.1 Ambiguous Search Queries

Ying, Scheuermann, Li, and Zhu (2007) argue that ambiguity is a severe problem in keyword-based search methods. Sanderson (2008) and Song, Luo, Nie, Yu, and Hon (2009) report that approximately 7% – 23% of search queries are ambiguous, with the median length of queries being one word. Search engines struggle to provide pertinent results from short queries which might not offer sufficient information. This causes the search engine to provide a diverse set of results (Luo, Liu, Zhang & Ma, 2014).

An example of an ambiguous search term is the word *java*. What is a person searching for when the search term *java* is entered into a search engine? Should the search engine return information on Java coffee or instead on the Java computer programming language? If the search is in fact for the Java programming language, precisely what is being searched for? Is the person looking for Java documentation explaining the syntax or rather a brief explanation of what Java is?

Search engines use various methods to mitigate the effect of ambiguous queries. As this paper’s focus is not on mitigating ambiguous queries, these methods will only be mentioned briefly:

- Word sense disambiguation is the process of identifying the sense of a word in a query to assist the search engine in returning relevant results (Gale, Church & Yarowsky, 1992; Voorhees, 1993; Ying *et al.*, 2007).
- Personalisation, specifically of web pages, is defined by “E-Business Solutions” (2011) and Linden (2007) as modified web pages that use a dynamically customised content delivery system in respect of the needs of the viewer. In other words, different search results are displayed to different users based on a user’s unique identifier, such as a user ID, membership number or subscription service (“E-business solutions”, 2011).
- Query expansion is a method employed by some search engines where suggestions are made to the user by adding new query terms to the users’ query. This can happen automatically or interactively (Carpineto, de Mori, Romano & Bigi, 2001; Mitra, Singhal & Buckley, 1998). Google, Yahoo! and Bing assist the user further by suggesting corrections to misspelt search queries and by noting what the user does in response (Loukides, 2010).
- Click-through data are captured by the search engine when a user clicks on a specific result in the result set. This can be used to personalise

web searches for a user (Leung, Ng & Lee, 2008).

- Clustering is the term used when related queries (documents) are organised into groups based on their overall similarity to one another. These clusters can be used by search engines to improve themselves in several aspects (Baeza-Yates, Hurtado & Mendoza, 2007).

## 4 USABILITY

There are many different definitions of usability, and many researchers/professionals in the field might have their own (Tullis & Albert, 2013). This section attempts to clarify this concept.

A broad definition of usability, as given by Tullis and Albert (2013), is the user's ability to use something to complete a task successfully. Rogers, Sharp and Preece (2011) agree with this definition and further describe it as an attribute of an interactive product that is easy to learn and effective and enjoyable to use.

The International Standards Organisation (International Standards Organization - ISO 9241-11, 1998) emphasises that people will be reluctant to use a software program if the usability of the program is poor. Furthermore, their definition identifies three goals of usability, namely *effectiveness* (the ability of a user to achieve specified objectives by using a system), *efficiency* (the resources which are exhausted to complete a task), and *satisfaction* (a user's attitude, positive or negative, towards a system and the lack of any discomfort).

Nielsen (2012) and Preece, Rogers and Sharp (2015) confirm these goals and add *safety* (protecting the user from undesirable situations), *utility* (the extent to which the product provides the right kind of functionality to achieve certain goals), *learnability* (how easy the system is to use), *memorability* (how easy it is to remember how to use the program, once learned), and *errors* (how many errors do users make, how severe are these errors, and how easily can users recover from the errors), as additional goals. Usability is measured to the extent to which a user can use a product to achieve these goals in a specified context of use (Rogers *et al.*, 2011).

Krug's (2006 p. 5) definition of usability is the following, "... making sure that something works well: that a person of average (or even below average) ability and experience can use the thing—whether it's a Web site, a fighter jet, or a revolving door—for its intended purpose without getting hopelessly frustrated" and this is the definition that is used for the purpose of this paper.

## 5 METHODOLOGY

The focus of this research paper is to determine the usability of *Google*, *Yahoo!* and *Bing*, when ambiguous search queries (*shoot*, *divide* and *seal*) are presented as input. This was the first step towards a larger research study where the objective was to determine the value of Brain-Computer Interface (BCI) measurements (user emotions) in computer usability testing, specifically measuring user experience (UX) when performing ambiguous search queries.

The following research question was asked: "What is the current usability state of the major search engines (*Google*, *Yahoo!* and *Bing*) when ambiguous search queries are used?". To answer the research question, 36 participants were recruited. All participants were first-year students enrolled in the computer literacy course at the University of the Free State main campus. The participants were almost equally split between male (19) and female (17), with the majority less than 21 years old (24). Eleven students were between 21 and 25 years of age, with one falling in the range of 26 to 30.

### 5.1 Pre-test Questionnaire

Participants completed a pre-test questionnaire eliciting their personal details, computer, Internet, and search engine experience. The purpose of these questions was to determine the participants' biographic information, as well as their self-rated experience for each of the categories.

The participants were grouped as follows: *Novice/First-Time Users*, *Knowledgeable Intermittent Users* and *Expert Frequent Users* (Shneiderman, 1998). Each participant had to answer three sets of similar questions for each category (Computer, World Wide Web/Internet and Search Engine Experience). The first question, "For how many years have you been using a computer?" presented the participant with the following options of which he/she had to pick one: *Never used a computer*; *Less than 1 year*; *1-3 years*; *More than 3 years but less than 5 years*; and *More than 5 years*. A follow-up question asked the participant, "How often do you use a computer?" The participant had to select one of the following options: *Daily*; *Weekly*; *Once every two weeks*; *Once a month*; and *Less than once a month*. Each option was awarded a numerical value: from 0 to 4 for the first question and 4 to 0 for the second question, which was then multiplied in order to calculate a final computer experience value. A participant with a score of 0 to 4 was classified as a *Novice/First Time User*.

*Knowledge Intermittent Users* had to score between 5 and 9, whereas an *Expert Frequent User's* score had to be between 10 and 16.

## 5.2 Randomised Controlled Trial

We conducted a randomised controlled trial of the usability of major search engines when using ambiguous search queries. Each participant completed three tasks, namely searches for information, using three pre-selected ambiguous search terms (*shoot*, *divide* and *seal*) when entered into three search engines (*Google*, *Yahoo!* and *Bing*). The participants carried out their tasks on three different occasions (*first*, *second* and *third*), using unique combinations of the three search engines and of the three search terms in a randomised order according to a Graeco-Latin Square design (Kempthorne, 1983 p. 187). The Graeco-Latin Square is a balanced design which allows one to investigate simultaneously the effect of three factors (here search term, search engine and occasion) in a cross-over study. As such it is eminently suited for usability experiments where multiple factors are of interest and within-subject comparisons are much more powerful than between-subject comparisons. The study adopted the binary measure of task success (Tullis & Albert, 2013), as a participant either succeeded or failed at the task.

The search terms selected for the usability test needed to show the same characteristics when submitted to the three search engines. Since the user was not allowed to change the search string via the keyboard, it was crucial that the “*Searches related to [original search string]*” in Google (“Google Search Engine”, 2018), “*Also Try*” in Yahoo! (“Yahoo! Search Engine”, 2018), and “*Related searches*” in Bing (“Bing search engine”, 2018) yielded the same results for the three search engines. These suggestions are usually provided by search engines to assist users in finding more relevant search queries. Google and Yahoo! display these suggestions at the bottom of the web page while Bing displays them both at the bottom and at the top right of the web page.

The trial was conducted in the usability laboratory of the Department of Computer Science and Informatics. The laboratory made it possible to control the lighting conditions, ambient temperature, interruptions, noise, and seating position of the participants.

Each of the ambiguous search terms (*shoot*, *divide* and *seal*) was linked to a specific question. The user completed the task successfully once he/she answered the question correctly. The questions were as follows:

- How many shooting games are listed in the section: “Popular Shooting Games” on the Armor Games website (*hint: Shooting Games/Armor Games*)?
- How do you insert a division symbol ( $\div$ ) in Microsoft Word? List the steps. (*hint: eHow*)
- Find the image of seal and his ex-wife, Heidi Klum, holding hands with two boys and a girl. [The image which they had to find, was provided below the question.]

Three steps were needed to complete the tasks with the search terms *shoot* and *seal*, while four were needed with the search term *divide*.

## 5.3 Post-test Questionnaire

Each of the 36 participants completed a post-test questionnaire, including a System Usability Scale (SUS) questionnaire for each search. The SUS questionnaire consisted of 10 statements, half of which were worded positively and the other half negatively. Since the introduction of the SUS questionnaire by Brooke (1996), it has been assumed that the ten statements of the SUS questionnaire were unidimensional, intended to measure only perceived ease-of-use. It was, however, proved that the SUS questionnaire has two factors measuring global system satisfaction as well as sub-scales of usability (8 statements) and learnability (2 statements, statements 4 and 10, respectively) (Lewis & Sauro, 2009; Sauro, 2011).

A five-point Likert scale was used to record the participants' level of agreement with each statement. The SUS score was then calculated as specified by Brooke (1996) and averaged for each participant to calculate the overall usability score of the search engine in question (Tullis & Albert, 2013). The final score indicated the participant's opinion of the usability of a specific search engine. Thus three SUS scores were available for each participant, recorded after the three occasions when a WWW search was done using the particular search engine/search term combinations allocated to each participant.

## 6 STATISTICAL ANALYSIS

The objective of the study was to characterise and compare the usability of the three search engines (*Google*, *Yahoo!* and *Bing*) when short, ambiguous search terms were used. Thus it was of primary interest to determine whether there were significant differences in mean SUS scores between the three Search Engines (*Google*, *Yahoo!* and *Bing*); additionally, the effects of

Search Term (*shoot, divide and seal*) and of Occasion (*first, second and third*) on mean SUS scores were investigated. Task success and the number of web pages needed to complete the task were also studied, but fall outside the scope of this paper.

To assess statistically the effect of *Search Engine*, *Search Term* and *Occasion*, respectively, on mean SUS scores, the following null hypotheses were formulated:

- $H_{0,1}$ : There are no differences between the mean SUS scores of the three Search Engines (*Google, Yahoo!* and *Bing*).
- $H_{0,2}$ : There are no differences between the mean SUS scores of the three Search Terms (*shoot, divide and seal*).
- $H_{0,3}$ : There are no differences between the mean SUS scores of the three Occasions (*first, second and third*).

The SUS scores were analysed using Analysis of Variance (ANOVA) fitting the factors *Participant*, *Occasion*, *Search Term* and *Search Engine*. From this ANOVA, F-statistics and P-values associated with the overall tests of the significance of the factors *Occasion*, *Search Term* and *Search Engine* are reported. The mean SUS scores for each level of the factors *Occasion*, *Search Term* and *Search Engine* are also reported.

Furthermore, the three *Search Engines* were compared by calculating point estimates for the three pairwise differences in mean SUS score between *Search Engines*, as well as 95% confidence intervals for the mean differences and the associated P-values. The three *Search Terms* and three *Occasions* were compared similarly.

The statistical analysis was carried out using the MIXED procedure of the SAS/STAT 13.1 software (SAS Institute Inc, 2013).

## 6.1 Overall Tests

The overall F-tests for the effect of *Search Engine*, *Search Term* and *Occasion* on the SUS score are reported in Table 1, together with the mean SUS score for each level of the factors *Occasion*, *Search Term* and *Search Engine*. To improve readability, the significant P-values ( $P < 0.05$ ) are underlined.

The results depicted in Table 1 show that the effect of *Search Engine* is statistically significant ( $P < 0.0001$ ). The estimate (mean) SUS scores for *Google, Yahoo!* and *Bing* are 81.11, 65.42 and 55.56, respectively. The null hypothesis,  $H_{0,1}$ , can thus be rejected. There is a statistically significant difference between the mean SUS scores of the three search engines (*Google, Yahoo!* and *Bing*)

Table 1: Effect of Search Engine, Search Term and Occasion on SUS Score: Mean Values and Overall F-tests.

Effect	Effect Level	Mean	F-statistic <sup>1</sup>	P-value <sup>1</sup>
Search Engine	Bing	55.56	32.42	<u>&lt;0.0001</u>
	Google	81.11		
	Yahoo!	65.42		
Search Term	shoot	70.00	6.17	<u>0.0035</u>
	divide	60.90		
	seal	71.18		
Occasion	first	65.14	1.06	0.3514
	second	67.15		
	third	69.79		

<sup>1</sup>F-test for null-hypothesis of no effect of *Occasion*, from ANOVA with *Participant*, *Search Engine*, *Search Term* and *Occasion* as fixed effects; F-statistic has 2 and 66 degrees of freedom.

Similarly, Table 1 shows that the effect of *Search Term* is significant ( $P < 0.0035$ ). The estimate (mean) SUS scores for *shoot, divide and seal* are 70.00, 60.90 and 71.18, respectively. The null hypothesis,  $H_{0,2}$ , can thus also be rejected. There is a statistically significant difference between the mean SUS scores of the three search terms (*shoot, divide and seal*).

Finally, Table 1 shows that the effect of *Occasion* is not significant ( $p < 0.3514$ ). The estimate (mean) SUS scores for the occasions *first, second and third* are 65.14, 67.15 and 69.79, respectively. The null hypothesis,  $H_{0,3}$ , can thus not be rejected. There is no statistically significant difference between the mean SUS scores of the three occasions (*first, second and third*).

In summary, the statistical analysis showed that there are statistically significant differences in the usability of the three *Search Engines*, as measured by the SUS questionnaire; similarly, *Search Term* had a significant effect on mean SUS score, but *Occasion* did not.

## 6.2 Pairwise Comparisons

Pairwise comparisons were done in order to determine any significant differences between the effects of pairs of *Search Engines*, *Search Terms* and *Occasions*. The following secondary hypotheses were formulated:

- $H_{0,1a}$ : There is no significant difference between the mean SUS score of *Google* and *Yahoo!*.
- $H_{0,1b}$ : There is no significant difference between the mean SUS score of *Google* and *Bing*.

- $H_{0,1c}$ : There is no significant difference between the mean SUS score of *Yahoo!* and *Bing*.
- $H_{0,2a}$ : There is no significant difference between the mean SUS score of *shoot* and *divide*.
- $H_{0,2b}$ : There is no significant difference between the mean SUS score of *shoot* and *seal*.
- $H_{0,2c}$ : There is no significant difference between the mean SUS score of *divide* and *seal*.

The results of the pairwise comparisons are summarised in Table 2. Again, significant P-values ( $P < 0.05$ ) are underlined. The results show that all three pairwise differences in mean SUS score between search engines are statistically significant. All three hypothesis  $H_{0,1a}$ ,  $H_{0,1b}$  and  $H_{0,1c}$  can thus be rejected. There is a statistically significant difference between the mean SUS scores of the three search engines.

Similarly, Table 2 shows that two of the three pairwise differences in mean SUS score between search terms, namely *shoot* versus *divide* and between *divide* versus *seal*, are statistically significant. The hypotheses  $H_{0,2a}$  and  $H_{0,2c}$  can thus be rejected. There is a statistically significant difference between the mean SUS scores of *shoot* and *divide* and between *divide* and *seal*.

Finally, none of the pairwise differences between the different occasions (*first*, *second* and *third*) is statistically significant, indicating that the order in which the participants used the three search engines and three search terms did not affect the SUS score.

## 7 DISCUSSION

The answer to the research question, “What is the current usability state of the major search engines (Google, Yahoo! and Bing) when ambiguous search queries are used?” can be summarised as follows.

There were statistically significant differences between the three search engines (Google vs Yahoo!, Google vs Bing and Yahoo! vs Bing) with respect to mean SUS score. Specifically, with regard to mean SUS score, Google was ranked the highest, with a mean SUS score of 81.11, followed by Yahoo! and then Bing with mean SUS scores of 65.42 and 55.56, respectively (Table 1). In other words, the participants experienced Google to be the easiest to use, followed by Yahoo! and Bing.

Regarding search terms, there were statistically significant differences between shoot versus divide and divide versus seal. These differences could have been caused by the fact that the search term divide was the only search term where the minimum number of web pages to view in order to find the answer was four, whereas the minimum number of web pages to view for search terms shoot and seal was three. In other words, using the search term divide, the participants experienced the task as being more difficult than when using the terms shoot or seal.

It was interesting to note that there were no significant differences in mean SUS score between the occasions.

Thus, the order in which the participants used the three search engines and three search terms did not affect on the SUS score. This suggests the absence of learning or tiring effects on the SUS scores.

## 8 CONCLUSION

The data reported here were obtained through a traditional data collection method, namely the SUS questionnaire. The data represented the participants’ perceived usability of each *Search Engine* (Google, Yahoo! and Bing), *Search Term* (*shoot*, *divide* and *seal*) and *Occasion* (*first*, *second* and *third*) while using ambiguous search terms.

Table 2: Comparison of Search Engine, Search Term and Occasion with Respect to SUS Score.

Effect	Contrast	Difference of means	95% CI for difference <sup>1</sup>	P-value <sup>1</sup>
Search Engine	Google vs Bing	25.56	19.16 to 31.95	<u>&lt; 0.0001</u>
	Yahoo! vs Bing	9.86	3.47 to 16.25	<u>0.0030</u>
	Google vs Yahoo!	15.69	9.30 to 22.09	<u>&lt; 0.0001</u>
Search Term	shoot vs divide	9.10	2.71 to 15.49	<u>0.0060</u>
	shoot vs seal	-1.18	-7.57 to 5.21	0.7135
	divide vs seal	-10.28	-16.67 to -3.89	<u>0.0020</u>
Occasion	first vs second	-2.01	-8.41 to 4.38	0.5315
	first vs third	-4.65	-11.04 to 1.74	0.1509
	second vs third	-2.64	-9.03 to 3.75	0.4127

<sup>1</sup>ANOVA with *Participant*, *Search Engine*, *Search Term* and *Occasion* as fixed effects.

The statistical analysis showed that with respect to mean SUS score, the participants found *Google* to be the most usable, followed by *Yahoo!* and *Bing*. The results corresponded with the participants' self-reported confidence level where the majority reported that they were confident using *Google* and less so *Yahoo!*, followed by *Bing*. The pre-test questionnaire also revealed that *Google* was the preferred searchengine for all participants. This fact was further confirmed when 13 and 15 participants indicated that they had never before used *Yahoo!* or *Bing*, respectively. Furthermore, the pre-test questionnaire revealed that approximately half of the participants' browser homepages were set to *Google*. Fourteen participants stated that they were used to how *Google* worked and that it returned results almost instantly.

The results from the post-test questionnaire showed that most participants preferred *Google*, followed by *Yahoo!* and *Bing*. These findings contribute to existing literature indicating that good usability could be one of the contributors to *Google* having the highest market share, followed by *Yahoo!* and *Bing*, worldwide as well as in South Africa.

With regard to the search terms, the statistical analysis showed that the task associated with the search term *divide* was more difficult than the tasks involving the terms *shoot* and *seal*. This finding could be attributed to the fact that *divide* was the only search term that required the participants to navigate to four web pages in order to find the answer, compared to the search terms *shoot* and *seal*, which only required three. Therefore, the fewer web pages a participant has to navigate, the more positive they find the experience.

Interestingly there were no statistically significant differences in mean SUS scores between the occasions, which suggests that learning and tiring effects did not significantly affect perceived usability. The post-test questionnaire confirmed this finding, as the majority of the participants indicated that their energy levels were unchanged after completing the testing session versus before starting the testing session. Some even indicated that they felt more energetic after completing the testing session.

In the light of the above, the participants perceived the usability of *Google* to be the highest, followed by *Yahoo!* and then *Bing*. Furthermore, tasks that involved navigating more web pages (*Search Term*) were more difficult. The order in which the tasks were completed (*Occasion*) did not affect the results.

## 9 CONTRIBUTION

With the ever-increasing size of the WWW and users' tendency to use short, ambiguous search queries resulting in incorrect or unexpected results being returned, users may frequently experience feelings of frustration or negativity – all indications of problems with usability.

This research paper succeeded in determining the usability of *Google*, *Yahoo!* and *Bing* when ambiguous search terms were presented as the input. It makes both a practical and methodological contribution to the body of knowledge.

In terms of the practical contribution, the results indicated that the usability aspects tested by the SUS questionnaire, namely measuring global system satisfaction as well as sub-scales of usability and learnability (Lewis & Sauro, 2009; Sauro, 2011), have an effect on the usability of a search engine in terms of users entering ambiguous search queries. The results further indicated the importance of keeping the number of webpages that the user has to navigate, as small as possible since the number of page visits seems to affect the users' perception of the usability of a search engine's handling of ambiguous search queries.

Regarding the methodological contribution of the paper, the use of the Graeco-Latin Square design (Kempthorne, 1983 p. 187) proved to be beneficial. This design allowed the researchers to simultaneously investigate and adjust for the orthogonal factors *Search Engine*, *Search Term* and *Occasion* while removing the typically very large between-subject variability from the statistical comparisons. Thus, in the present application, the Graeco-Latin Square (Kempthorne, 1983 p. 187) was an efficient design for the estimation of contrasts between search engines, search terms, and occasions with high precision (high power). Therefore, its use in a future study of the same or similar type is recommended.

## 10 FUTURE RESEARCH

The research presented in this paper paved the way for a follow-up paper focussing on the effect of *Search Engine*, *Search Term* and *Occasion* on Brain-Computer Interface metrics for emotions, when ambiguous search queries are used (Nel, De Wet & Schall, 2019).

This paper presented the differences in the mean SUS scores to determine the usability of Google, Yahoo! and Bing when ambiguous search queries were used; follow-up papers will report on the:



- The effect of Search Engine, Search Term and Occasion on the participants' probability of task success when ambiguous search queries are used.
- The effect of Search Engine, Search Term and Occasion on the number of web pages visited to complete a task when ambiguous search queries are used.

## REFERENCES

- Alpert, J. & Hajaj, N. 2008. *We knew the web was big...* [Online], Available: <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html> [2018, November 26].
- Azzopardi, L. 2007. Position paper: Towards evaluating the user experience of interactive information access systems. *Proceedings of SIGIR 2007 Workshop: Web Information Seeking and Interaction*. 60–64.
- Baeza-Yates, R., Hurtado, C. & Mendoza, M. 2007. Improving search engines by query clustering. *Journal of the American Society for Information Science and Technology*. 58(12).
- Bing search engine*. 2018. [Online], Available: [www.bing.com](http://www.bing.com) [2018, November 26].
- Breytenbach, J. & McDonald, T. 2010. Soeknjindekking van Suid-Afrikaanse en Afrikaanse webruimtes. *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie*. 29(3).
- Brooke, J. 1996. *SUS: A Quick and Dirty Usability Scale. Usability Evaluation in Industry*. P. Jordan, B. Thomas, B. Weerdmeester, & I. McClelland (eds.). London: Taylor & Francis.
- Carpineto, C., de Mori, R., Romano, G. & Bigi, B. 2001. An information-theoretic approach to automatic query expansion. In Vol. 19. *ACM Transactions on Information Systems ACM Transactions on Information Systems (TOIS)*. 1–27. [Online], Available: [http://www.accessmylibrary.com/coms2/summary\\_0286-10728134\\_ITM](http://www.accessmylibrary.com/coms2/summary_0286-10728134_ITM).
- Dictionary.com. 2018. *Information retrieval*. [Online], Available: <http://dictionary.reference.com/browse/information+retrieval> [2018, November 21].
- Dou, Z., Song, R. & Wen, J. 2007. A large-scale evaluation and analysis of personalized search strategies. *Proceedings of the 16th international conference on World Wide Web - WWW '07*. 581.
- E-business solutions*. 2011. [Online], Available: [http://www.it-architects.co.uk/a\\_-\\_z\\_glossary\\_index/E-Business\\_Solutions\\_Glossary/e-business\\_solutions\\_glossary.html](http://www.it-architects.co.uk/a_-_z_glossary_index/E-Business_Solutions_Glossary/e-business_solutions_glossary.html) [2011, March 08].
- Edosomwan, O.J. & Edosomwan, T. 2010. Comparative analysis of some search engines. *South African Journal of Science*. 106(11/12). [Online], Available: <http://www.sajs.co.za/index.php/SAJS/article/view/169/439>.
- Elbassuoni, S., Luxenburger, J. & Weikum, G. 2007. Adaptive personalization of web search. *Proceedings of SIGIR 2007 Workshop: Web Information Seeking and Interaction*. 1–4.
- Fowler, D.S. 2018. *How many websites are there in the world?* [Online], Available: <https://tekeye.uk/computing/how-many-websites-are-there> [2018, October 22].
- Gabbert, E. 2017. *Keywords vs. search queries: What's the difference?* [Online], Available: <http://www.wordstream.com/blog/ws/2011/05/25/keywords-vs-search-queries> [2018, November 26].
- Gale, W.A., Church, K.W. & Yarowsky, D. 1992. Using bilingual materials to develop word sense disambiguation methods. In Montréal, Canada *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. 101–112.
- Goodman, E. & Cramer, M. 2010. [Online], Available: [http://www.comscore.com/Press\\_Events/Presentations\\_Whitepapers/2010/The\\_Future\\_of\\_Search\\_The\\_Emerging\\_Power\\_of\\_Real-time\\_Personalized\\_Search](http://www.comscore.com/Press_Events/Presentations_Whitepapers/2010/The_Future_of_Search_The_Emerging_Power_of_Real-time_Personalized_Search).
- Google Search Engine*. 2018. [Online], Available: [www.google.co.za](http://www.google.co.za) [2018, November 26].
- International Standards Organization - ISO 9241-11. 1998. *Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability*.
- Kemphorne, O. 1983. *Design and Analysis of Experiments*. Revised ed. Florida: Robert E Krieger Publishing Company.
- Koetsier, J. 2013. *How Google searches 30 trillion web pages, 100 billion times a month*. [Online], Available: <https://venturebeat.com/2013/03/01/how-google-searches-30-trillion-web-pages-100-billion-times-a-month/> [2017, October 11].
- Krug, S. 2006. *Don't Make Me Think! A Common Sense Approach to Web Usability*. 2nd ed. Vol. 277. Berkeley, California USA: New Riders Publishing.
- de Kunder, M. 2018. *The size of the World Wide Web (The Internet)*. [Online], Available: <http://www.worldwidewebsize.com/> [2018, October 22].
- Leung, K.W., Ng, W. & Lee, D.L. 2008. Personalized concept-based clustering of search engine queries. *IEEE Transactions on Knowledge and Data Engineering*. 20(11):1505–1518.
- Lewis, J.R. & Sauro, J. 2009. The factor structure of the system usability scale. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 5619 LNCS:94–103.
- Linden, G. 2007. *Personalized search primer and Google's approach*. [Online], Available: [http://www.readwriteweb.com/archives/personalized\\_search\\_primer.php](http://www.readwriteweb.com/archives/personalized_search_primer.php) [2011, March 08].
- Loukides, M. 2010. *What is data science? Analysis: The future belongs to the companies and people that turn data into products*. [Online], Available:

- <http://radar.oreilly.com/2010/06/what-is-data-science.html> [2010, July 01].
- Luo, C., Liu, Y., Zhang, M. & Ma, S. 2014. Query ambiguity identification based on user behavior information. *Information Retrieval Technology*. (863):36–47.
- Meaning of “search query” in the *English Dictionary*. 2018. [Online], Available: <http://dictionary.cambridge.org/dictionary/english/search-query> [2018, November 12].
- Merriam-Webster, I. 2018. *Dictionary: Information retrieval*. [Online], Available: [http://www.merriam-webster.com/dictionary/information retrieval](http://www.merriam-webster.com/dictionary/information%20retrieval) [2018, March 17].
- Mitra, M., Singhal, A. & Buckley, C. 1998. Improving automatic query expansion. In New York: ACM New York *SIGIR '98 Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Mouton, J. 2001. *How to succeed in your Master's and Doctoral Studies. A South African guide and resource book*. Van Schaik publishers.
- Nel, W., De Wet, L. & Schall, R. 2019. The effect of search engine, search term and occasion on brain-computer interface metrics for emotions when ambiguous search queries are used. *CHIRA 2019 - Proceedings of the 3rd International Conference on Computer-Human Interaction Research and Applications*. (Chira):28–39.
- Nielsen, J. 2012. *Usability 101: Introduction to usability*. [Online], Available: <http://www.nngroup.com/articles/usability-101-introduction-to-usability/> [2018, March 12].
- Oberoi, I.S. & Chopra, M. 2010. *Web Search Engines – A Comparative Study*. [Online], Available: [http://www.idt.mdh.se/kurser/ct3340/ht09/ADMINISTRATION/IRCS09-submissions/ircse09\\_submission\\_14.pdf](http://www.idt.mdh.se/kurser/ct3340/ht09/ADMINISTRATION/IRCS09-submissions/ircse09_submission_14.pdf).
- Oxford English Dictionary*. 2010. 2nd ed. Oxford University Press.
- Patel, N. 2015. *Understanding the difference between queries and keywords and what to do about it*. SEJ Summit in Santa Monica: SearchEngine Journal. [Online], Available: <https://www.searchenginejournal.com/understanding-difference-queries-keywords/126421/>.
- Preece, J., Rogers, Y. & Sharp, H. 2015. *Interaction Design - Beyond Human-Computer Interaction*. 4th ed. West Sussex, United Kingdom: John Wiley & Sons Ltd.
- Rogers, Y., Sharp, H. & Preece, J. 2011. *Interaction Design - Beyond Human-Computer Interaction*. 3rd ed. West Sussex, United Kingdom: John Wiley & Sons Ltd.
- Rouse, M. 2016. *Search string*. [Online], Available: <http://whatis.techtarget.com/definition/search-string> [2017, October 12].
- Sanderson, M. 2008. Ambiguous queries: Test collections need more sense. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 499–506.
- SAS Institute Inc. 2013. *SAS/STAT 13.1 User's Guide*. Cary, NC: SAS Institute Inc.
- Sauro, J. 2011. *Measuring Usability With The System Usability Scale (SUS)*. [Online], Available: <https://measuringu.com/sus/> [2020, May 07].
- Schwartz, B. 2016a. *Google's search knows about over 130 trillion pages*. [Online], Available: <https://searchengine-land.com/googles-search-indexes-hits-130-trillion-pages-documents-263378> [2017, October 11].
- Schwartz, B. 2016b. *Google knows of 130 trillion pages on the web - 100 trillion more in 4 years*. [Online], Available: <https://www.seroundtable.com/google-130-trillion-pages-22985.html> [2017, October 11].
- Search Engine Market Share South Africa*. 2020. [Online], Available: <https://gs.statcounter.com/search-engine-market-share/all/south-africa/#quarterly-202001-202002-bar> [2020, May 07].
- Search Engine Market Share Worldwide*. 2020. [Online], Available: <https://gs.statcounter.com/search-engine-market-share/#quarterly-202001-202002-bar> [2020, May 07].
- Search string*. 2017. [Online], Available: <https://www.computerhope.com/jargon/s/searstri.htm> [2017, October 12].
- Search term*. n.d. [Online], Available: <http://www.dictionary.com/browse/search-term> [2017a, October 12].
- Search term*. n.d. [Online], Available: <http://www.businessdictionary.com/definition/search-term.html> [2017b, October 17].
- Shneiderman, B. 1998. *Designing the User Interface. Strategies for Effective Human- Computer Interaction*. 3rd ed. USA: Addison Wesley Longman, Inc.
- Song, R., Luo, Z., Wen, J., Yu, Y. & Hon, H. 2007. Identifying ambiguous queries in web search. *Proceedings of the 16th international conference on World Wide Web - WWW '07*. 1169.
- Song, R., Luo, Z., Nie, J.Y., Yu, Y. & Hon, H.W. 2009. Identification of ambiguous queries in web search. *Information Processing and Management*. 45(2):216–229.
- Teevan, J., Dumais, S.T. & Horvitz, E. 2007. Characterizing the value of personalizing search. *SIGIR '07*.
- TheFreeDictionary by Farlex. 2015. *Information retrieval*. [Online], Available: <http://encyclopedia2.thefreedictionary.com/Information+Retrieval> [2015, March 17].
- Tullis, T. & Albert, B. 2013. *Measuring the User Experience - Collecting, Analyzing, and Presenting Usability Metrics*. 2nd ed. Amsterdam: Elsevier/Morgan Kaufmann.
- Voorhees, E. 1993. Using WordNet to disambiguate word senses for text retrieval. In Pittsburgh, PA, USA *ACM-SIGIR '93*. 171–180.
- Yahoo! Search Engine*. 2018. [Online], Available: <https://za.yahoo.com/> [2018, November 26].
- Ying, L., Scheuermann, P., Li, X. & Zhu, X. 2007. Using WordNet to disambiguate word senses for text classification. In Vol. Part III. Berlin Heidelberg: Springer-Verlag *ICCS*. 781–789.