User Preference and Search Engine Latency

Jake D. Brutlag, Hilary Hutchinson, Maria Stone
Google, Inc

Abstract

Industry research advocates a 4 second rule for web pages to load [7]. Usability engineers note that a response time over 1 second may interrupt a user's flow of thought [6, 9]. There is a general belief that, all other factors equal, users will abandon a slow search engine in favor of a faster alternative. This study compares two mock search engines that differ only in branding (generic color scheme) and latency (fast vs. slow). The fast latency was fixed at 250 ms, while 4 different slow latencies were evaluated: 2s, 3s, 4s, and 5s. When the slower search engine latency is 5 seconds, users state that they perceive the fast engine as faster. When the slower search engine latency is 4 or 5 seconds, users choose to use the fast engine more often. Based on pooling data for 2 s and 3 s, once slow latency exceeds 3 seconds, users are 1.5 times more likely to choose the fast engine.

KEY WORDS: web search, latency, response time

1. Introduction

Speed of service is recognized as a desirable attribute for any web application or service. However, what counts as "too slow" varies based on the task at hand. For interactive tasks (e.g. panning a map), response times of 100 ms or less are necessary [6, 9]. However, for less interactive tasks (e.g. reading email, searching the web, or downloading a document), acceptable response times anywhere from 1-16 seconds have been reported [5, 11, 4, 7]. There is a some evidence that for less interactive tasks, standards are becoming more demanding as web users become more savvy and broadband penetration increases [7].

User perception of latency for the specific task of web search is of particular interest to the authors. Given that user perception of latency is task dependent, there is much to be learned by narrowing focus to the web search task.

Our primary research question is "how fast is fast enough" for the delivery of web search results pages. For this study, we assumed that that the end-to-end speed-of-light web search latency was 250 ms. This 250 ms includes server processing time, network transit time, and browser rendering time. Do users find web search latency in excess 250 ms as acceptable as 250 ms? We selected the 2 to 5 second range of alternative latencies to study for two reasons. First, we assumed this range would be practical for a controlled experiment with a small number of participants, with each participating in a one hour session. The larger the latency impact, the smaller the number of participants needed to measure the effect. Second, we wanted the range to include latency we believed would almost certainly provoke a response.

A priori, we were not certain of the most appropriate outcome measure. We elected to measure satisfaction, preference, and perception by asking questions. We also elected to measure choice (observed preference) by allowing the participants to choose between two search engines with different latencies after they had used both.

2. Study Description

The study, a controlled experiment conducted in Spring 2007, compared two mock search engines that differed only in subtle branding (yellow vs. blue coloring) and latency (fast vs. slow). The fast latency was fixed at 250 ms, while 4 different slow latencies were evaluated: 2s, 3s, 4s, and 5s.

Forty adults from the San Francisco Bay Area participated in the study. Some effort was made to ensure diversity across gender, age, and typical network connectivity (dial up, T-1, cable, and DSL connections). All participants were familiar with the Google search engine and were compensated.

Each participant tested only one of the slow latencies. Sixteen participants experienced a slow latency of 2s; 8 participants were assigned to each of the other 3 slow latencies. For half of participants, the blue engine was the fast engine. For the other half, yellow was the fast engine. Participants were not informed of the speed difference.

2.1 Procedure

Each participant performed 140 searches with provided search scenarios and query keywords. With provided query keywords, this is a feasible number for a one hour session. The searches were organized into 14 blocks of 10 searches each.

In the first block, all participants saw the same scenarios/queries, and were permitted to choose between the two engines. For this block only, both search engines returned results at the slow latency.

In the next 12 blocks, participants did not receive a choice of search engine. Across all of these blocks, the designated fast engine returned results at the fast latency of 250 ms while the other continued to be slow.

The searches in these 12 blocks were drawn from a pool of 30 paired searches organized into blocks of 5 (5 pairs = 10 searches). We designated the first search of each pair "query set 1" and the second search of each pair "query set 2". For each participant, the first 6 blocks (of 12) were selected at random without replacement from the 6 blocks of paired searches. The second 6 blocks (of 12) were a second, independent, random sample without replacement from the 6 blocks of paired searches. In the first 6 blocks, query set 1 was assigned to one of the two engines, and query set 2 to the other engine. In

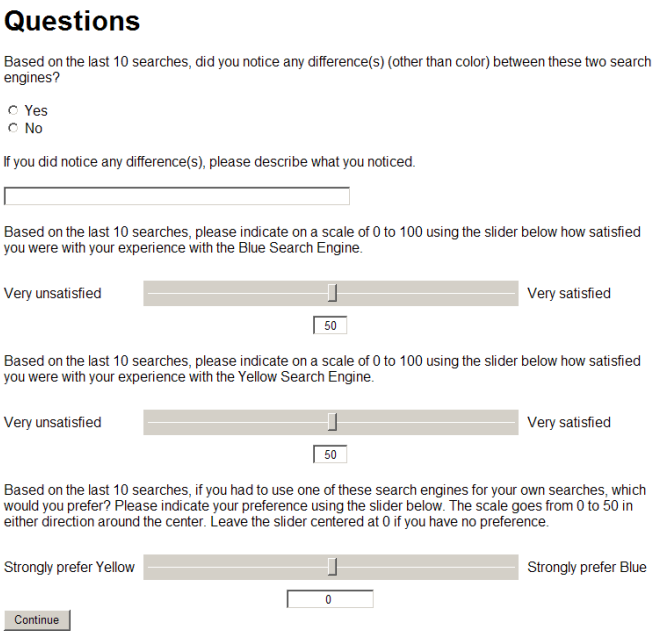

Figure 1: Questions appearing after each block of 10 searches

the second 6 blocks, assignments of query set to engine were swapped. So over the course of the 12 blocks, participants executed each search on both engines.

Within each of the 12 blocks, although the number of searches on each engine was the same, the searches were conducted in a random order, subject to the restriction that a participant conduct no more than two consecutive searches with the same engine. Note that this means that while a participant saw both searches of a pair within a block, searches of a pair were not necessarily seen consecutively.

After each of the 12 blocks, participants rated satisfaction and preference on the Questions screen illustrated in Figure 1.

For the last block of 10 searches, participants were again permitted to choose between the two engines. The designated fast search engine continued to return results at the fast latency of 250 ms. At the end of the session, participants rated satisfaction and preference overall and finally, their perception of the speed of the two search engines.

For each search, participants were given a one-sentence scenario about what they were searching for, then clicked through to the search engine home page (chosen or assigned) where a query keyword(s) related to the scenario was already filled into the search box. Participants then pressed the search button and were taken to a results page with 10 results after the appropriate latency had elapsed.

Participants had 15 seconds to select the result that best answered the search scenario. This was a cover task introduced to direct the participant's focus to what we assume is the most important aspect of the search experience: result relevance. The 15 s time interval was the same for all searches, regardless of latency; however, the timer did not begin until the results page was fully rendered in the browser. We anticipated in most cases participant would make a decision well within 15 s, and therefore complete all 140 searches in about an hour.

2.2 Hardware and Software

A custom-designed Java servlet application served as both the study software and the two mock search engines. Participants used the IE6 web browser to interact with this application, which served the scenarios, mock home pages, results pages with injected latencies, and question screens in the appropriate order. In order to control latency, the application ran on the same computer as the web browser and served pre-generated (static HTML) results pages. All participants used the same computer during the study: an IBM Thinkpad 43p with external mouse, WinXP Pro SP2.

A custom browser plug-in logged page load timing events to measure the actual latencies generated. The latencies were tuned on the study laptop ahead of time. Based on this tuning, a constant of 80 ms was subtracted from all injected latencies to account for the standard delay in loading a results page without any latency injected. Tuning indicated that the standard deviation of the actual latencies delivered was $\pm$ 20-30 ms of the target. We considered this variation acceptable.

2.3 Scenarios and Queries

The study utilized a set of English search queries popular on Google in Dec 2006. The search results for each query were the results Google provided for the query, excluding ads. We paired the queries and wrote scenarios. Using internal tools, we ensured each query of a pair had results of similar quality (relevance). Each query of a pair also had a similar scenario, as defined by number of keywords and topic. For example, here is one query/scenario pair:

basketball You want to buy tickets to an NBA basketball game

baseball You want to buy tickets to an MLB baseball game

For the 30 paired searches without a choice of engine, the queries and scenarios were made to conform to a taxonomy of search goals: information queries, navigation queries, and resource queries [10, 2]. For information queries, the user is looking for general or specific knowledge; 12 paired searches were informational. For navigation queries, the user is looking for a particular website: 6 paired searches were navigational. For resource queries, the user is looking for a specific non-informational goal, such as downloading a file or looking for entertainment: 12 paired searches used resource queries.

All of the 20 searches for which the participant was allowed a choice were information queries.

2.4 Branding

Branding must be significant enough for participants to distinguish between the search engines. On the other hand, the two brands must be neutral to ensure user response is not dictated by branding. We selected the blue and yellow color branding through consultation with a user experience designer.

Figure 2: Blue Search Engine home page

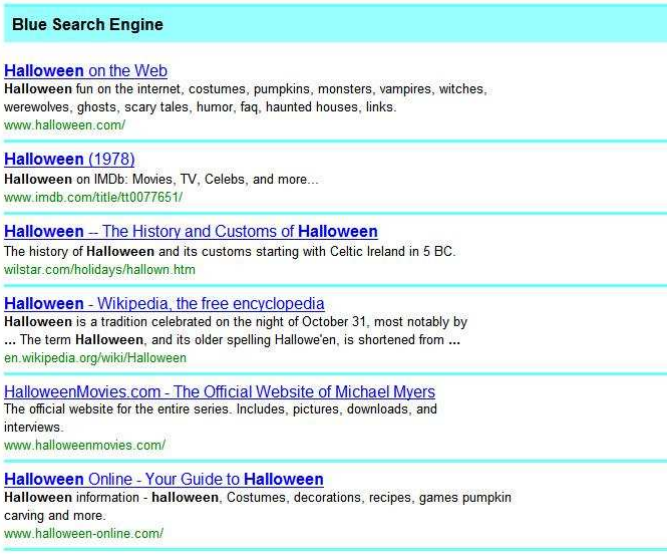

Figure 3: Blue Search Engine results page

Figure 4: Yellow Search Engine home page

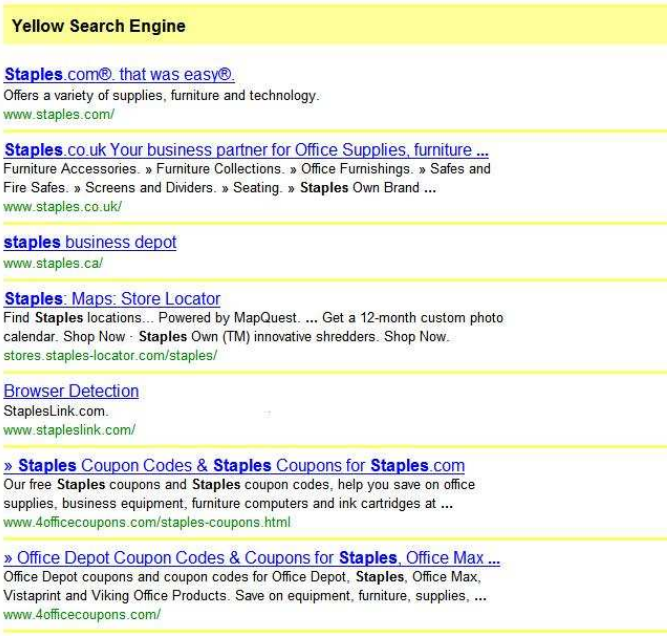

Figure 5: Yellow Search Engine home page

Figures 2 through 5 illustrate the mock home pages and search results pages for each search engine. The home page was meant to mimic the simple design of the Google home page and give participants a realistic starting point for their search. The results page always contained 10 results and no ads, and stripes of color were used to reinforce the branding when participants scrolled the page.

3. Analysis Methodology

3.1 Independent Variables

Latency. Each participant experienced one of the slow search engine latencies: 2, 3, 4, or 5 seconds.

Color For half of participants, blue was the fast search engine, for the other half, yellow was the fast search engine.

Query Set Although all participants completed each set with both search engines, half of the participants first completed query set 1 on the fast engine and half of the participants first completed query set 2 on the fast engine. For analysis of block level responses (satisfaction and preference ratings collected after each block of 10 searches) this independent variable is the block of paired queries (query block of the 6 such blocks).

Order Throughout the study, there were occasions when the name of one search engine would appear ”first/left” and the other ”second/right”. For example, for the 20 searches in which the participant was allowed to choose, we alternated which button appeared on the left and which appear on the right. But one of the two buttons has to appear on the left for the first search. Similarly,

participants used sliders to express preference (see Figure 1). One of the labels appeared on the left and the other appeared on the right (this ordering was maintained every time this screen appeared for the same participant). For all of these "order" variables, half of the participants experienced one scheme, half experienced the other.

These design variables motivated 2 (color) x 2 (query set) x 2 (order) = 8 participants per slow latency for a counterbalanced design. Because latency differences are more difficult to discern for latencies closer together, we elected to have 16 participants for the 2 second slow latency level.

3.2 Dependent Variables

Stated preference We measured stated preference on a 0-50 scale in either direction around 0, with 50 anchoring each end of the scale and indicating strong preference for one of the search engines. Zero indicated no preference. Attached to each preference score is the direction of preference: blue or yellow. The experiment solicited preference after each of the 12 no-choice blocks and again at end of the session. The 12 no-choice block preferences were intended to be independent, with instructions asking participants to base their score only the searches in the immediately preceding block. The final preference score was intended to measure preference for the entire evaluation, with instructions worded to this effect.

Although preference was solicited on a 0-50 scale in terms of brand, these preference scores are easily converted into a preference score on the scale 0 to 100 for the fast engine of the pair. For example, if a participant score is Blue 40, and blue is the slow engine of the pair, this score translates into a preference of 10 for the fast engine (in this case yellow). Analysis focuses on preference for the fast engine rather than brand (color) preference.

Satisfaction We measured satisfaction with each search engine on a 0-100 scale, with 100 indicating very satisfied and 0 indicating very unsatisfied. The experiment solicited satisfaction after each of the 12 no-choice blocks and again at the end of the session. The 12 no-choice block satisfaction scores were intended to be independent, with instructions asking participants to base their score only the 10 searches in the immediately preceding block. The final satisfaction score was intended to measure satisfaction for each engine over the entire evaluation, with instructions worded to this effect.

Satisfaction is arguably the most subjective outcome measured in this study. For analysis, we consider the difference between the satisfaction score for the fast engine and the satisfaction score for the slow engine.

Perception We measured perception on a 0-50 scale in either direction around 0, with 50 anchoring each end of the scale and indicating one search engine was much faster. Zero indicated no difference. Attached to each score is the direction of difference: blue or yellow. A score of Blue 40, for example, means the participant perceived the blue search engine as the faster of the pair with "how much faster" quantified by a score of 40 out of a possible 50. This outcome was only solicited once at the end of evaluation in order to avoid introducing bias in other outcomes.

Like stated preference, these perception scores are converted into scores measuring the perception of the fast engine as the faster of the pair.

Choice We measured choice (observed preference) by the frequency that a participant selected a search engine during the choice blocks. For example, if a participant chooses blue 6 out of 10 times in a choice block, their observed preference for blue is 6 out of 10. Equivalently, their observed preference for yellow is 4 out of 10. The choice outcome is observed twice, once for the initial choice block, in which both engines have the slower latency, and once for the second choice block at the end of the session, in which the two engines have different latencies. Although there were no explicit instructions, it was intended that participants choices in the final choice block reflect any preference acquired during the course of the evaluation.

Like stated preference and perception, choice frequencies are converted into the frequency of selecting the fast engine, rather than the frequency of selecting blue or yellow.

3.3 Statistical Methodology

Regression is the statistical method utilized in this analysis. Regression models an dependent (outcome) variable as a function of one or more independent (predictor variables) plus a random error, e [12]. For example, we can model preference for the fast engine, y, as a function of the latency of the slow search engine, x:

$$y = 50 + \beta x + e \tag{1}$$

The coefficient of x, β, describes the association between the slow search engine latency and the outcome, preference for the fast engine. Specifically, β is the expected change in the preference score associated with a 1 second increase in the latency of the slow engine. Note the coefficient is a statement about the expected value of preference score, or equivalently the average of a sample of preference scores. Individual scores will vary, and this is modeled via the random error, e.

β is estimated from the data, the values of x and y. Each value of y is an outcome measured and the corresponding x is based on the study design. If the estimated β is too close to 0, then we conclude than there is not enough evidence for an association between the latency of the slow engine latency and the preference score. This assessment of whether β is too close to 0 is a hypothesis test. In the process of estimating the coefficient β from data, we also estimate a standard error for β. The standard error is a statistical yardstick for measuring if β is too close to 0. If β is outside of 2 standard errors of 0 (roughly), we say the association between slower search engine latency and preference score is "statistically significant".

In general, a statistically significant association does not imply causation. But in the context of a controlled experiment, such as this study, we can make causal conclusions; that is, we can conclude that increasing the latency of the slower search engine causes a change in preference score.

In this study, we chose slow latencies of 2 s, 3 s, 4 s, and 5 s to be compared to a fast latency of 250 ms. Rather than model preference score as linear function of latency, we model each of the levels separately. This approach had two advantages:

- It allows for a non-linear association between the slow engine latency and preference score.
- It allows for separate hypothesis tests for each latency level. This gives us the ability to pinpoint the interval during which the latency of the slow engine starts to impact preference score (or other outcome variable).

Let $x_1, ...x_4$ be indicators of whether the slower search engine latency is 2 s, 3 s, 4 s, or 5 s. One of these predictor variables will be 1 and other three 0 for each outcome. Then the model equation is:

$$y = 50 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e \qquad (2)$$

The fact that we can conduct a separate hypothesis test for each latency coefficient does not preclude a joint hypothesis test of all latency coefficients. For the joint hypothesis test, the null hypothesis is "all coefficients are 0". If we reject the null hypothesis, then we conclude "at least 1 coefficient is non-zero". Refer to [12] for how to conduct a hypothesis test of multiple coefficients.

Equation 2 has a fixed intercept of 50. Usually, a regression model has a coefficient for the intercept which, like the other coefficients, is estimated from the data. However, in this case it is not possible to estimate an intercept coefficient. In order to estimate an intercept coefficient, there would have to be data outcomes for which $x_1, ...x_4$ are all 0. Then, the intercept coefficient would estimate the expected preference score for the fast search engine, given two search engines with identical latency. But if the study design is valid, this should be 50 = no preference. Hence, rather than include a latency pair in the study with identical "fast" and "slow" search engine latencies, we simply assume that the intercept is 50.

The outcome variables, with the exception of the choice outcome, are subjective and not calibrated across participants. The regression models do not account for this, so we must scrutinize any statistically significant result. If a result is due to unusually high ratings from 1 or 2 participants, we err on the side of caution and identify the result as inconclusive. The analysis of block ratings is more sensitive to the influence of individual participants than the analysis of ratings collected at the end of the study session. In the former, each participant contributes 12 scores, while in the latter, each participant only contributes one score.

Table 1: Block Preference and Satisfaction by Slow Latency

	Pref for Faster			Sat Faster - Slower		
	[0,49]	50	[51,100]	[-100,-1]	0	[1,100]
2 s	32	117	43	32	104	56
3 s	37	47	12	35	45	16
4 s	19	47	30	21	47	28
5 s	10	69	17	10	64	22

Table 2: Block Preference and Satisfaction by Query Block

	Pref for Query Set 1			Sat Set 1 - Set 2		
	[0,49]	50	[51,100]	[-100,-1]	0	[1,100]
1	18	41	21	21	37	22
2	10	50	20	10	51	19
3	9	51	20	12	49	19
4	14	45	21	16	39	25
5	17	48	15	20	44	16
6	17	45	18	21	40	19

4. Data Overview

4.1 Block Outcomes

Block outcomes refer to the satisfaction and preference ratings collected after each block of 10 searches. Table 1 gives response frequencies by latency of the slow engine. Table 1 suggests no clear pattern as the slow latency increases. For example, there appears to be slight preference for the fast engine and higher satisfaction with the fast engine when the slow latency is 2 s; however, Table 1 also suggests a slight preference for the slow engine and higher satisfaction with the slow engine when the slow latency is 3 s. Most responses reflect no difference in preference or satisfaction between the two engines. Each participant contributes 12 responses to the frequencies in table 1, and it is possible that 1 or 2 participants are driving an apparent pattern. Table 1 also ignores the magnitude of preference and satisfaction scores, although such scores are not calibrated across participants.

Table 2 gives response frequencies by the block of paired queries (query block). In table 2, *Pref for Query Set 1* refers to the preference score for whichever engine the participant was forced to use for the first query of each pair. *Sat Set 1 - Set 2* refers to difference between the satisfaction score for whichever engine the participant was forced to use for the first query of each pair and the satisfaction score for the other engine (used for the second query of each pair). The first column of table 2 refers to the block of paired queries.

The frequencies of Table 2 put the frequencies of Table 1 in perspective. Comparing the two tables suggest the specific queries a participant sees on give search engine are at least as important in determining block preference and satisfaction rating as latency.

4.2 Final Outcomes

Final outcomes refer to the satisfaction, preference, and perception ratings collected at the end of each participant session, as well as selection frequency for the final choice block. The

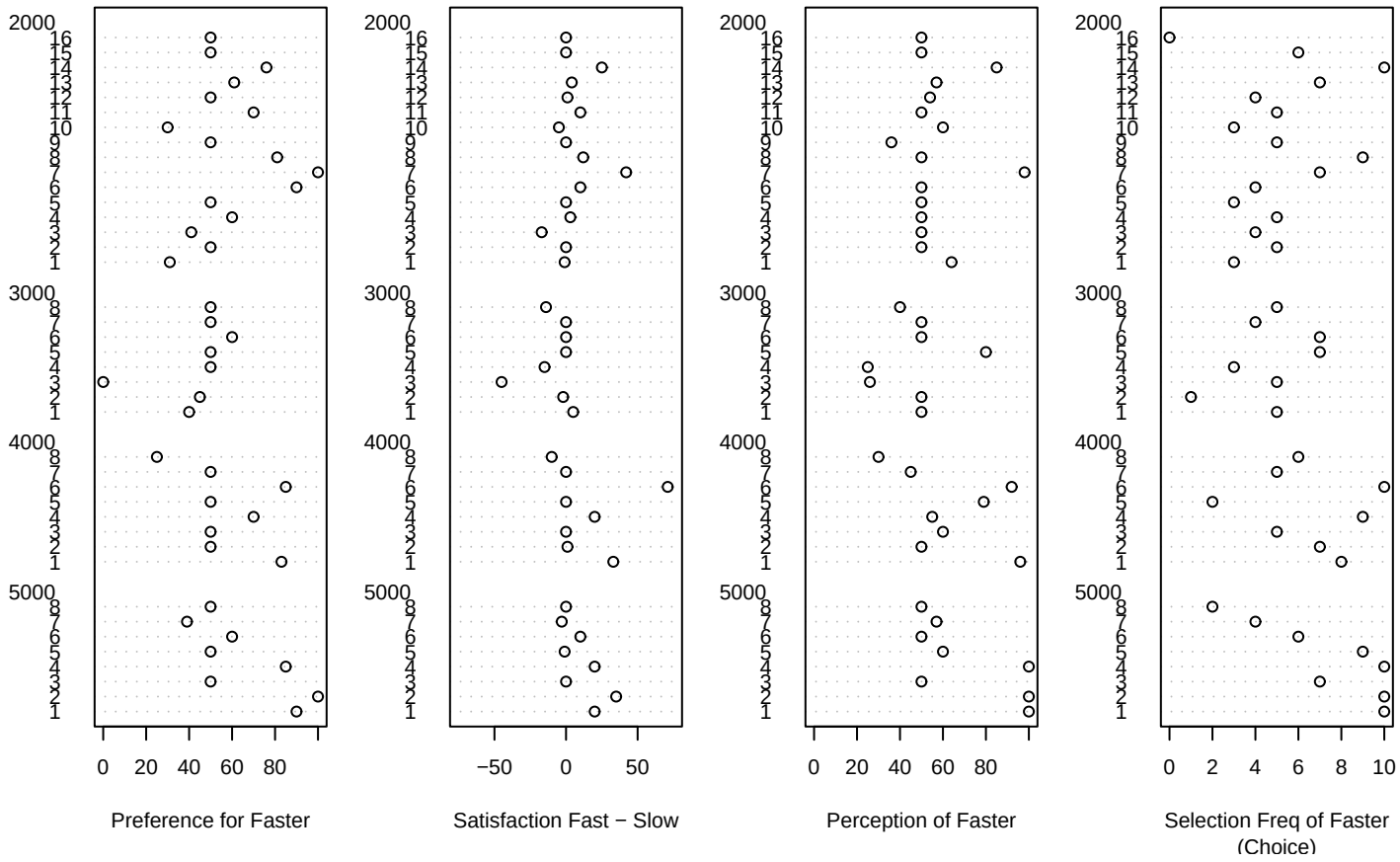

Figure 6: Final Outcomes by Slow Latency and Participant

Table 3: Correlations between Final Outcomes

	Sat. Diff.	Perception	Choice
Preference	0.83	0.70	0.59
Sat. Diff.	-	0.76	0.58
Perception	-	-	0.55

dot charts in Figure 6 display these outcomes by participant, categorized by the latency of the slower search engine (in milliseconds).

In Figure 6, it is clear that the most common stated preference and perception response is 50, indicating no preference and no perception of one engine as faster than the other. Similarly, the most common satisfaction difference is 0. For perception and choice, there are 2 participants with slow latency of 4 s (1,6) and 3 participants (1,2,4) with slow latency of 5 s that both perceive the fast engine as faster and who selected the fast engine at least 8 times during the final choice block.

Participant responses across the four outcomes are correlated. Table 3 lists the pairwise correlations[1] for these outcomes.

[1]Pearson correlation coefficients.

5. Association between Stated Preference and Latency

5.1 Block Preference

There is no detectable association between block preference scores and latency. We applied the following regression model:

$$y_{kl} = 50 + \beta_{1i}x_{1i} + \beta_{2j}x_{2j} + \beta_3 x_3 + \beta_4 x_4 + e_{kl} \qquad (3)$$

where

y_{kl} stated preference for the fast engine after block l, $l = 1, ...12$ for participant k, $k = 1, ...8$

x_{1i} indicator of latency of the slow search engine, with $i = 1, ...4$ corresponding to latency levels 2s,...5s. For example, if a participant k is assigned to 2 s latency level, then $x_{11} = 1$ and $x_{12} = x_{13} = x_{14} = 0$ for that participant.

x_{2j} indicator of query block j corresponding to block l. If block l is not assigned query block j, $x_{2j} = 0$. If block l is assigned query block j, then $x_{2j} = 1$ if query set 1 is assigned to the fast search engine (hence set 2 is assigned

Table 4: Coefficients for Block Preference

Coefficient	Value	Std. Error
Slow 2 s β_{11}	1.12	2.36
Slow 3 s β_{12}	-8.83	3.33
Slow 4 s β_{13}	3.78	3.33
Slow 5 s β_{14}	1.76	3.33
Query Block 1 β_{21}	4.44	1.54
Query Block 2 β_{22}	2.3	1.54
Query Block 3 β_{23}	2.88	1.54
Query Block 4 β_{24}	0.57	1.54
Query Block 5 β_{25}	0.13	1.54
Query Block 6 β_{26}	4.13	1.54
Color β_3	0.35	1.49
Order β_4	0.89	1.49

to the slow engine) and $x_{2j} = -1$ if query set 2 is assigned to the fast search engine (hence set 1 is assigned to the slow search engine).

x_3 indicator of color, $x_3 = 1$ if blue is the fast search engine for this participant, $x_3 = -1$ if yellow is the fast search engine for this participant.

x_4 indicator of order, $x_4 = 1$ if fast search engine has order 1, $x_4 = -1$ if slow search engine has order 1.

e_{kl} random correlated error; for $k \neq k'$ (different participants), $Cor(e_{kl}, e_{k'l}) = 0$. For $l \neq l'$ (different blocks for the same participant k), $Cor(e_{kl}, e_{kl'}) = \rho > 0$.

The estimates[2] of the model coefficients are in Table 4.

Each β_{1i} coefficient is the deviation in average block preference above or below 50 (=no preference) for the fast engine with 250 ms latency when the slow engine has latency level i. The model assumes the average preference score for the "faster" of two search engines with identical latency is 50 (no preference). For example, the estimate of β_{11} is 1.12. This means on average the block preference score is 51.12 for the fast engine when the slow engine had a latency of 2 s.

Each β_{2j} coefficient is the deviation in average block preference above or below 50 for the fast engine for a block assigned query block j and query set 1 assigned to the fast engine. For a block assigned to query block j and query set 2 assigned to the fast engine, the deviation is $-\beta_{2j}$. For example, the estimate of β_{21} is 4.44. This means on average the block preference score is 54.44 for the fast search engine for a block assigned the block 1 set of paired queries and the first query of each pair assigned to the fast engine. In contrast, if the second query of each pair is assigned to the fast engine, then on average the preference score is 45.56 for the fast engine.

The β_3 coefficient is the deviation in average block preference above or below 50 for the fast engine if the fast engine brand is blue. $-\beta_3$ is the deviation if fast engine is yellow.

The β_4 coefficient is the deviation in average block preference above or below 50 for the fast engine if the fast engine brand has order 1. $-\beta_4$ is the deviation if fast engine has order 2.

[2]Maximum likelihood (as opposed to REML) estimates, $\hat{\rho} = 0.28$.

Table 5: Coefficients for Final Preference

Coefficient	Value	Std. Error
Slow 2 s β_{11}	8.75	5.25
Slow 3 s β_{12}	-6.87	7.43
Slow 4 s β_{13}	7.87	7.43
Slow 5 s β_{14}	15.5	7.43
Query Set β_2	-1.1	3.32
Color β_3	-2.05	3.32
Order β_4	-0.2	3.32

As a group, the latency coefficients are not statistically significant[3]. As a group, the query block coefficients are statistically significant[4], confirming that the queries and/or the search results presented for those queries do influence block preference ratings.

5.2 Final Preference

There is no detectable association between final preference scores and latency. We applied the following regression model:

$$y = 50 + \beta_{1i}x_{1i} + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e \qquad (4)$$

where

y stated preference for the fast engine

x_{1i} indicator of latency of the slow search engine, with $i = 1, ...4$ corresponding to latency levels 2s,...5s. For example, if a participant is assigned to 2 s latency level, then $x_{11} = 1$ and $x_{12} = x_{13} = x_{14} = 0$ for that participant.

x_2 indicator of query set, $x_2 = 1$ if a participant sees query set 1 first on the fast engine (in the the second 6 blocks of the 12 no-choice blocks, this participant sees query set 1 on the slow engine) and $x_2 = -1$ if a participant sees query set 1 first on the slow engine.

x_3 indicator of color, $x_3 = 1$ if blue is the fast search engine, $x_3 = -1$ if yellow is the fast search engine.

x_4 indicator of order, $x_4 = 1$ if fast engine has order 1, $x_4 = -1$ if slow engine has order 1.

e random uncorrelated error.

The estimates of the model coefficients are in Table 5.

Each β_{1i} coefficient is the deviation in average final preference above or below 50 (=no preference) for the fast engine with 250 ms latency when the slow engine has latency level i. The model assumes the average preference score for the "faster" of two search engines with identical latency is 50 (no preference). For example, the estimate of β_{11} is 8.75. This means on average the final preference score is 58.75 for the fast engine when the slow engine had a latency of 2 s.

[3]Likelihood ratio test statistic 8.2 on 4 df.

[4]Likelihood ratio test statistic 21.3 on 6 df.

Table 6: Coefficients for Block Satisfaction

Coefficient	Value	Std. Error
Slow 2 s β_{11}	2.45	1.93
Slow 3 s β_{12}	-7.12	2.74
Slow 4 s β_{13}	4.31	2.74
Slow 5 s β_{14}	1.98	2.74
Query Block 1 β_{21}	4.8	1.47
Query Block 2 β_{22}	1.92	1.47
Query Block 3 β_{23}	1.59	1.47
Query Block 4 β_{24}	1	1.47
Query Block 5 β_{25}	2.5	1.47
Query Block 6 β_{26}	4.56	1.47
Color β_3	-1.04	1.22
Order β_4	0.78	1.22

The β_2 coefficient is the deviation in average final preference above or below 50 for the fast engine if a participant sees query set 1 first on the fast engine. $-\beta_3$ is the deviation if a participant sees query set 1 first on the slow engine.

The β_3 coefficient is the deviation in average final preference above or below 50 for the fast engine if the fast engine brand is blue. $-\beta_3$ is the deviation if fast engine is yellow.

The β_4 coefficient is the deviation in average final preference above or below 50 for the fast engine if the fast engine brand has order 1. $-\beta_4$ is the deviation if fast engine has order 2.

As a group, the latency coefficients are not statistically significant[5].

6. Association between Satisfaction and Latency

6.1 Block Satisfaction

There is no detectable association between block differences in satisfaction scores and latency. The regression model for detecting an association between a difference in satisfaction scores and latency is identical to model equation 3:

$$y_{kl} = \beta_{1i}x_{1i} + \beta_{2j}x_{2j} + \beta_3 x_3 + \beta_4 x_4 + e_{kl} \qquad (5)$$

All terms are defined as in equation 3, except y_{kl}. In equation 5, y_{kl} is the (satisfaction for the fast engine - satisfaction for the slow engine) for satisfaction ratings from block l, $l = 1, ...12$ and participant k, $k = 1, ...8$. Note that the intercept of equation 5 is 0.

The estimates[6] of the model coefficients are in Table 6. The interpretations of coefficients are similar to those for equation 3.

As a group, the latency coefficients are statistically significant[7]. The large negative coefficient for a slow latency of 3 s indicates the slow engine receives higher satisfaction scores than the fast engine (on average) when the latency is 3 s. The statistical significance of this coefficient is due the satisfaction scores of participants 3 and 4. Omitting either participant from the analysis shrinks the coefficient to -5.5. Both of these participants gave a higher satisfaction score to the slow search engine for 9 out of 12 blocks. Rather than reach a conclusion based on these two participants alone, we ignore this result.

As a group, the query coefficients are statistically significant[8]. Queries and/or the search results presented for those queries do influence the difference in satisfaction scores.

[5]F-test statistic is 2.28 on 4 and 33 df.

[6]Again, these are ML estimates, $\hat{\rho} = 0.21$.

[7]Likelihood ratio test statistic is 10.2 on 4 df.

Table 7: Coefficients for Final Satisfaction

Coefficient	Value	Std. Error
Slow 2 s β_{11}	5.25	4.48
Slow 3 s β_{12}	-8.87	6.33
Slow 4 s β_{13}	14.37	6.33
Slow 5 s β_{14}	10.12	6.33
Query Set β_2	0.42	2.83
Color β_3	-2.07	2.83
Order β_4	0.22	2.83

6.2 Final Satisfaction

There is no detectable association between differences in final satisfaction scores and latency. We applied the following regression model:

$$y = \beta_{1i}x_{1i} + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e \qquad (6)$$

All terms are defined as in equation 4, except y. In equation 6, y is (satisfaction score for the fast engine - satisfaction score for the slow engine), for scores collected at the end of the study session. Note that the intercept of equation 6 is 0.

The estimates of the model coefficients are in Table 7. The interpretations of coefficients are similar to those for equation 4.

As a group, the latency coefficients are statistically significant[9]. The coefficient for slow latency of 4 s indicates greater satisfaction with the fast engine when the latency is 4 s. However, the magnitude of this coefficient is due solely to participant 6. Omitting this participant shrinks β_{13} to 5.82. As suggested by Figure 6, the difference in satisfaction scores is large for this participant not just among participants experiencing 4 s slow latency, but among all participants. The influence of this participant is further exaggerated because satisfaction score differences are less variable than preference or perception scores. Therefore, we ignore this result.

7. Association between Perception and Latency

When the slow search engine latency is 5 s, some participants perceive the fast search engine as the faster of the pair. We applied the following regression model:

$$y = 50 + \beta_{1i}x_{1i} + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e \qquad (7)$$

All terms are defined as in equation 4, except y. In equation 7, y is perception score that the participant perceives the fast engine as the faster of the pair at the end of the study session.

[8]Likelihood ratio test statistic 26.4 on 6 df.

[9]F-test statistic is 2.76 on 4 and 33 df.

Table 8: Coefficients for Perception

Coefficient	Value	Std. Error
Slow 2 s β_{11}	6.5	5.02
Slow 3 s β_{12}	-3.62	7.1
Slow 4 s β_{13}	13.37	7.1
Slow 5 s β_{14}	20.87	7.1
Query Set β_2	0.07	3.17
Color β_3	0.08	3.17
Order β_4	-2.28	3.17

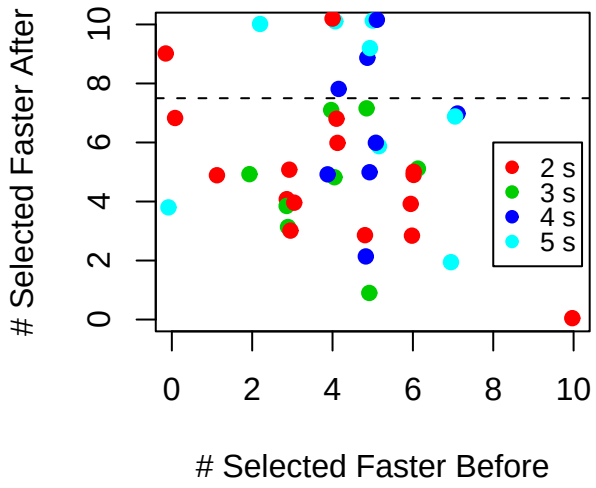

Figure 7: Choice Outcome by Slow Latency

The estimates of the model coefficients are in Table 8. The interpretations of coefficients are similar to those for equation 4.

As a group, the latency coefficients are statistically significant[10]. The coefficient for slow latency of 5 s indicates a perception of the fast engine as faster when the slow engine latency is 5 s. This coefficient is statistically significant[11] and supported by participants 1,2, and 4. These are same three participants who expressed a preference for fast engine (see Figure 6), but they are more certain in their perception than their preference. Even if one of the three is omitted from the analysis, the coefficient remains statistically significant.

8. Association between Choice and Latency

The choice outcome has several advantages compared to the other outcome measures:

- It is based on observing participants rather than soliciting their opinion. This eliminates the concern about a lack of calibration across users.
- In the initial choice block, there is no latency difference between the search engines. Participant choices in this choice block serve a baseline.
- The initial choice block presumably captures a participant's color preference. A participant's specific preference is more accurate than a population preference for blue or yellow, which is what the color variable represented in the regression models for stated preference, satisfaction, and perception.

Figure 7 is a scatterplot of the number of times each participant selected the fast engine in the initial choice block (before it was fast) and the final choice block. The data points are jittered so that overlapping points can be distinguished. The horizontal spread of the data (choice before) is less than the vertical spread of the data (choice after), suggesting that latency changes do impact observed preference. Points above the dashed horizontal line are participants choosing the fast engine 8 or more times after the latency change. For 4 s and 5 s, there are 3 and 4 participants respectively. While there are also 2 participants for the 2 s slow latency, there are 16 participants for 2 s latency, and $2/16$ is less than half of $3/8$.

As suggested by Figure 7, when the slow search engine latency is 4 s or 5 s, participants are more likely to choose the fast search engine than the slow search engine. We applied the following logistic regression model:

$$log(p/(1-p)) = \beta_{1i}x_{1i} + \beta_{2j}x_{2j} + e \qquad (8)$$

where:

p proportion of choices for the fast engine (probability of choosing the fast engine)

x_{1i} indicator of participant, $i = 1, ...40$ For participant i, $x_{1i} = 1$ for both of the choice outcomes (initial block and final block) observed for that participant. For another participant, i', $i' \neq i$, $x_{1i} = 0$.

x_{2j} indicator of latency of the slow search engine assigned to this participant for the second choice outcome, with $j = 1, ...4$ corresponding to latency levels 2s,...5s. If participant i is assigned to slow latency j, then only for the final choice block is $x_{2j} = 1$. For the initial choice block, $x_{2j} = 0$.

e = random uncorrelated error

The participant coefficients β_{2j} obviate the need for a color variable. Based on the fact that query set and order were not statistically significant in final preference, final satisfaction, or perception models, we omit these variables from this model.

Table 9 lists the estimates of the coefficients of equation 8.

The participant coefficients, β_{1i}, can be interpreted as baseline odds or baseline probability for each participant to select the fast engine via the transformations:

$$odds = exp(\beta_{1i})$$
$$probability = \frac{exp(\beta_{1i})}{1 + exp(\beta_{1i})}$$

[10] F-test statistic is 3.53 on 4 and 33 df.
[11] The t-statistic is 2.94, and the p-value is 0.006.

Table 9: Coefficients for Choice

Coefficient	Value	Std. Error	Odds Ratio/Mult.	Prob.
Participant Coefficients β_{1i}				
1, 2s	-0.42	0.47	0.66	0.4
1, 3s	-0.34	0.48	0.71	0.42
1, 4s	0.08	0.49	1.09	0.52
1, 5s	0.52	0.55	1.67	0.63
2, 2s	-0.01	0.47	0.99	0.5
2, 3s	-0.99	0.52	0.37	0.27
2, 4s	0.54	0.51	1.71	0.63
2, 5s	0.24	0.53	1.27	0.56
3, 2s	-0.21	0.46	0.81	0.45
3, 3s	0.07	0.48	1.07	0.52
3, 4s	-0.33	0.48	0.72	0.42
3, 5s	0.24	0.53	1.27	0.56
4, 2s	-0.62	0.47	0.54	0.35
4, 3s	-0.99	0.52	0.37	0.27
4, 4s	0.54	0.51	1.71	0.63
4, 5s	-0.25	0.51	0.78	0.44
5, 2s	-0.62	0.47	0.54	0.35
5, 3s	0.07	0.48	1.07	0.52
5, 4s	-0.97	0.51	0.38	0.28
5, 5s	0.24	0.53	1.27	0.56
6, 2s	-0.84	0.49	0.43	0.3
6, 3s	0.27	0.48	1.31	0.57
6, 4s	0.79	0.54	2.21	0.69
6, 5s	-0.48	0.51	0.62	0.38
7, 2s	-0.01	0.47	0.99	0.5
7, 3s	-0.76	0.5	0.47	0.32
7, 4s	-0.54	0.49	0.58	0.37
7, 5s	-2.25	0.64	0.11	0.1
8, 2s	-0.42	0.47	0.66	0.4
8, 3s	-0.76	0.5	0.47	0.32
8, 4s	-0.13	0.48	0.88	0.47
8, 5s	-0.94	0.52	0.39	0.28
9, 2s	-1.07	0.51	0.34	0.26
10, 2s	-1.07	0.51	0.34	0.26
11, 2s	-0.01	0.47	0.99	0.5
12, 2s	-0.84	0.49	0.43	0.3
13, 2s	-0.84	0.49	0.43	0.3
14, 2s	0.64	0.5	1.9	0.66
15, 2s	-0.21	0.46	0.81	0.45
16, 2s	-0.21	0.46	0.81	0.45
Latency Coefficients				
Slow 2 s β_{11}	0.43	0.23	1.53	0.6
Slow 3 s β_{12}	0.27	0.33	1.31	0.57
Slow 4 s β_{13}	0.67	0.34	1.95	0.66
Slow 5 s β_{14}	1.42	0.37	4.14	0.81

The transformed coefficients are listed in columns 4 and 5 of Table 9. Participant coefficients are labeled by participant number and the slow latency. For example, the coefficient for participant 1 for 2 s is -0.42. This participant's estimated baseline odds for preferring the fast engine are 0.66 or about 7:10. Expressed as a probability, the estimated probability of selecting the fast engine for this participant is 0.4. Of course, we assume a priori preference for the fast engine is actually the branding preference. For participant numbers 1-4, blue is the fast engine, so the coefficient suggests this participant is biased towards yellow, although the coefficient is not statistically significant.

The latency coefficients, β_{2j}, can be interpreted as odds multiplier via the transformation odds multiplier = $exp(\beta_{2j})$. To understand an odds multiplier, consider a participant with baseline odds of choosing search engine A to B of 2:3, given search engine A and B have the same latency of 4s. This preference for B is presumably based on the branding. Now suppose the latency for A improves to 250 ms. From Table 9, β_{13}=1.95 or approximately 2. The baseline odds are multiplied by the odds multiplier, so the odds become 4:3. With the latency change, the participant now prefers A. So $exp(\beta_{2j})$ is the expected change in the odds ratio when the latency of one of two search engines is improved to 250 ms from a previous shared latency of j.

The latency coefficients, β_{2j}, can also be interpreted as probabilities under the assumption that there is no prior preference for either search engine (that is, the baseline odds=1:1) via the transformation probability = $exp(\beta_{2j})/(1+exp(\beta_{2j}))$. That is, assuming a participant has no preference between two search engines, $exp(\beta_{2j})/(1+exp(\beta_{2j}))$ is the estimated probability the participant will choose the fast engine if the latency of the fast engine is 250 ms and the latency of the slow engine is j.

As a group, the participant coefficients are not statistically significant[12], suggesting most participants aren't predisposed to choose blue or yellow. Only three individual participant coefficients, participant 7 for 5 s and participants 9 and 10 for 2 s, are statistically significant. The first is biased towards blue and the second two are biased towards yellow.

As a group, the latency coefficients are statistically significant[13] The coefficients for 4 s and 5 s are both statistically significant. For 4 s, the conclusion rests on the three participants above the dashed line in Figure 7. These are participants 1, 4, and 6. If any of these three are omitted from the analysis, the coefficient is no longer significant. The conclusion for 5 s is stronger. If any of the participants above the dashed line in Figure 7 (participants 1,2,4 and 5) are omitted, the coefficient remains significant.

Some participants do choose a search engine with 250 ms latency over a search engine with 4s or 5s latency.

[12]Deviance 67.85 on 40 df.
[13]Deviance 23.93 on 4 df.

9. Choice as Function of Latency

The primary research question of this study is "how fast is fast enough": how large must the latency gap be between a speed-of-light search engine (latency 250 ms) and a slower search engine before there is a noticeable impact on user preference/choice? Based on the results of the previous section, the answer provided is "less than 4 seconds".

We now refine this answer, by assuming choice is a monotonic increasing function of the slow search engine latency. The data from this study is insufficient to validate this assumption. Nevertheless, we adopt it as a reasonable assumption.

The coefficients in Table 9 suggest a monotonic increasing function, except that the coefficients for 2 s and 3 s are inverted. That is, the odds multiplier for 3 s, 1.31, is less than the odds multiplier for 2 s, 1.53 even though 2 s $<$ 3 s. In order to address this, we fit a logistic regression model that pools the data for 2 s and 3 s - that is, we assume all the participants at 2 s and 3 s experienced the same latency for the slow search engine. This does not change any of the coefficient estimates presented previously, except that there are no longer coefficients for Slow 2 s and Slow 3 s but rather a single coefficient for "Slow 2 or 3 s". Table 10 gives the coefficient estimate.

Table 10: Coefficient for Slow Latency 2 or 3 s

Coef. Value	Std. Error	Odds Multiplier	Prob.
0.37	0.19	1.45	0.59

The coefficient for "Slow 2 or 3 s" is statistically significant. We associate this coefficient with a slow search engine latency of 2.5 s.

A monotonic increasing function of choice is obtained from the logistic regression model via linear interpolation between the 3 coefficients for 2.5 s, 4 s, and 5 s. If we add ± 2 standard errors to each coefficient prior to the applying the odds multiplier or probability transformation, we obtain confidence intervals at latency 2.5 s, 4 s, and 5 s. Applying linear interpolation to these confidence intervals generates confidence bands for the function. Figure 8 graphs the interpolated function for the odds multiplier and probability of selecting the fast engine given no prior preference.

Superimposed on the interpolated choice functions (solid line) and confidence bands (dashed lines) are points corresponding to the coefficients of the logistic regression model in Table 9.

The question "how fast is fast enough" can now be answered by inverting the choice function. First, decide what constitutes a "noticeable impact" on observed user preference. Second, express this as probability of selecting the fast search engine. Third, invert the choice function and read the latency target off of the x-axis. For example, if noticeable impact means the odds of choosing the faster engine are 1.5 to 1 (60%), then the corresponding latency target is 3 seconds.

The choice function does not really distinguish between "% of users" and "% of user searches"; in theory we can interpret the probability either way. However, figure 6 suggests that users either perceive (consciously or unconsciously) the latency difference and act on that perception, or they do not.

The lower confidence bound flat-lines between 2.5 and 4 s. The fact that the lower confidence bound is greater than or equal to the "no change in preference" line (odds multiple of 1) or "no preference" line (probability of 0.5) reflects the fact the pooled coefficient is statistically significant. Our confidence of an odds multiplier of 1.45 at "2 or 3 s" now matches our confidence in the odds multiplier of 1.95 at 4 s, although the confidence at an odds multiplier of 1.45 required 3 times the number of participants (24 vs. 8).

One may pose the question: would more participants allow us to estimate the choice function with confidence at 2 s or perhaps an even slower search engine latency? While the theoretical answer is yes, in practice the number of participants becomes cost prohibitive. Doubling the number of participants reduces the standard error by the factor $1/\sqrt{2}$. To detect odds of preference for the faster engine of 5:4 (odds multiplier 1.25) requires between 32 and 64 participants. Smaller differences in latency may indeed have some impact on user choice, but detecting such an impact is not feasible given this study design.

10. Conclusions

This study compared two mock search engines, one delivering search results in 250 ms and a slower search engine delivering search results in either 2, 3, 4, or 5 seconds. The key findings are:

- User perception, satisfaction, stated preference, and choice (observed preference) are moderately correlated.
- Regardless of slow search engine latency, user stated preference is inconclusive.
- Regardless of slow search engine latency, the difference in user satisfaction scores between the search engines is inconclusive.
- When the slower search engine latency is 5 seconds, some users state they perceive the faster engine as faster.
- When the slower search engine latency is 4 or 5 seconds, some users choose to use the faster engine more often.
- Based on pooling data for 2 s and 3 s, once latency exceeds 3 seconds for the slower engine, users are 1.5 times as likely to choose the faster engine.

11. Future Work

Given users can perceive latency differences on the order of a few hundred milliseconds [6, 9], users in this study seem rather insensitive to latency differences an order of magnitude larger (seconds). In part this is a limitation of the study design. The small sample and one hour exposure period are practical constraints. Similar constraints may have in part motivated previous studies of web page performance to use latencies on the order of seconds [5, 4].

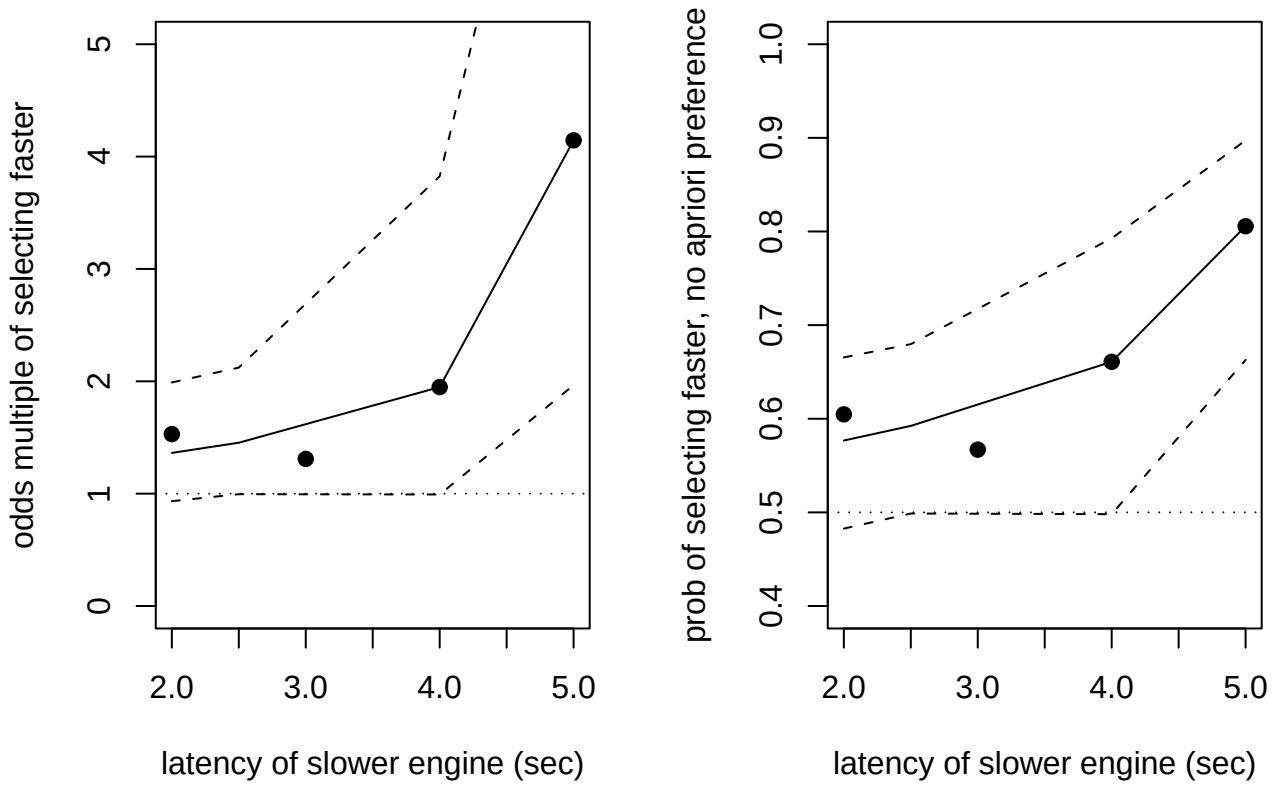

Figure 8: Choice as an Increasing Function of Slow Latency

However, within these constraints, there are potential design improvements. In a controlled experiment, it is difficult to replicate the time pressure a user might experience in the real world. The current design emulated time pressure by informing participant of the number of remaining searches as they progressed and including explicit instructions to complete the searches "quickly". An improved design could offer an incentive to finish quickly.

In this study, the choice outcome proved more effective than the other outcome measurs, in part due to the design decision to collect choice data in the first block. In hindsight, a before and after comparison of stated preference and satisfaction difference could be almost as valuable and is advisable for a future study.

There are directions of future work for small sample controlled experiments. In the real word, latency is not fixed for every visit to a search engines, and multiple characteristics of the latency distribution may influence search engine preference. These characteristics are easily manipulated in a controlled experiment. For example, users may be more sensitive to variable rather than fixed changes in latency, recent rather than older exposures to high latency, and sudden rather than gradual changes. They may also adapt to slower or faster latency over time. These are but some of the theories others have studied [8, 3, 1] in a more general context, and a future study could investigate them in the web search context.

References

[1] Bhatti, N., Bouch, A., Kuchinsky, A. (2000). "Integrating user-perceived quality into web server design," *Computer Networks*, **33**, 1-16.

[2] Broder, A. (2002). "A taxonomy of web search." *ACM SIGIR Forum*, **36**, 3-10.

[3] Fischer, A., Blommaert, F. (2001). "Effects of time delay on user satisfaction," *Proceedings of the International Conference on Affective Human Factors Design*, 407-414.

[4] Galletta, D., Henry, R., McCoy, S., Polak, P. (2004). "Web site delays: How tolerant are users?" *Journal of the Association for Information Systems*, **5**, 1-28.

[5] Nah, F. (2004). "A study on tolerable waiting time: how long are web users willing to wait?" *Behaviour & Information Technology*, **23**, 153-163.

[6] Johnson, J. (2000), *GUI Bloopers*, Academic Press, chapter 7.

[7] Jupiter Research (2006), *Retail Web Site Performance: Consumer Reaction to a Poor Online Shopping Experience*, Vendor Research commissioned by Akamai. (http://www.akamai.com/4seconds).

[8] Kahneman, D. Tversky, A. (2000), *Choices, Values and Frames*, Cambridge University Press, chapter 38.

[9] Nielsen, J. (1993), *Usability Engineering*, Academic Press, chapter 5.

[10] Rose, D. Levinson, D. (2004). "Understanding user goals in web search." *Proceedings of the 13th international conference on World Wide Web*, 13-19.

[11] Sevcik, P. (2003). "How fast is fast enough?" *Business Communications Review*, **33**, March 2003.

[12] Weisberg, S. (1985). *Applied Linear Regression*, John Wiley & Sons.