A Four Group Cross-Over Design for Measuring Irreversible Treatments on Web Search Tasks

Li Ma
Stanford University
Stanford, CA 94305
ma2@stanford.edu

David Mease
Google
Mountain View, CA 94043
dmease@google.com

Daniel M. Russell
Google
Mountain View, CA 94043
drussell@google.com

Abstract— **When trying to measure the effect of irreversible treatments such as training interventions, the choice of the experimental design can be difficult. A two group cross-over experimental design cannot be used due to longitudinal effects during the course of the experimental run, which can be especially large in dynamic web search environments. A standard case/control two group design also can be problematic because it is negatively impacted by variability among participants. Our solution uses a four group cross-over design that combines features from standard cross-over and case/control designs. We illustrate the effectiveness of this design through a case study in which participants are shown a video on how to use "control-F" to search for text within a web page. We quantify the improvement of our four group cross-over design compared to the standard case/control two group design with respect to measuring the effect of this video on participants. Finally, we compare the magnitude of our estimated "control-F" effect to similar studies on web ranking and user-interface changes, revealing that teaching this skill produces a potentially large improvement in web searchers' ability to search rapidly.**

1. Introduction

Web search engines constantly strive to enhance users' search experiences by improving the quality of the results returned as well as the user interface. However, there is also tremendous unrealized potential for improving the search experience by educating the users on useful features. Designing experiments to measure the impact of such educational interventions is complicated by the fact that the effect of interest is *irreversible*. That is, once a participant is taught a skill, he or she can not "unlearn" the skill.

In this paper we will focus on the general problem of evaluating the impact of any irreversible treatment. As an example we will consider a study in which participants are shown a video instructing them how to use "control-F" to search for text within a web page. We will use time until task completion as our evaluation metric. This metric has been used successfully for evaluation of ranking [8] as well as user interface changes [3] in web search.

We will consider standard options for experimental designs used in other task completion time studies. For an irreversible treatment effect, the implication is that the treatment must always be applied subsequent to the control whenever both are applied to the same participant. As a consequence, a standard cross-over design would confound other changes that occur over the course of the experiment with the treatment effect of interest. As such, we propose a four group cross-over design. We illustrate the proposed design using the "control-F" intervention and show that our design produces a more stable estimate of the treatment effect than when using a standard case/control two group design.

The motivation for using the "control-F" video as our educational intervention comes from a survey conducted during 2009, in which we found that relatively few web searchers were aware of (or use) "control-F" (or its equivalents, Command - F on the Macintosh, or menu items Edit > Find). More specifically, only 9.6% of a random sample (n=1000) of web searchers knew the "control-F" idiom. We also have found many instances of natural web search problems that require scanning a long landing page—a behavior for which knowing "control-F" is an important skill. Not knowing this skill causes simple search problems to become needlessly complex. Our observations are consistent with Buchanan and Loizides [2] who showed that "control-F" can significantly improve common reading tasks encountered during information seeking tasks. However, despite long history of "control-F" in document processing systems of many kinds, even for those who report knowing "control-F", they simply might not recall the keyboard shortcut at the appropriate time in use context [6]. Informing or reminding users of this option can greatly improve their search efficiency.

The paper is organized as follows. Section 2 describes the instructional video shown to the participants. Section 3 describes different candidate experimental designs for evaluating the effect of this video. Section 4 describes the data collection process and the nature of the data. Section 5 gives details of the statistical analysis and the overall results of the experiment. Section 6 provides some additional analysis for the individual tasks. Section 7 compares the magnitude of the effect estimated in this study to those in other task completion time studies. Finally, Section 8 presents some concluding remarks.

2. The Instructional Video

The instructional video we will use as an our example in this paper is available at

http://www.youtube.com/watch?v=D-FkKwRD4iY

The video is 37 seconds in length. It instructs participants how to use the "control-F" feature to find text on a page and provides two examples. The video shows the keyboard keys as well as a user demonstrating the feature, with a commentator providing the instruction. For example, part of the commentary states,"a great trick to know is how to use the on page find command to locate just the words you're looking for on the page."

3. Experiment Design

We carried out our study with paid participants, each assigned a set group of predefined web search tasks, similar to the studies in [3], [8]. The design issues we explore in this section center around which of these paid participants are assigned which tasks, as well as what participants are shown the video at what time. Section 4 will provide more specific information about the nature of the tasks, the number of participants and the participant instructions. For the remainder of this section, we consider the general experimental design problem of assigning participants to tasks and treatments.

Case/control two group design: For measuring the benefit of web search instructional intervention, some standard designs immediately come to mind. The most basic is what we will refer to as the case/control two group design. In such a design the study participants are randomly split into two disjoint groups: control and treatment. The the treatment group is given the intervention (shown the video), and the control group is not. The impact of the intervention can then be measured by comparing the metric of interest across the two groups.

Drawback of case/control two group design: An obvious drawback of this case/control two group design is that differences between study participants are confounded with the treatment impact measurement. In our study the metric of interest is the time until task completion. If there is substantial variability between participants in terms of task completion time, this will directly lead to variability in the treatment effect estimation under such a design. If the participants in the treatment group happen by chance to be faster than the control group, this will lead to a larger estimate for the treatment effect. Likewise, if slower participants happen to be selected for the treatment group, this will lead to a lower estimate for the treatment effect.

We will demonstrate in Section 5.3 that this participant variability is in fact quite substantial for our study.

Standard cross-over design: Xu and Mease [8] used a two-group *cross-over design* to overcome this difficulty for measuring the user impact of a web ranking change on task completion time. A cross-over design [5, p.197] is a well known technique that can significantly improve the precision of estimating treatment effects when the variability of experiment subjects (participants) is large relative to the treatment effect size of interest. In our setting a cross-over design would imply that half of the participants complete the first half of the tasks without the instructional video while the other half of the participants complete the other half of the tasks with the instructional video. Then the conditions are reversed so that in the end, all participants have done all tasks and all participants have been exposed to both the treatment and control conditions. Additionally, no participant has done the same task more than once, which is an important restriction in our setting. A critical consideration is that participant variation is often quite large, as shown in [8].

Drawback of standard cross-over design: Unfortunately, the cross-over design is problematic for our setting due to an irreversible intervention effect. The main feature of the cross-over design is that all participants receive both the experiment and control conditions. However, as mentioned in Section 1, for irreversible intervention effects the treatment must always be applied subsequent to the control condition when both are applied to the same participant. Thus the cross-over design will confound the treatment effect with other changes that occur over the course of time through the study. For example, it is quite possible that there will be improvements in the results returned by the search engine over the course of the study. These improvements should not be attributed to the instructional intervention. Also, participants themselves may naturally become faster through the course of the study simply from practice or increased familiarity with the interface. We will refer collectively to such auxiliary changes over time as "longitudinal effects".

Note that longitudinal effects are not an issue in the case/ control two group design. In that design, the control group and treatment group are disjoint so two groups can be measured simultaneously.

In summary, the cross-over design is inadmissible for studying an irreversible intervention since it confounds the treatment effect with longitudinal effects. On the other hand, the case/control two group design is an option, but can suffer from variability amongst the study participants.

3.1 Solution

To deal with the confounding longitudinal effects, we propose a *four group* cross-over design as illustrated in Figure 1. We split the participants randomly into four groups to carry out two separate cross-over experiments. For the first two groups (A and B) we use a cross-over experiment that measures the effect of the instructional video, which as we have said will be naturally confounded with longitudinal effects since the treatment must always be applied after the control. With the second two groups (C and D) we carry out the very same cross-over experiment, except that we do not show the instructional video. Thus, this experiment measures only the longitudinal effects that are confounded with our treatment effect in the first cross-over experiment. By accounting for longitudinal effects learned from the second experiment, we can extract a pure estimate of the treatment effect from the first experiment.

It is not obvious that this design should be better than the case/control two group design. Although we are reducing

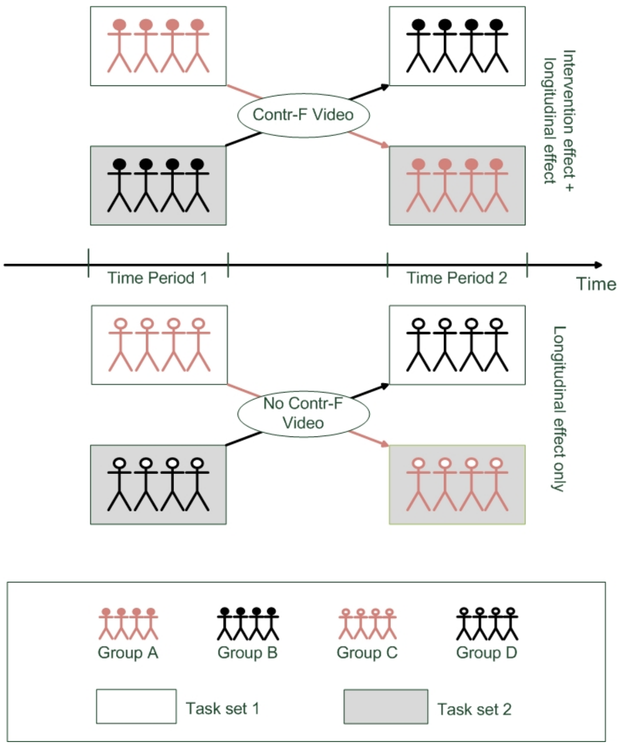

Fig. 1. Illustration of our 4 group design. For groups A and B (top) we show the “control-F” video between time periods 1 and 2 and cross over the participants to the opposite tasks. For groups C and D (bottom) we also cross over the participants to the opposite tasks from time period 1 to 2 but we do not show the “control-F” video.

the impact of participant variation through the use of the cross-over, we have also “wasted” a portion of our data for estimating the longitudinal effects, which would not be present at all in the case/control two group design. In our analysis we will compare our four group cross-over design with the case/control two group design and give guidance as to which design is better under what conditions. We note that the data for our study was collected using our four group cross-over design. The data collection and the details of the experiment are described in the following sections.

Task number	Task description
#1	Where is the Altar stone at Stonehenge believed to have come from?
#2	In the Tin Tin comics, the name of a character originally named Blumenstein was changed after the original publication. What was it changed to?
#3	Why did Captain Cook visit Tahiti?
#4	What food was popular in the Medieval Belgian Village at the 1964 World Fair according to Wikipedia?
#5	In 1984 the band named The Cure released an album entitled “The Top”. The lead singer played all the instruments on this album except one. What is that instrument?
#6	In the song “Day at the Races” by Jurassic 5, what are the next 3 lines following the line “Beatin up the block with the ghetto hop that knock and make you wanna crash the spot”?
#7	Which 2 French New World colonies did the British capture during the Seven Years War?
#8	When a dam is being built, something called a cofferdam is used to keep the work area dry. How many cofferdams were used to build the Hoover Dam? One, two or three?
#9	Kenneth Massey keeps a web page where he produces an overall ranking of college football teams based on combining many different rankings. On that web page, what is the current rank for Indiana?
#10	In Martin Luther King’s famous “I Have a Dream” speech, what are the 4 sentences immediately following “But we refuse to believe that the bank of justice is bankrupt. We refuse to believe that there are insufficient funds in the great vaults of opportunity of this nation. So we have come to cash this check—a check that will give us upon demand the riches of freedom and the security of justice. We have also come to this hallowed spot to remind America of the fierce urgency of now.”

TABLE I
THE 10 TASKS CHOSEN FOR THE “CONTROL-F” STUDY.

4. DATA

4.1 Data collection

The first step in carrying out our study was to select web search tasks. We chose the ten tasks listed in Table I. The tasks were chosen as tasks which we felt would benefit from knowledge of “control-F”. For all ten tasks, search engines are generally quite successful at returning relevant web pages; however, the answer is somewhat difficult to find within the text of the web page without using “control-F”. For our four group cross-over design in Figure 1, Tasks #1-#5 were used as the first set of tasks, and Tasks #6-#10 made up the second set.

The participants were instructed to use Google and, as in [3], to keep searching until they found the answer or 7 minutes had passed for each task. A timer was displayed on the screen, and the total task time was recorded once the participant pressed the “Finish searching” button.

In the case of applying the treatment, additional instructions were included to prompt the participants to watch the “Contr-F” video. Specifically, the participants were told the following:

“All 5 tasks ... require you to find information which might be buried in text on a web page. For that reason, it will be helpful if you learn how to use “control-F” to search for text within a web page. Here is a link to a video you should watch to help you become proficient with “control-F”: (Link to video is given.) Please watch this video before beginning the tasks.”

All tasks were completed by 50 raters for each treatment and participant group combination shown in Figure 1. For example, in the first time period Task #2 (which is in task set 1) was completed by 50 participants from group A and 50 participants from group C. In the second time period that task was completed by 50 participants from group B and 50 participants from group D. The four groups contained more than 50 participants so that when a given participant did not complete all five tasks for some reason, the additional participants could be assigned to these tasks.

4.2 Graphical summaries

Figure 2 presents the average log (base 10) task completion time for each of the 10 tasks for the four participant groups. The upper plot corresponds to participant groups A and B and thus shows the combined impact of the treatment effect and the longitudinal effects. Because the majority of the points fall beneath the 45 degree line, the plot suggests that the combination of the treatment and longitudinal effects results in an overall decrease in task time. The lower plot shows the data for participant groups C and D and thus isolates the impact of longitudinal effects. The points in this plot also fall beneath the 45 degree line but to a lesser degree than those in the first plot.

These observations suggest that our treatment effect of interest is in fact positive in the sense that it reduces task completion time. In the next section we will carry out model-based analysis to formally quantify the impact of the treatment effect on task completion time as well as examine its statistical significance.

5. Statistical Modeling and Analysis

5.1 The model

In this section we describe our statistical model for the data. The model uses the log (base 10) of the task time as the response. (The log transformation is commonly used to correct for the right-skew in time data.) We model this response as a function of the participant effects, the task effects, the longitudinal effects and the treatment ("control-F" video) effect. We model the participant effects as random effects and model the task, longitudinal and treatment effects as fixed effects. Formally, we write our model as

$$\begin{aligned}\log(Time)_{ij} = \mu &+ Participant_i + Task_j \\ &+ Longitudinal_{t(i,j)} + ContrF_{k(i,j)} \\ &+ (Task \times ContrF)_{jk(i,j)} \\ &+ (Task \times Longitudinal)_{jt(i,j)} + \varepsilon_{ij}\end{aligned} \tag{1}$$

where *Task*, *Longitudinal* and *ContrF* are the task, longitudinal and treatment fixed effects respectively. The random participant effects are represented as *Participant*. We assume that these random participant effects are independent and follow a normal distribution with mean zero and a common variance denoted by σ_P^2. The term ε represents errors that are not captured by the other terms in the model. These are also assumed to be independent and follow a normal distribution with mean zero and a common variance denoted by σ_e^2.

We also include some interaction terms in our model. The term $Task \times ContrF$ represents the task-treatment interaction. This interaction term is included because we believe some

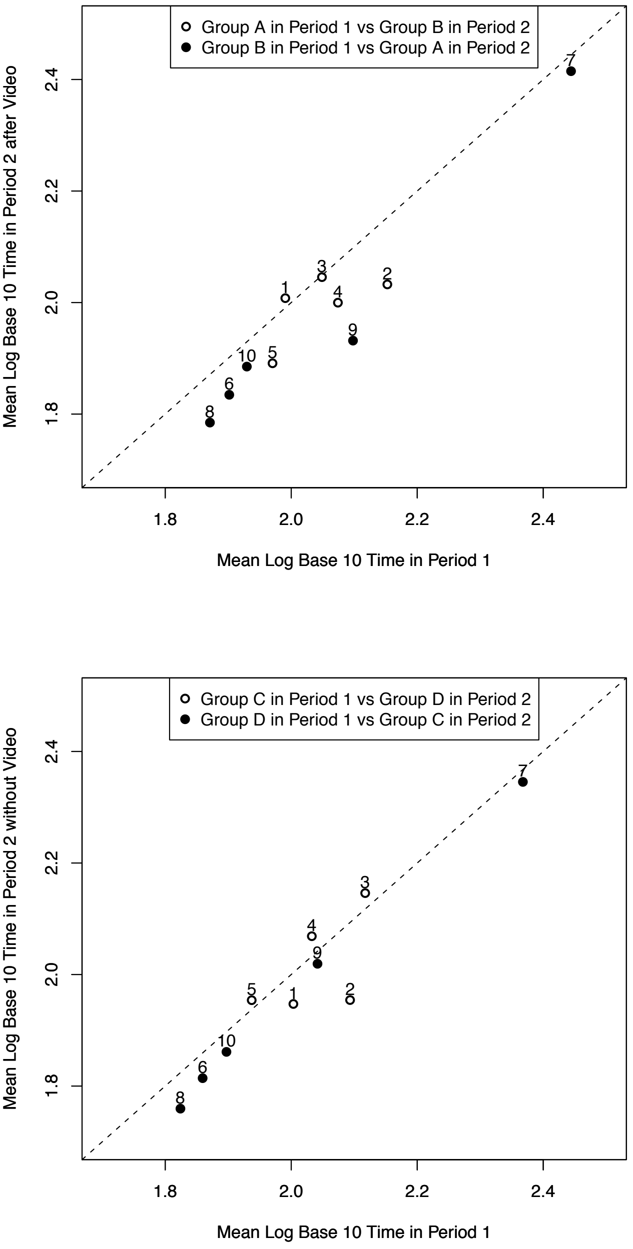

Fig. 2. Average log (base 10) task completion time by task. Upper graph: participant groups A and B. Lower graph: participant groups C and D. In both graphs the majority of the points fall beneath the 45 degree line, suggesting that the task times were faster in Period 2 than in Period 1. This is more pronounced in the upper graph (which measures the treatment effect in conjunction with the longitudinal effect) and less pronounced in the lower graph (which isolates the longitudinal effect).

tasks may benefit from knowledge of "control-F" more than others. We will discuss this interaction in more detail in Section 6. The term $Task \times Longitudinal$ in our model represents the task-longitudinal interaction. This interaction allows for some tasks to become faster to complete over time more than others. The main reason for including this interaction is that we believe Google's ranking may change more for the queries used in some tasks than in others.

We use i as the index for participants and j as the index for tasks. The index for treatment, $k(i, j)$, is either 0 or 1 depending on whether or not the treatment effect has been applied on Participant i when completing Task j. The time index $t(i, j)$ equals 1 or 2 depending on whether Participant i completed Task j in time period 1 or time period 2.

The model we have presented is over-parameterized unless we impose some constraints. To deal with this, we fix baseline values of zero for the effect of no "control-F" and also for the effect of the first time period. Also, we constrain the sum of the task effects to zero, and constrain the sum (over tasks) of the interaction terms to zero. Mathematically,

$$ContrF_0 = 0, \quad Longitudinal_1 = 0, \quad \sum_j Task_j = 0$$

and

$$\begin{aligned} \sum_j (Task \times ContrF)_{j0} &= 0, \\ \sum_j (Task \times ContrF)_{j1} &= 0, \\ \sum_j (Task \times Longitudinal)_{j1} &= 0, \\ \sum_j (Task \times Longitudinal)_{j2} &= 0. \end{aligned}$$

With these constraints, μ now represents the overall average log task time at time period 1 without the intervention effect. $Longitudinal_2$ represents the average longitudinal effect difference from time period 1 to time period 2, and $ContrF_1$ represents the average treatment ("control-F") effect, which is the key quantity of interest.

We note that our model is an example of a linear mixed effects model. More information about this class of models can be found in [7].

5.2 Results

The model is fitted to the collected data by a method called restricted maximum likelihood (REML) [7, Section 2.2] using the `lme4` package in the statistical language R. The estimates for the overall treatment and longitudinal effects are given in Table II. Note that these are still on the log scale. The 95% confidence intervals (CIs) are computed using the `mcmcsamp` function in the `lme4` package, which is deemed a more appropriate way of computing CIs for mixed effects models than using a normal approximation [1].

Parameter	Estimate (SE)	Approx. 95% CI
$ContrF_1$	-0.031 (0.021)	[-0.062, 0.016]
$Longitudinal_2$	-0.040 (0.016)	[-0.072, -0.010]

TABLE II
ESTIMATED OVERALL TREATMENT AND LONGITUDINAL EFFECTS ON LOG TASK-COMPLETION TIME.

In Table III we translate the estimated effects and their confidence intervals from the log scale into the original time scale. We see that the "control-F" instructional video overall reduced task-completion time by about 7%, although the effect was not quite significant at the 5% level since the confidence interval overlaps zero. In Section 7 we will compare the magnitude of the 7% figure to effect sizes from other studies on time until task completion for different treatments.

Table III also shows the estimate for the longitudinal effect. The participants are estimated to take about 9% less time to complete tasks in the second time period even without viewing the instructional video. This can be due to Google's results improving over time as well as the participants becoming more efficient. We will discuss this further in Section 8. Since this 9% longitudinal effect size is similar in magnitude to the "control-F" instructional video effect of 7%, we see the importance of using groups C and D in the four group cross-over design. If we had only used groups A and B (i.e., a standard cross-over design), we would overestimate the "control-F" treatment effect by more than a factor of 2 by confounding it with the longitudinal effect. This underscores why our four group cross-over design is a better choice than standard cross-over for irreversible treatments in the presence of non-negligible longitudinal effects; however, the question of whether or not the case/control two group design might still be preferable to our four group cross-over design has not been answered. We address that question in the following subsection.

Effect on time	Estimate	Approx. 95% CI
Control-F	-7%	[-13%, 4%]
Longitudinal	-9%	[-15%, -2%]

TABLE III
ESTIMATED OVERALL TREATMENT AND LONGITUDINAL EFFECTS ON TASK-COMPLETION TIME.

5.3 Comparison to the case/control two group design

In this section, we provide a heuristic comparison of our four group cross-over design to the case/control two group design in terms of the statistical precision for estimating the treatment effect.

Suppose a study (under either design) involves $4m$ participants and $2n$ tasks, and each participant will complete each task exactly once. This is slightly different from our "control-F" study in which some participants were not given all 10 tasks for practical reasons, but the general conclusions still apply.

Under the two group design, the experimenter would randomly split the $4m$ participants into a case group and a control group each consisting of $2m$ participants. After applying the intervention treatment to the treatment group (e.g. the "control-F" video), the two groups are to complete the $2n$ tasks during the same time period, with the task-completion times recorded. Because there is only one time period in this setting, model (1) simplifies to

$$\log(Time)_{ij} = \mu + Participant_i + Task_j + ContrF_{k(i)} + (Task \times ContrF)_{jk(i)} + \varepsilon_{ij}. \quad (2)$$

From this, we can derive that the variance of the (best linear unbiased) estimator for the treatment effect under the two group design is

$$\frac{\sigma_P^2}{m} + \frac{\sigma_e^2}{2mn}, \quad (3)$$

while the variance of the (best linear unbiased) estimator of the treatment effect under our four group cross-over design can be shown to be no more than

$$\frac{2\sigma_e^2}{mn}. \quad (4)$$

This variance upper bound is derived in the Appendix. It is based on replacing the estimator with a sub-optimal simplified version. The actual variance for the REML estimator is more complex and depends on the effect sizes of each of the individual tasks, and is not exactly that of the best linear unbiased estimator (BLUE). Nevertheless, using this bound is sufficient for demonstrating the main determinants of the relative attractiveness of the two designs.

Taking the ratio of (3) to (4) gives

$$1 + \frac{n}{2}\left(\frac{\sigma_P^2}{\sigma_e^2} - \frac{3}{2n}\right). \quad (5)$$

Thus, if

$$\sigma_P^2/\sigma_e^2 > 3/2n \quad (6)$$

the estimator from the four group cross-over design has a smaller variance than that from the two group design. In other words, if the participant variance σ_P^2 is large, our design is preferable. This is because our design reduces the impact of participant variation through the cross-over by having the same participants participate under both control (no "control-F") and treatment ("control-F") conditions.

From (5) we can also infer that the relative advantage of the four group cross-over design increases as the number of tasks ($2n$) included in the study increases. For studies involving moderate to large numbers of tasks, the improvement can be dramatic. More intuition as to why this is so can be gathered from reading Section 6.

For our "control-F" study data we can verify that (6) holds by noting that $m = 25$ and $n = 5$. The estimated variance value was 0.022 for σ_P^2 and for 0.059 for σ_e^2. Therefore, our estimated ratio for σ_P^2/σ_e^2 is approximately 0.37 which is larger than $3/2n = 0.30$. Thus, we conclude the four group cross-over design leads to better statistical precision for the overall treatment effect in our study.

To see the precise nature of this improvement for our current study, we can use the estimates of σ_P^2 and σ_e^2 (0.022 and 0.059, respectively) in (3) with $m = 25$ and $n = 5$. This gives an estimated variance under the two group design of about 0.0011, which corresponds to a standard error of $\sqrt{0.0011} = 0.033$. This is much larger than the standard error of our current estimate for the intervention effect under the four group cross-over design which is 0.021 (see Table II). We can use these standard errors to make the comparison in terms of confidence intervals (based on a normal distribution) on the original time scale. Figure 3 shows how much wider the confidence interval around the 7% treatment effect estimate would have been had we used the two group design instead of our four group cross-over design.

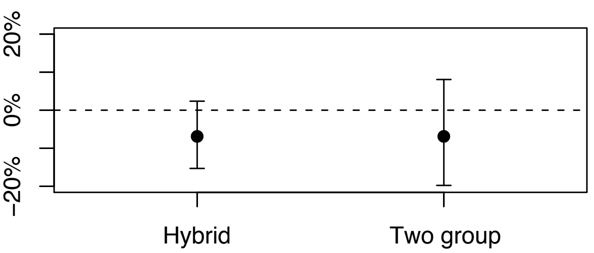

Fig. 3. Approximate 95% confidence intervals of the treatment effect for the two designs, assuming the estimated effect size is 7%.

Overall for our study, the four group cross-over design is preferable to the two group design. The reduction in the impact of participant variation due to the cross-over more than makes up for the need to estimate the longitudinal effects that are introduced.

6. Individual Task Analysis

In addition to the overall treatment effect *ContrF*, our model (1) also contains the interaction term $Task \times ContrF$ which allows for heterogeneous treatment effects over the different tasks as described in Section 5.1. Although these task-specific effects are generally secondary in importance to the overall treatment effect, they can be of interest when one wishes to study which tasks (or types of tasks) are more or less influenced by the treatment. For instance, perhaps one wishes to learn for which specific tasks to trigger certain search strategy suggestions. In a case like this, one would want to compare the task-specific effects to see which types of tasks are most likely to benefit from this suggestion. In this section we discuss this type of task-specific analysis. We carry out the analysis for our "control-F" study, and we compare the performance of our four group cross-over design to the two sample design with respect to this type of analysis.

Under the parametrization of the model (1), the task-specific treatment effect for Task j is

$$ContrF_1 + (Task \times ContrF)_{j1}.$$

Figure 4 gives the estimates of our ten task-specific treatment effects in terms of percentage increase in task-completion time along with the Bonferroni-corrected 95% confidence intervals. (Bonferroni correction [4, p.350] makes the confidence intervals wider to compensate for the fact that we are making multiple (ten) confidence intervals instead of just a single one.) We can see that the “control-F” instructional video had a stronger effect on Tasks 3, 4, 5 and 9 than on the others. The largest of these is Task 9. For this task the figure shows that participants take 25% less time on average to complete the task after controlling for the longitudinal effects. The fact that Task 9 had the largest effect is intuitive since it can only be answered by viewing a list of over 100 college football teams which are not in a fixed order. This task seems very difficult without knowledge of “control-F”, perhaps even more so than the other tasks.

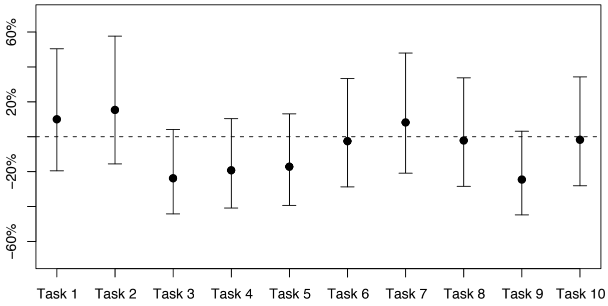

Fig. 4. Confidence intervals for the task-specific treatment effects.

Although our four group cross-over design proved to be preferable to the two groups design for estimating the overall impact of our treatment effect, the same is not true for analysis of the individual task effects. Rather, analysis based on our statistical model (1) and the results from our data shows that had we used the two group design, the confidence intervals for the individual tasks in Figure 4 would actually have been narrower. To get some intuition as to why this is true, consider the data from our four group cross-over design for only a single task and trace its path through Figure 1. For example, in time period 1 Task 1 is completed by groups A and C without seeing the video. In time period 2, it is completed by group B after watching the video and by group D without watching the video. So for Task 1, no cross-over has occurred under our four group design. In fact, no design can allow cross-over on individual tasks because a participant can complete each task no more than once. Thus, on the individual task level, the four group design spends a quarter of the data (i.e. groups C and D in the second time period) on estimating longitudinal effects, which would not have existed under the two-group design. In other words, the cross-over feature of our four group cross-over design only reduces the effect of participant variation in aggregate across all tasks, but not for any given individual task. Consequently, if the estimation of individual tasks effects is of primary interest, the case/control two group design is actually preferable. This should also help explain why the relative advantage of our four group cross-over design for the overall treatment effect increases as the number of tasks increases as noted in Section 5.3.

7. Comparison with Other Task Time Studies

Our “control-F” study has been used up to this point mainly to provide an example of an irreversible treatment effect. However, it is also interesting to compare the magnitude of the specific effect observed in our study to the magnitude of the task completion time effects in other studies. For instance, we can compare the effect of the presence or absence of the “control-F” video in our study to the effects of ranking and user interface changes. This helps us understand the relative value of showing the video to users. It is a nice feature of the task completion time metric that it facilities such comparisons among a diverse set of treatments.

We will use the experiments in [3] as a comparison point. The methodology used in that study is similar to the methodology in our study as noted earlier. The experiments in [3] measured the task time impact of removing the top 5 Google results as well as the impact of removing the abstracts from the original 10 Google results (but retaining the title and the URL). It was found that removing the top 5 Google results increased task completion time by 23.9% while removing only the abstracts increased task completion time by 3.5%. This suggests that our observed 7% impact from “control-F” falls in between these two. A graph showing all 3 effects with 95% confidence intervals is given in Figure 5. We note that because our study was conducted on a smaller scale than that in [3], the confidence interval for our study is much wider than those in [3]. Consequently, we cannot conclude with 95% confidence that the “control-F” impact is larger than the effect of removing abstracts, but the point estimates suggest that to be true.

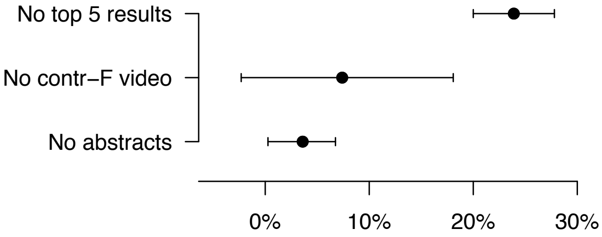

Fig. 5. Comparison of task completion time effect estimates for three different treatments.

Of course there are some methodological differences in these two studies as well as in the statistical models used, but overall the comparison is logical. Perhaps the largest difficulty in making the comparison is that our study deals with 10 tasks which were selected based on their potential benefit from “control-F”, while [3] deals with essentially random tasks. But, if we assume the effects observed on these random tasks in [3] would be similar on our 10 tasks, then the graph in Figure 5 is relevant for our 10 “control-F” tasks.

8. Discussion

In this paper we have demonstrated that our four group cross-over design can be advantageous for measuring the impact of irreversible treatments such as training interventions. In particular, this research design solves a major problem when studying the effects of interventions on populations that may change rapidly, or over corpora (such as the web) that change rapidly and dramatically. Without some compensation in the experimental design, the effects of treatments of interest can be obscured by the overall improvement in underlying data resources or improvements in the user-experience (e.g., when web search results improve).

Using the four group cross-over design we have observed that the magnitude of our specific "control-F" treatment effect appears to be fairly large, suggesting there is an opportunity to substantially improve the search experience of users through this type of educational intervention.

Generalizing to real users: While the "control-F" effect is real and measurable, it is difficult to know how the results of our "control-F" study will generalize to actual users outside of a laboratory environment. First, we do not know what fraction of users' tasks would benefit from knowledge of "control-F" due to the complexity of understanding web search tasks in general. The 10 tasks we chose for our study were selected because we believed that they could benefit from this knowledge, but we do not know how representative they are of naturally-occurring tasks.

There are other factors that complicate the generalization of the conclusions. For example, the participants are paid to read and follow instructions, so the treatment effect might be stronger for them than for actual users. But on the other hand, we have reason to believe the participants overall may have more web search expertise than average users, partly because inclusion in the study is based on self-selection. One particularly important consequence is that a higher proportion of the participants may have knowledge about "control-F" prior to the study, mitigating the effect of the instructional video. The impact of prior knowledge was not considered in our study for practical reasons.

Reducing longitudinal effects: While our four group cross-over design provides a means to control for potential longitudinal effects, ideally the experimenter should still try to minimize the longitudinal effects as much as possible. In our "control-F" study, the longitudinal effects arise from both the participants and search engine improving over the course of the study. One way to minimize the impact of participant improvement is to provide some initial tasks during a pre-period that are not actually included in the study. The idea here is that most of the participant improvement happens during the first few tasks, but after this the improvement is small. To minimize the longitudinal effect caused by the search engine improving, one should ensure that there is not a large gap between time periods 1 and 2. In our study, this was not possible due to practical reasons, but in other studies it may be feasible.

References

[1] R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, November 2008.

[2] G. Buchanan and F. Loizides. Investigating document triage on paper and electronic media. In *ECDL*, pages 416–427, 2007.

[3] R. Khan, D. Mease, and R. Patel. The impact of result abstracts on task completion time. In *WWW Workshop on Web Search Result Summarization and Presentation*, 2009.

[4] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer-Verlag New York Inc, 3rd edition, 2005.

[5] R. Mead. *The design of experiments*. Cambridge University Press, 1988.

[6] S. C. Peres, F. P. Tamborello, M. D. Fleetwood, P. Chung, and D. L. Paige-Smith. Keyboard shortcut usage: The roles of social factors and computer experience. In *Human Factors and Ergonomics Society*, pages 803–807, 2004.

[7] J. C. Pinheiro and D. M. Bates. *Mixed Effects Models in S and S-Plus*. Springer, April 2002.

[8] Y. Xu and D. Mease. Evaluating web search using task completion time. In *SIGIR*, pages 676–677, 2009.

Appendix

The variance of the treatment effect estimates: For the two group design, without loss of generality, assume that Participants 1 to $2m$ are put into the control group and Participants $2m+1$ to $4m$ consist the treatment group. Also, under the two group design, there is no longitudinal effect, so we drop the longitudinal subscripts as well as the terms involving *Longitudinal* from Model (1). The best linear unbiased estimator (BLUE) for the treatment effect

$$
\begin{aligned}
\widehat{ContrF_{TG}} &= \frac{1}{4mn}\Big(\sum_{i=1}^{2m}\sum_{j=1}^{2n} y_{ij11} - \sum_{i=2m+1}^{4m}\sum_{j=1}^{2n} y_{ij01}\Big) \\
&= ContrF_1 - ContrF_0 + \frac{1}{2m}\sum_{i=1}^{4m} P_i \\
&\quad + \frac{1}{4mn}\Big(\sum_{i=1}^{2m}\sum_{j=1}^{2n} \varepsilon_{ij11} - \sum_{i=2m+1}^{4m}\sum_{j=1}^{2n} \varepsilon_{ij01}\Big).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
Var(\widehat{ContrF_{TG}}) &= \frac{1}{4m^2}\cdot 4m\sigma_P^2 + \frac{1}{16m^2n^2}\cdot 8mn\sigma_e^2 \\
&= \frac{\sigma_P^2}{m} + \frac{\sigma_e^2}{2mn}.
\end{aligned}
$$

For the four group cross-over design, the exact form of the estimator is relatively difficult to write out due to the unbalanced nature of the design. However, one can easily calculate the variance of a simple (suboptimal linear) estimator

$$
\begin{aligned}
\widetilde{ContrF_H} &= \frac{1}{2mn}\Bigg(\Big(\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij12} - \sum_{i=1}^{m}\sum_{j=n+1}^{2n} y_{ij01} \\
&\quad - \sum_{i=m+1}^{2m}\sum_{j=1}^{n} y_{ij01} + \sum_{i=m+1}^{2m}\sum_{j=n+1}^{2n} y_{ij12}\Big) \\
&\quad - \Big(\sum_{i=2m+1}^{3m}\sum_{j=1}^{n} y_{ij02} - \sum_{i=2m+1}^{3m}\sum_{j=n+1}^{2n} y_{ij01} \\
&\quad - \sum_{i=3m+1}^{4m}\sum_{j=1}^{n} y_{ij01} + \sum_{i=3m+1}^{4m}\sum_{j=n+1}^{2n} y_{ij02}\Big)\Bigg) \\
&= ContrF_1 - ContrF_0 \\
&\quad + \frac{1}{2mn}\Bigg(\sum_{i=1}^{m}\sum_{j=1}^{n} \varepsilon_{ij12} - \sum_{i=1}^{m}\sum_{j=n+1}^{2n} \varepsilon_{ij01} \\
&\quad - \sum_{i=m+1}^{2m}\sum_{j=1}^{n} \varepsilon_{ij01} + \sum_{i=m+1}^{2m}\sum_{j=n+1}^{2n} \varepsilon_{ij12} \\
&\quad - \sum_{i=2m+1}^{3m}\sum_{j=1}^{n} \varepsilon_{ij02} + \sum_{i=2m+1}^{3m}\sum_{j=n+1}^{2n} \varepsilon_{ij01} \\
&\quad + \sum_{i=3m+1}^{4m}\sum_{j=1}^{n} \varepsilon_{ij01} - \sum_{i=3m+1}^{4m}\sum_{j=n+1}^{2n} \varepsilon_{ij02}\Bigg),
\end{aligned}
$$

and so the variance of the BLUE under the four group cross-over design,

$$
\begin{aligned}
Var(\widehat{ContrF_H}) &\le Var(\widetilde{ContrF_H}) \\
&= \frac{1}{4m^2n^2}\cdot 8mn\sigma_e^2 = \frac{2\sigma_e^2}{mn}.
\end{aligned}
$$