Secretary Search under the Microscope

Astrid Matthey and Ondřej Rydval*

February 13, 2009

Preliminary version, please do not quote!

Abstract

We analyze individual behavior in a secretary search problem. Our experimental design allows us to directly observe individual search strategies, rather than inferring them from stopping times as in earlier studies. The results suggest that subjects' search is i) too short on average, confirming previous findings, ii) very heterogeneous, both across and within subjects, and iii) influenced by perceived regularities in the sequences of applicants that subjects encounter. Overall, subjects deviate from the optimal strategy in many different ways, in some cases with substantial consequences for expected payoffs.

JEL classification: D01, D83, C91

Keywords: search, secretary problem, experiments

*Max-Planck-Institute of Economics, Jena. matthey@econ.mpg; rydval@econ.mpg.de
We wish to thank Neil J. Bearden, Werner Güth, and Štěpán Jurajda for helpful comments, and Chris Göring and Karolin Schröter for invaluable research assistance. The usual disclaimer applies.

1 Introduction

In many search situations - such as searching for mates, housing, parking space, or employees - the available options (applicants) appear sequentially, and decisions about them must be made without knowing their quality relative to the options yet to come. Once an option is rejected, the chance of it being available later on decreases, possibly to zero. Because of the prototypical example, the literature refers to this kind of search situation as a *secretary search* task.

We experimentally study such a secretary search task, the design of which is similar to Bearden et al. (2006). Our subjects sequentially inspect 40 randomly ordered applicants, each of which must be either accepted or rejected, with no possibility of recalling past rejected applicants. The applicants are ranked in terms of their quality, but when an applicant appears, subjects only observe its rank relative to all past applicants. The search stops as soon as an applicant is accepted. Only if the accepted applicant is among the best three out of the 40, it yields a positive payoff. The task is repeated over 24 rounds featuring various sequences of applicants (applicant streams).

The main innovation of our design is to elicit search strategies explicitly. At the beginning of each search period, subjects state the worst relative rank that they are willing to accept for this period's applicant. If the applicant then turns out to have a better or equal relative rank, it is accepted and the search stops; otherwise the search continues. A similar willingness-to-accept procedure was used in a very different search task by Sonnemans (1998). The procedure yields more explicit data on search behavior and thereby helps to extend the knowledge gained in earlier studies, where search strategies were inferred solely from subjects' acceptance (stopping) decisions (see Bearden et al., 2006, and Zwick et al., 2003, for overviews).

Another design innovation pertains to the set of applicant streams that our subjects encounter. In previous studies, applicant streams were drawn randomly. We study whether applicant stream characteristics affect search behavior, and hence whether any conclusions regarding subjects' strategies should be viewed conditional on particular applicant streams used. We design two treatments differing in certain characteristics of the applicant streams that we expect to affect the length of search. Our applicant stream design further deals with the bias caused by non-random attrition (dynamic sample selection) of subjects depending on their search strategies: We design several special applicant streams that allow us to observe a large number of search periods for a majority of subjects regardless of their search strategy.

The results extend our understanding of search behavior in several respects. First, the willingness-to-accept data provides much more direct evidence on whether subjects actually use strategies which are similar in their structure to the theoretically optimal multi-threshold rule (MTR). Multi-threshold rules reject any applicant for the first several periods, then after a certain period (threshold 1) accept any applicant with a relative rank of 1, then after some more periods (threshold 2) accept any applicant with a relative rank of 1 or 2, and similarly for higher-order MTR thresholds. There are as many such thresholds as there are positive-payoff applicants. We are able to directly observe MTR thresholds in each round, which allows us to assess their between- and within-subject heterogeneity. Previous studies could derive individual MTR thresholds only indirectly from stopping decisions,

and had to assume subjects use MTR-like search strategies that are stable over rounds.

The majority of our subjects appear to use MTR-like search strategies (at least in later rounds), but their thresholds are on average lower than the risk-neutral optimal ones. These general findings are in line with those of most previous studies (see Seale and Rapoport, 1997, 2000 for early examples). However, we observe sizeable between-subject heterogeneity, with some subjects' thresholds being much lower and others' thresholds higher than optimal. Furthermore, only few subjects have stable strategies over most rounds, while a majority of subjects adjust their thresholds substantially, some even in successive rounds. We show, however, that the cost of deviating from the optimal strategy even quite substantially - though still in an MTR manner - has rather small expected payoff consequences.

A minority of our subjects clearly do not use MTR-like strategies. Their willingness-to-accept sequences are non-monotonic, discontinuous, or include relative ranks higher than the number of positive payoffs. All of these behaviors are irrational in theory, but are permitted by our elicitation procedure that does not enforce MTR-like strategies. Neither of the behaviors could have been documented by previous studies by nature of their design, but we show that they frequently involve non-negligible expected payoff cost. We attempt to deal with them in a way that takes into account the stochastic nature of search behavior and preserves as much of subjects' original intentions as possible.

Finally, we observe that our treatment design affects the length of search in the expected direction: subjects who encounter more applicant streams that reward "longer" search tend to have higher MTR thresholds compared to subjects who encounter more applicant streams that reward "shorter" search. Subjects thus appear to adjust their search strategies to the perceived regularities of the applicant streams on average. However, the between-subjects pattern of the adjustment is very heterogeneous. Having encountered good applicants in early periods in past rounds, some (most) subjects apparently decide to search "shorter," while others speculate that this trend will be reversed in future rounds and consequently decide to search "longer." These findings overall support our conjecture that any conclusions about the nature of search behavior should be conditioned on the set of applicant streams that subjects encounter, though uncovering a more generally applicable set of relevant applicant stream characteristics remains a topic of future research.

The remainder of the paper is organized as follows. Section 2 explains the design. The data is described in section 3.1 and analyzed in section 3.2. Section 4 concludes.

2 Design

2.1 The generalized secretary search problem

The Generalized Secretary Problem (Bearden et al, 2006) can be described by the following characteristics:

1. There are n applicants for a single position who can be ranked (with no ties) in terms of their quality, where n is known to the decision maker (DM).

2. The DM interviews the applicants sequentially in a random order (all $n!$ orderings are equally likely).

3. For each interviewed applicant, the DM only observes its relative rank, r, and has to decide whether to accept or reject the applicant.

4. Once rejected, an applicant cannot be recalled. The n^{th} applicant is accepted if reached.

5. The DM's payoff for selecting an applicant with an absolute rank of a is $\pi(a)$, where $\pi(1) \geq ... \geq \pi(n)$.

This problem has many extensions such as positive search cost and possibility of recalling past applicants, which we do not discuss.

The optimal policy takes the form of a multi-threshold rule (MTR): one should not accept any applicant for the first several interviews/periods, then comes a period (i.e., threshold 1) from which onwards one should accept an applicant with $r = 1$, then comes another period (i.e., threshold 2) from which onwards one should accept an applicant with $r \leq 2$, and similarly for higher-order thresholds. There are A such thresholds where A is the number of positive-payoff applicants with $\pi(a) > 0$. The size of the optimal MTR thresholds depends on n and $\pi(a)$. (The optimal thresholds of course also vary with risk aversion, but surprisingly little given a sensible utility function, which is probably why the literature always refers only to the risk-neutral optimal thresholds.)

Using various parameterizations and frames, secretary search experiments elicit stopping decisions (stopping times) over many search trials/rounds. From the stopping decisions they usually attempt to derive individual MTR thresholds and compare them to the optimal thresholds (assuming MTR-like behavior stable over rounds, and using restricted grid search or a similar procedure to derive individually best-fitting thresholds); or to compare the overall goodness of fit of the MTR and various non-MTR search rules/heuristics (again permitting between-subject but not within-subjects heterogeneity in the estimated parameters). In most parameterizations of the task, subjects apparently stop searching too early - especially higher-order individual MTR thresholds are much lower than the optimal ones - and MTR seems to fit better than non-MTR search rules.

To give a more concrete picture, in a recent study that we draw upon, Bearden, Rapoport and Murphy (Management Science, 2006) conduct two secretary search experiments framed as an apartment search, with the above described setup and the following parameters: n=60 (n=40); A=6 (A=3) with $\pi(a)$ decreasing geometrically (linearly); 60 rounds featuring different streams of applicants picked randomly from the set of $n!$ streams; 2 (1) randomly selected rounds paid out; and 62 (30) subjects. Assuming stable behavior over rounds, the authors estimate individual MTR thresholds as well as thresholds for two non-MTR search rules described below. The restricted grid search estimation procedure minimizes the number of rounds where a search rule wrongly predicts the stopping time, which actually yields non-uniqueness in the MTR estimates for about half of the subjects (for whom the authors assign estimates closest to the optimal policy).

The non-MTR search rules, the form of which is quite typical for this literature, are as follows. The Horse Race Decision Rule (HRR) assumes that one has A

separate thresholds for the number of occurrences of each $r \leq A$, and that one stops searching and accepts the current applicant as soon as any of the thresholds is reached. The Successive Undesirable Applicant Decision Rule (SUAR) assumes that one has A separate thresholds for the minimum number of successive undesirable applicants (i.e. applicants with $r > A$) passed before any given $r \leq A$ is chosen, and that one stops searching and accepts the current applicant as soon as any of the thresholds is reached.

For both HRR and SUAR, when a subject reaches her threshold for, say, $r = 3$ (and hence stops searching), if she were instead offered an applicant with $r < 3$, she would not have preferred it and hence would not have stopped searching (unless reaching the threshold for this $r < 3$). In this sense, the HRR and SUAR search rules seem incentive incompatible. Both of them are generalizations of two central search rules studied in earlier secretary search experiments with a single positive payoff ($A = 1$), where subjects report in post-experimental questionnaires that their search was affected by the average rate of candidate arrival (AROCA) and the number of periods since the last candidate arrived (PSLC), where a candidate is an applicant with $r = 1$ (Seale and Rapoport, 1997, 2000). While search rules based on AROCA or PSLC are clearly incentive compatible for search tasks with $A = 1$, their HRR and SUAR generalizations for search tasks with $A > 1$ seem less sensible. Still, something like AROCA and PSLC could affect search behavior even when $A > 1$, for instance if one re-defines a candidate as any applicant with $r \leq A$. We return to this issue when describing the streams of applicants used in our experiment.

2.2 Our design and elicitation procedure

We draw on the second experiment of Bearden at al. (2006), with $n = 40$ and $A = 3$, and prizes of 15, 10 and 5 EUR (their prizes were \$12, \$7 and \$2). We have 21 rounds (plus 3 initial warm up rounds without payoffs), two of which are drawn randomly (by subjects themselves and without replacement) and paid out. The optimal risk-neutral MTR thresholds for our case are periods 14, 28 and 35, which is generally slightly lower than Bearden et al.'s (2006) thresholds that you can see in the first row of their Table 3 reproduced below. The remaining rows show the estimated thresholds (averaged across 30 subjects) for the MTR, HRR and SUAR search rules. The last column shows the number of rounds (averaged across subjects) for which the individually best-fitting thresholds wrongly predict the stopping time (out of the total of 60 rounds). The only substantial difference

Table 3 **Optimal (r_x^*) and Mean Observed Estimated Threshold Values (r_x) Across Subjects and Trials for Three Different Decision Rules for Experiment 2**

x	1	2	3	Mean violations
r_x^*	14	29	37	—
Mean r_x: MTR	13	22	30	12
Mean r_x: HRR	4	3	2	44
Mean r_x: SUAR	5	7	6	49

in our design is that our subjects do not choose in every period whether to accept

or reject the current applicant after observing its relative rank, r. We instead ask subjects to state at the *beginning* of each period - i.e., *before* the current applicant appears - which worse/highest relative rank, h, they would be willing to accept. If the current applicant's r turns out lower than or equal to h, the applicant is selected; otherwise the task continues with the next period (subjects can anytime enter $h = 0$ to make sure that the task continues). Obviously, a deterministic subject following the optimal risk-neutral MTR policy would switch from $h = 0$ to $h = 1$ in period 14 (threshold 1), then (if not stopping earlier) to $h = 2$ in period 28 (threshold 2), and then (if not stopping earlier) to $h = 3$ in period 35 (threshold 3). Thus our h-setting, willingness-to-accept elicitation procedure is in the spirit of an MTR search rule, and one could argue that it undesirably encourages or induces MTR-like behavior as opposed to non-MTR rules, but see more on this below. A version of this elicitation procedure is used by Sonnemans (JEBO 1998), but in a different search task: with unlimited number of alternatives, positive search cost and the possibility of recalling past applicants, and with subjects setting h as part of their fixed strategy for the whole round (i.e., "In this round I want to stop only if ..").

Our main motivation for using this elicitation procedure was to look more closely at search behavior. Given the typically limited sample sizes and number of trials, it seems rather problematic to use solely stopping times to reliably derive MTR thresholds and to assess whether subjects actually use MTR search rules (in a deterministic or stochastic sense). For similar reasons, previous studies also cannot reveal much about heterogeneity in search behavior in cross section, and especially over time, given that their derivation of MTR and non-MTR thresholds is based on assuming that search behavior is stable over rounds, which we do not find (and nor do Bearden et al., 2006, when comparing stopping times in early and late rounds of their experiments). Our elicitation procedure should permit a closer look at how people deviate from the optimal search rule, and in turn how costly various types of deviations are. To preview, we confirm the finding of previous studies that subjects are early searchers (on average).

Another strand of motivation arises from our concern with previous studies drawing applicant streams *randomly* from the set of $n!$ available streams. What if applicant stream characteristics considerably affect search behavior (of some subjects) and its adjustment over time? If so, which applicant stream characteristics matter, and should any conclusions regarding subjects searching too short (as is usually claimed) be viewed conditional on particular applicant streams used? We attempt to shed some initial light on this issue by designing two treatments differing in certain characteristics of our applicant streams which we hypothesize might affect the length of search. Furthermore, because any comparison of search behavior is undesirably affected by non-random attrition of subjects depending on their search rules, we design special applicant streams that allow us to observe a large number of search periods for a majority of subjects pretty much regardless of their search rule.

We anticipate criticism along the lines that our h-setting elicitation procedure encourages or induces MTR-like behavior, as opposed to non-MTR search rules such as those mentioned above. However, the procedure actually gives subjects quite a lot of leeway for using non-MTR search rules. We do not advise or force subjects to enter h in a monotonic manner and subjects also do not observe their previous h entries (in any given round) when entering their current h. Also, subjects are free

to enter h in a non-continuous (jumping) fashion, i.e., skipping some intermediate h values (i.e., $h = 1$ or $h = 2$ or both). Subjects can also enter $h > 3$ anytime; although such behavior cannot be justified on payoff maximizing grounds, subjects may have non-monetary "aspirations". For instance, when approaching n and realizing that the positive-payoff applicants have most likely been passed, subjects may simply wish to *choose* an applicant themselves even if this implies a zero payoff, rather than being assigned a random applicant in period 40.

We in fact observe non-monotonic, jumping and h¿3 behavior in our sample. It is difficult to judge whether these kinds of behavior are a consequence of our elicitation procedure since they could not have been detected in previous studies focusing solely on stopping times. We could in principle run treatments testing for the influence of our elicitation procedure on search behavior, such as showing subjects their past h entries (for the current round), or reminding or forcing subjects to be monotonic, non-jumping and enter only $h \leq 3$, but we are not persuaded that such reminders or enforcement are generally desirable. In fact, the post-session questionnaire responses of the concerned subjects indicate that they made their decisions consciously and did not feel restricted by our elicitation procedure.

Our experimental instructions, framed as a stack-of-cards search problem, are included in the appendix (translation from German). The instructions also contain an understanding test; more on this in the results section. We used a sequence of video pilots to fine-tune the instructions: pairs of subjects in a video cabin first discussed the instructions and the task and then jointly made decisions. Watching the videos proved invaluable for detecting potential misunderstanding of the instructions and the actual computerized experiment. One thing that has become apparent is that the secretary search problem is a complex task and requires the use of many examples in the instructions.

The videos also helped us choose a sensible parameterization of the task: to cut a long story short, a sensible combination of n, A, $\pi(a)$ and the number of payoff-relevant rounds, for which the video subjects' first reaction was *not* that the game was just a "Glueckspiel" gamble with virtually hypothetical earnings (as may have been the case in previous experiments with $A = 1$, higher n than ours, and a lower fraction of payoff-relevant rounds). Furthermore, we used the videos to get insights into which applicant stream characteristics our subjects attend to, both during search and when inspecting feedback, which we then used as guidance for designing our applicant streams.

All experimental sessions were implemented using z-tree (Fischbacher, 2007)

2.3 Design of applicant streams in the two treatments

As outlined above, one of our motivations is to get an insight into how and which applicant stream characteristics affect search behavior, both during search in a given round and also in subsequent rounds through applicant stream feedback (at the end of each round, subjects observe the whole stream of absolute ranks). Our basic idea is to design two treatments differing in potentially important characteristics of applicant streams which might affect length of search. Yet identifying such characteristics is a tough challenge. The earlier literature (e.g., Seale and Rapoport, 2000) suggests that, judged from subjects' stopping times, something like AROCA and PSLC affects the way people search during a given round (see

the previous section), and that subjects adjust their search behavior over time by (mostly) searching slightly longer. However, given that stopping times reveal rather little about actual search rules and their heterogeneity across subjects and over time, it is hard to use the past results to derive relevant applicant stream characteristics. Moreover, universally relevant characteristics most likely do not even exist given the heterogeneity of search behavior that we document below.

Our video pilots as well as post-session questionnaire responses from the actual experiment confirm our doubts. There is considerable variation in what subjects attend to during search, though if anything, subjects frequently do seem to attend to some form of PSLC or AROCA, with predominant attention to applicants with $r = 1$, and some report to use it when adjusting their MTR thresholds in an *ad hoc* way. We also observe considerable variation in the use and interpretation of feedback: some subjects attend to the position of some or all positive-payoff applicants, but others look only at which positive-payoff applicants they passed or which they did not reach, and others apparently ignore feedback completely; as one would expect, attention to feedback also seems to depend on whether or not subjects earn a positive payoff in a given round.

Based on the video pilots, we have taken a sort of pragmatic approach to selecting applicant stream characteristics for the experiment. Implicitly assuming that subjects use MTR-like search rules, we design two types of applicant streams for which it pays off to search (in an MTR manner) either "long" (meaning that one's MTR thresholds are positioned in line with the optimal policy or slightly later) or "short" (meaning that one's MTR thresholds are positioned in earlier periods compared to the optimal policy). We then design a Long (Short) treatment that contains more rounds with Long (Short) applicant streams, nested among two other applicant stream types described below. This of course means that applicant streams are not generated entirely randomly, but as will be clear below, there is a considerable random element in the generation of a particular applicant stream faced by a particular subject in a particular round.

Rather than the above described payoff-reinforcement perspective, one may of course view our treatment design from a feedback perspective, and perhaps subjects do as well: As will be clear below, our treatment design implies how frequently each positive-payoff applicant on average appears in a particular segment of the applicant streams. As a result, subjects may (think they) learn something about such regularities and act upon them in their subsequent search, which might affect their length of search. In other words, even if we observe a treatment effect, we do not know whether it works through payoff reinforcement ("I adjust my thresholds up or down according to what most frequently paid off in the past") or learning about applicant stream regularities ("I adjust my thresholds according to where I most frequently observed positive-payoff applicants on previous rounds' feedback screens"). We should mention that some subjects actually report interpreting feedback (i.e., the position of positive-payoff applicants) in a way that most likely affects behavior in the opposite direction than payoff reinforcement would: they expect that "If good applicants recently appeared in later periods, they will most likely appear earlier next time." This further underscores the uncertainty as to which applicant stream characteristics are potentially relevant and in what sense.

As a short de-tour (this paragraph can be skipped without any loss of information), we had actually started this project with the objective of studying which parts of feedback information people mostly use (rationally or irrationally) to adjust their

search behavior over time (rounds). For instance, we implemented a video pilot with two treatments where applicant streams were identical across treatments up to any stopping point, but what differed across the treatments (conditional on stopping in a particular period) was the position of the positive-payoff applicants (if any) in the unreached remainder of the applicant stream. We wished to study whether, ceteris paribus, people on average react differently when observing that they could have done better had they searched just a little bit longer in one treatment, as opposed to having to search much longer in the other treatment. In another video pilot, we varied the amount of applicant stream feedback by having a round where subjects did not get to see the unreached remainder of the applicant stream, and then another round where subject could choose whether to see the unreached remainder. This was in order to assess whether subjects on average use the information contained therein to adjust their search behavior (for credibility reasons, subjects could verify the complete applicant streams ex-post). What these pilots showed is that the way people react to feedback is simply very heterogeneous.

We now return to our Short and Long treatments and describe the applicant stream types that subjects face. As shown in figure 1 below, we have three sub-types of **Long applicant streams** and three sub-types of **Short applicant streams**. The first column shows the sub-type labels while the next five columns show the criteria for generating each of the sub-types, in terms of the positions of the five applicants with absolute ranks 1-5 (i.e., not only the three positive-payoff applicants, for reasons explained beneath the table). The last column lists the number of applicant streams of each sub-type that each subject faces in the Long and Short treatments. The notes beneath the table briefly explain the logic of each sub-type and contain additional remarks. You will notice that searching short (long) makes likely but does not guarantee ending up with a positive payoff in rounds with Short (Long) applicant streams, and vice versa for ending up with a zero payoff. For instance, the Long1 and Long3 sub-types permit positive payoffs even for subjects with very short and moderately short search, respectively, but this is inevitable given the succession of relative ranks we wished to implement, and note that following the optimal policy in both cases yields an even higher payoff. Further, positive-payoff applicants in Short applicant streams come early, but not too early in order not to encourage very short search. We did not wish to make the distinction between Long and Short applicant streams too extreme, given how little we know about subjects' search rules. In retrospect, the extent of search heterogeneity (including non-monotonic and jumping behavior) that we observe in the experiment suggests that this was a sensible decision.

Figure 1 shows that besides Long and Short applicant streams, subjects also face **Neutral applicant streams**. Neutral streams make earning a positive payoff quite probable regardless of one's search rule (MTR thresholds). These streams serve the purpose of avoiding that subjects get too frustrated when, for example, an inherently short searching subject ends up in the Long treatment and makes little or no money when facing frequent Long streams. The two sub-types of Neutral streams also serve the purpose of "filling the gaps" in the sense that, taking all stream types together, any of the three positive-payoff applicants can in principle occur in any period 4-39. Each subject faces the same number of the first and second Neutral sub-types because these may slightly encourage relatively longer and shorter search, respectively.

Sub-type of applicant stream	Periods of occurrence of abolute rank #					Number of streams in Long/Short treatment
	#1	#2	#3	#4	#5	
Long 1	29-39	29-39	2-7	--	--	2/1
Long 2	29-39	29-39	29-39	1-10	--	2/1
Long 3	29-39	29-39	22-28	after #1 & #2	1-10	3/1
Short 1	4-11	17-25	26-33	--	--	1/2
Short 2	4-11 after #2	4-11	23-32	before #2 & #1	after #3	1/3
Short 3	4-11 after #3	23-32	4-11	--	--	1/2
Neutral1	18-28	8-17	8-17 after #2	--	--	4/4
Neutral2	8-17	29-35	18-28	1-7	--	4/4
Assessment1	2-7	37-40	2-7 after #4	2-7 after #1	--	0-2/0-2
Assessment2	2-7	37-40	2-7 after #5	37-40	2-7 after #1	0-2/0-2
Assessment3	37-40	2-7	2-7 after #2	37-40	2-7 after #3	0-2/0-2
Assessment4	37-40	2-7	37-40	2-7 after #2	2-7 after #4	0-2/0-2
Assessment5	37-40	37-40	2-7	2-7 after #5	2-7 after #3	0-2/0-2
Assessment6	37-40	37-40	2-7	2-7 after #3	37-40	0-2/0-2

Sub-type of applicant stream	Explanation of sub-types and additional notes
Long 1	after #3, j>1 till at least p.29; optimal search earns #1 or #2, short search likely earns nothing
Long 2	after #4, j>1 till at least p.29; optimal search earns #1 or #2 or #3, short search likely earns nothing
Long 3	after #5, j>1 till at least p.22; optimal search earns #1 or #2, short search likely earns #3 or nothing
Short 1	short search likely earns something, optimal or longer search earns nothing
Short 2	#4 & #5 can swap; short search likely earns something, optimal or longer search earns nothing
Short 3	short search likely earns something, optimal or longer search earns nothing
Neutral1	after #2, j>1 at least till p.18; short search can earn #2 or #3 or #1, optimal search can earn #2 or #1
Neutral2	after #1, j>1; short search can earn #1 or #3 or #2, optimal search can earn #1 or #2
Assessment1	after #1 & #4 & #3, j>3; unless very short search or h>3, short and long search earn #2
Assessment2	after #1 & #5 & #3, j>3; unless very short search or h>3, short and long search earn #2 or nothing
Assessment3	after #2 & #3 & #5, j>3; unless very short search or h>3, short and long search earn #1 or nothing
Assessment4	after #2 & #4 & #5, j>3; unless very short search or h>3, short and long search earn #1 or #3
Assessment5	after #3 & #5 & #4, j>3; unless very short search or h>3, short and long search earn #1 or #2
Assessment6	#6 in p.2-7 after #4; after #3 & #4, j>2; short and long search can earn #1 or #2 or nothing

Figure 1: Applicant stream characteristics

The last type of applicant streams that subjects face are **Assessment applicant streams**. These serve the purpose of providing us with as much search data as possible (i.e., as many search periods observed before stopping) for a majority of subjects regardless of their search rule. This is unlike the other types of applicant streams, especially Short and Long streams, which can be expected to generate non-random attrition of subjects in earlier search periods, with the attrition depending on particular search rules that subjects use (i.e., dynamic sample selection on unobservables). This was problematic in previous studies which generated applicant streams randomly and which focused solely on stopping times. As a consequence, various types of subjects (search rules) were likely over- or under-represented in calculating the various MTR and non-MTR thresholds.

Without having Assessment streams, the dynamic sample selection would also affect our own comparison of behavior across the Long and Short treatments, even if both treatments were homogenous with respect to the distribution of subjects' search rules (MTR thresholds). A simplistic way to think about this is that (abusing notation) each subject has an initial (first-period) h_i equal to "personal fixed-effect + Beta differing across treatments"; then even with the upward adjustment of h_i over periods being identical for all subjects, and with the h_i distribution being identical across treatments, one still gets (ceteris paribus) sample selection since more subjects drop out in any given period in the treatment with higher Beta.

In this experiment, rather than ex post dealing with the dynamic sample selection

econometrically by estimating a proper duration model and focusing on predicted search behavior (which would likely require a lot more data), we mainly focus on search behavior in the rounds where subjects face the Assessment streams (i.e., Assessment rounds). The Assessment streams contain sequences of relative ranks ensuring that, under normal circumstances, almost any subject (search rule) searches till at least period 37 out of 40. This may be frustrating for subjects who, for the other types of applicant streams, are generally not used to searching so long, but Assessment streams frequently (though not always) compensate for this by allowing subjects to eventually earn a positive payoff. Other than making subjects search long, Assessment streams are not entirely unusual or improbable and in fact share various general characteristics with the other stream types (see below). Nevertheless, to minimize the potential cost associated with any given Assessment sub-type generating an unexpected behavioral reaction and undesirably affecting subsequent search behavior, we generate six different Assessment sub-types and spread them across subjects; furthermore, each subject faces two of the first triplet and two of the second triplet of sub-types because the sub-types within each triplet are similar (the first (second) triplet resembles Short (Long) applicant streams to some extent).

Figure 2 below shows how the four types of applicant streams are ordered in the Short and Long treatments. As mentioned above, we have 3 warm up rounds (which contain the same stream types in the same order) and 21 payoff-relevant rounds (the number of which may be insufficient for observing "asymptotic" behavior but was adjusted based on the observed attention span of our video subjects). There are of course tradeoffs associated with any stream type ordering. Our requirements were that (a) there are pairs of Assessment rounds situated at an early, intermediate and late stage of the experiment, to permit comparing behavior in both adjacent and more distant rounds; (b) the intermediate and late Assessment rounds are each preceded by a sequence of Short and Long rounds (applicant streams) which differ across treatments in their composition and location (but intentionally not excessively; also, within a treatment, the intermediate and late sequences of Short and Long rounds are intentionally interspersed with Neutral rounds in a different manner); and (c) the pairs of Assessment rounds are preceded by at least one Neutral round to synchronize stream types across treatments prior to the Assessment rounds. In addition, Assessment streams A1 in rounds 3 and 11, and likewise Assessment streams A2 in rounds 4 and 20, are identical, in order to permit a cleaner within-subjects comparison (but repeating both A1 and A2 sufficiently apart to limit the possibility that subjects recognize past applicant streams). Our randomization of applicant streams across subjects

	Round																						
Treatment	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Short	N	**L**	**S**	N	N	A1	A2	**S**	N	**L**	**S**	**S**	N	A1	A	N	**S**	**L**	**S**	**S**	N	A	A2	N
Long	N	**L**	**S**	N	N	A1	A2	**L**	N	**S**	**L**	**L**	N	A1	A	N	**L**	**S**	**L**	**L**	N	A	A2	N

Figure 2: Sequence of applicant streams

and rounds is only partial: there are six groups of 10-12 subjects (three groups in each treatment) within which subjects face exactly the same set of streams. Across groups, we vary Long and Short stream sub-types across rounds in a balanced way, to increase the chances that any treatment effect we observe is due to our general treatment design rather than using particular sub-types and their succession. (As

indicated in the last column of figure 2, each group faces a given Short or Long sub-type 1-3 times depending on the treatment.) We then generate several Short and Long applicant streams for each sub-type and use any one stream in only one group, to limit the impact of a particular Short or Long stream (as opposed to the impact of a stream sub-type). We also vary across groups the succession of Neutral and Assessment sub-types and generate several applicant streams for each sub-type. For comparison purposes, however, the first, second and third group from each treatment, respectively, face identical Assessment streams in identical rounds.

In addition, certain characteristics are standardized for all types of applicant streams, mainly to ensure that no applicant stream is too improbable in certain respects. The first criterion is purely pragmatic: we make sure that the three positive-payoff applicants (and sometimes also the best two zero-payoff applicants) are non-adjacent and non-equidistant. The second criterion is motivated by our and others' previous observation that subjects pay attention to various forms of PSLC and AROCA. There are many ways one could standardize applicant streams along these dimensions, given that both PSLC and AROCA most likely have different influence on different subjects and in different segments of an applicant stream. We utilize our observation from the video pilots that subjects mostly pay attention to the occurrence of applicants with $r = 1$. In a large set of applicant streams simulated without any restrictions, there occur on average 3.2 applicants with $r = 1$ before period 14 (i.e., before the optimal threshold 1), with the count being 2, 3 and 4 in about 25%, 30% and 20% of streams, respectively. We consequently impose a criterion that all our applicant streams have 2-4 occurrences of applicants with $r = 1$ before period 14 (more specifically, the number of occurrences is evenly distributed between 2, 3 and 4 for Long and Neutral streams, while the distribution is by necessity shifted slightly upwards and downwards for Short and Assessment streams, respectively). When counting applicants with $r = 1$ across all periods, we have 2-7 such applicants, with an average of 3.8 and a standard deviation of 1.2 which is slightly below the statistics for unrestricted applicant streams.

3 Results

3.1 Data description

3.1.1 Preliminaries

Our sample contains 60 subjects, 31 for the Short treatment and 29 for the Long treatment. In fact, 71 subjects participated in the experiment, but 11 of them made various mistakes in the understanding test (see the instructions in the Appendix), mostly in the test questions concerning the updating of relative ranks or the procedure of setting h. We let these subjects correct the mistakes and participate in the experiment, but based on their behavior and post-session questionnaire responses, we concluded that they indeed misunderstood various aspects of the task and hence excluded them from the analysis.

We start off by looking at stopping times and comparing them to the second experiment of Bearden at al. (2006), which we draw upon. The mean stopping time for their 30 subjects is 26.54 periods (out of 40 periods) across all 60 rounds, with

a standard deviation (of the subject means) of 3.09 periods. The mean stopping time is 25.89 and 27.20 periods for the first and second 30 rounds, respectively (see their Table 2 not reproduced here); based on this, Bearden et al. claim that subjects gradually search longer. Without making any adjustments to our raw search data at this point, the mean stopping time for our 60 subjects is 26.95 periods (23.34 periods when excluding Assessment streams), 26.66 (23.01) periods in the Short treatment and 27.27 (23.69) periods in the Long treatment, with a standard deviation (of the subject means) of around 5 both overall and in each treatment. Excluding Assessment streams, the mean stopping time is 22.70 periods in rounds 1-10 and 24.07 periods in rounds 11-21. Thus our stopping-time patterns are quite similar to those reported by Bearden et al., though we see larger between-subjects stopping-time heterogeneity in our data. This may partly be due to the fact that Bearden et al.'s subjects all face the same applicant streams (though the order is varied between subjects), whereas our six groups of subjects face different applicant streams.

3.1.2 Non-monotonic, jumping and $h > 3$ behavior, and our treatment thereof

As outlined above, an important feature of our data is the occurrence of non-monotonic, jumping and $h > 3$ behavior. Whatever lies behind setting $h > 3$, its occurrence is relatively inconsequential for our analysis and conclusions. As explained in more detail later on, such cases are not as frequent as non-monotonic and jumping behavior, they usually occur in late periods of a subject's search in a given round, and, in our view most importantly, they predominantly occur only after a subject has already entered a stream of $h = 3$, which altogether fits well with our interpretation of $h > 3$ as reflecting non-monetary "aspirations" (see above).

We are more concerned with the occurrence of non-monotonic and jumping behavior. As quantified later on, both of these types of behavior occur for a subset of subjects and in only some rounds for these subjects (i.e., only in some subject-rounds) and do not vanish over time. We encounter all possible forms of jumping behavior, in isolation (i.e., "pure" jumpers, whom we regard as a special form of an MTR search rule) or combined with non-monotonicity:

- "0-2" jumpers skip $h = 1$, so a stream of $h = 0$ is followed by a stream of $h = 2$ and possibly $h = 3$; we later interpret such subject-rounds as missing MTR threshold 1
- "0-3" jumpers skip $h = 1$ and $h = 2$, so a stream of $h = 0$ is followed by a stream of $h = 3$; we later interpret such subject-rounds as missing MTR thresholds 1 and 2
- "1-3" jumpers skip $h = 2$, so a stream of $h = 0$ and $h = 1$ is followed by a stream of $h = 3$; we later interpret such subject-rounds as missing MTR threshold 2

Non-monotonic behavior occurs whenever a subject enters h in a non-monotonic manner in a given round. Most cases of non-monotonicity are rather minor and in

our view indicate one-off "random" errors (attention lapses, perceptual errors, occasional experimentation, etc.) rather than non-MTR search rules. This interpretation is also supported by evidence from post-session questionnaires where most of the concerned subjects verbally describe monotonic MTR-like search strategies.

We are less confident about how to approach more extensive cases of non-monotonicity involving more successive periods and more values of h. Even these cases might at least partly arise from search behavior being by its nature stochastic. Both non-monotonic and jumping behavior might also be due to the flatness of the payoff function in the sense of Harrison (1989, 1992), though the expected payoff losses associated with these kinds of behavior are sometimes (though not always) non-negligible (more on this in section 3.1.3).

From subjects' post-session questionnaire responses, we in fact see that the prime reason for extensive non-monotonic and jumping behavior is a (possibly erroneous) indifference between setting various values of $0 < h \leq 3$. We say "possibly erroneous" because it is often unclear whether or not the concerned non-monotonic and jumping subjects grasped the expected-payoff consequences of setting different values of h (in a given period). Most of the concerned subjects report indifference between the three positive payoffs. However, they erroneously conclude that this implies indifference between setting the three values of $0 < h \leq 3$, ignoring that setting higher versus lower h affects the chances of stopping on positive-payoff as well as on zero-payoff applicants. Once again, it is unclear whether this kind of erroneous reasoning exists in previous studies which elicited only stopping times.

If the primary source of both non-monotonic and jumping behavior is indeed the above described kind of erroneous reasoning, then we see little justification for treating these behaviors as distinct behavioral types, and for a priori viewing them as manifestations of non-MTR search rules. Our data approach follows this logic.

In particular, we first assume that all subjects use a stochastic MTR search rule. We use a simple prediction/imputation procedure (described below) to impute a monotonic h stream for each subject-round in which the raw h stream is non-monotonic; the other (monotonic) subject-rounds, including pure jumpers, are unaffected by this procedure. Conditional on the procedure being correct, we then quantify the extent of non-monotonicity and assess which cases seem too extensive to be regarded as representing an MTR search rule. Using the predicted h streams, and having excluded the most extensive cases of non-monotonicity, we then compute MTR thresholds.

The prediction/imputation procedure minimizes, separately for each subject-round, the sum of absolute deviations between the raw and predicted h stream across all available periods (prior to the minimization, we recode all $h > 3$ to $h = 3$, but this turns out inconsequential). The only restriction we impose on this procedure is that it never "creates" new pure jumpers; there are in any case only few subject-rounds where a predicted jump-containing h stream would fit strictly better than a predicted jump-free h stream. The procedure sometimes leads to non-uniqueness where several predicted (jump-free) h streams minimize the sum of absolute deviations, especially in subject-rounds where the raw h stream contains a sequence of alternating h values. In such cases, we use three sub-rules which essentially "smooth out" the predicted h streams, similar to what a more sophisticated prediction method would do. For example, whenever the raw h stream

alternates longer than two successive periods (e.g., $h = 0,,0,1,0,1,0,1,...,$ or $h = 0,,0,2,0,2,0,2,...$), one of the sub-rules smoothes it out by stacking together first the lower and then the higher h values.

Our prediction/imputation procedure (or its variants with different objective functions; see also section 3.1.3) is of course somewhat simplistic. The nature of the secretary search task calls for dynamic duration modeling, but we would question, first, how feasible it is to implement it, and second, how much potential it has for improving prediction. As to the latter issue, minor cases of non-monotonicity are fixed by our procedure just fine, but a more formal estimation approach could be useful for extensive cases of non-monotonicity, for which it could also provide a formal measure of goodness of fit. As to the former issue, given how much heterogeneity in search behavior we document below, both across subjects and across rounds (even adjacent ones), it is hard to imagine how one would sensibly pool subject-rounds in order to make a more data-intensive estimation procedure feasible.

We use our prediction/imputation procedure for the six Assessment rounds listed in figure 2 above (i.e., rounds 3, 4, 11, 12, 19 and 20) that, with few exceptions, provide us with almost complete h streams and hence allow us to compute most MTR thresholds for a majority of subject-rounds. We also use the same procedure for rounds 13-18 that generally provide us with much shorter h streams but nevertheless permit computing at least threshold 1 (which is affected by non-random attrition only to a minor extent).

We are now ready to quantify the extent of non-monotonic behavior in terms of the sum (and per-period average) of absolute deviations between the raw and predicted h streams. Out of the total of 60*6=360 Assessment subject-rounds, 12% feature some form of non-monotonicity. In 2.5% of Assessment subject-rounds (involving 5 subjects), the sum of absolute deviations is above 8 (which coincides with the average absolute deviation being above 0.2 per period); we exclude these subject-rounds from computing MTR thresholds 2 and 3 (but not threshold 1) because non-monotonicity seems too extensive to reliably determine the position of the two thresholds. For the remaining non-monotonic cases, the sum of absolute deviations is 3-6 for 3% and below 3 for 6.5% of Assessment subject-rounds; these subject-rounds are included when computing MTR thresholds (we initially penalized the non-monotonic subject-rounds by weighting them inversely proportionally to the extent of non-monotonicity, but this yielded statistics for MTR thresholds virtually identical to the unweighted ones reported below).

The prediction/imputation procedure does not deal with pure jumpers and never "creates" new pure jumpers (except for rare non-unique cases as discussed above). We do not include "0-2" and "0-3" jumpers (subject-rounds) in the calculation of MTR threshold 1, and "1-3" and "0-3" jumpers (subject-rounds) in the calculation of MTR threshold 2. Also, we compute MTR threshold 2 both with and without "0-2" jumpers (subject-rounds), and MTR threshold 3 both with and without "0-2", "0-3" and "1-3" jumpers (subject-rounds). As a consequence, each threshold's computation documented below draws on a different subject-round data set.

Just to illustrate the extent of jumping behavior, out of 288 Assessment subject-rounds useable for computing MTR threshold 3 (the remaining 360-288=72 subject-rounds lack threshold 3 data due to early stopping or other problems explained below), there are 34.7% subject-rounds affected by jumping behavior: 6.6% are "0-2" jumpers, 14.6% are "1-3" jumpers and 13.5% are "0-3" jumpers. As explained

above, we compute threshold 3 both with and without these subject-rounds included.

3.1.3 Cost of non-monotonic, jumping and $h > 3$ behavior

Figure 3 below displays the expected cost of deviating from the optimal solution in the ways observed in our data. The "exp pay" column shows the per-round expected payoff (in EUR) of a particular search rule (i.e., the expected payoff calculated over all 40! applicant streams, using Bearden et al.'s code, as opposed to the expected payoff calculated over the set of applicant streams that our subjects actually face, which might also be informative but we have not yet done these calculations). The "exp cost" column shows the decrease in the expected payoff (in EUR) compared to the optimal MTR policy. The "% of optimal" column similarly shows the expected payoff as a percentage of the optimal policy's expected payoff. The first row titled "optimal thresholds" recapitulates that the optimal threshold 1 ($h = 1$) is 14 periods, the optimal threshold 2 ($h = 2$) is 28 periods, and the optimal threshold 3 ($h = 3$) is 35 periods; the expected payoff of the optimal MTR policy is 8.25 EUR.

The top part of figure 3 (rows 1-7) displays how our subjects typically deviate from the optimal policy, focusing first on monotonic, non-jumping and $h \leq 3$ MTR-like behavior. Row 1 reports that the average empirical MTR thresholds (calculated in later sections) are 12, 22 and 27 periods, i.e., subjects are early searchers on average. However, the last three columns indicate that the expected cost of this average early search is very small. Row 2 shows that when each of the average empirical thresholds is lowered by one standard deviation (about 4 periods for each threshold), the resulting MTR search rule still earns 81% of the optimal policy's expected payoff. Rows 4-6 additionally show that lowering the first average MTR threshold matters most and is as costly as jointly lowering the higher two thresholds. Row 7 illustrates the (infrequently observed) case where a subject enters equally long streams of $h = 0$ and $h = 1$; this rather odd search rule has similar expected payoff implications as the early-search rule in row 2. In sum, the cost of deviating from the optimal policy quite substantially - though still in an MTR-like manner - is undesirably small in the sense of Harrison (1989, 1992).

The next part of the figure displays typical behavior of "0-2", "0-3" and "1-3" jumpers. The examples refer to subjects who consistently jumped in most or all rounds, and we report their search rules in the last several rounds of the experiment. Rows 8-12 show that being a "0-2" jumper reduces one's expected payoff to about 80% of the optimal payoff (except for the extreme jumper in row 10 who pays a high penalty for setting h=2 way too early). Rows 13-16 show that being a "0-3" jumper reduces one's expected payoff to about 70% of the optimal payoff, while rows 17-19 show that being a "1-3" jumper is much less costly (or costless in row 19). Thus, with the exception of "1-3" jumpers, jumping behavior is generally costlier than the non-jumping early-search behavior in rows 1-7, though even extreme jumpers still earn a relatively high expected payoff.

We next turn to the cost of non-monotonic behavior, with implications for our prediction/imputation procedure. Rows 20, 23, 26, 29, and 32 display typical non-monotonic h streams and their expected payoff - the expected cost of these (extensively) non-monotonic search rules is relatively high. Right beneath each of the five rows, we list the associated imputed h streams and their expected payoff

- this shows that our prediction/imputation procedure substantially changes the expected payoff compared to the raw, non-monotonic h stream. This is of course a consequence of the imputation procedure having an entirely different objective function, i.e., minimizing the sum of absolute deviations between the raw and predicted h streams. The last row in each triplet (beneath the "imputed" row) shows the predicted h stream if we instead reduced (in a sensible manner) the expected payoff difference between the raw and predicted h stream. Doing so completely changes the structure of the predicted h streams and inflates the sum of absolute deviations between the raw and predicted h streams. A compromise objective function that takes into account both the sum of absolute deviations and the expected payoff difference would inevitably involve a handful of arbitrary restrictions (which do not apply universally).

3.2 Analysis of MTR thresholds

We next use the predicted h streams to compute MTR thresholds for each subject-round. As discussed above, our approach to eliciting and estimating the thresholds involves potential advantages as well as disadvantages compared to the previous literature which infers MTR thresholds from stopping times. As a benchmark, but with the above mentioned caveats in mind, we relate our threshold estimates to those reported for Bearden et al.'s (2006) experiment we draw upon. Their restricted grid search analysis of stopping times, applied separately to each individual and assuming stability of thresholds over all 60 rounds, yields aggregate (across-subjects) means of the MTR thresholds equal to periods 13, 22 and 30, as opposed to their optimal MTR thresholds equal to periods 14, 29 and 37 (recall that their optimal MTR thresholds 2 and 3 are slightly higher than ours; one should add that the estimated individual MTR thresholds yield wrong predictions of stopping times in 12 out of 60 rounds on average). As to between-subjects variability, the authors claim that it is significant but do not report any figures.

3.2.1 MTR threshold 1

Unlike for higher-order thresholds, threshold 1 is by definition not affected by non-random attrition resulting from early stopping (with one exception where a subject in one round enters $h = 1$ in period 1; this observation is excluded). This allows us to analyze more than just the Assessment rounds. We in fact analyze rounds 11-20, giving subjects plenty of time (3 warm-up and 10 real rounds) to settle on a stable search strategy (if any), and also excluding the last round 21 to minimize potential "end-game effects".

Non-monotonic behavior affects the computation of threshold 1 only to a minor extent: most imputations potentially affecting the position of threshold 1 amount to fixing one-off cases of non-monotonicity. Out of the 60*10=600 subject-rounds used for computing threshold 1, there are 2.7% of subject-rounds (involving 10 subjects) in which a stream of $h > 0$ is interrupted by one period of $h = 0$ or vice versa, or in which $h = 0$ and $h = 1$ briefly alternate around the position of threshold 1. There are further 5.3% of subject-rounds (involving 10 subjects, partly overlapping with the first group) in which a stream of $h = 1$ is interrupted by one or several h¿1 values or vice versa, but only some of these instances occur around the position of threshold 1 (i.e., around the start of the $h = 1$ stream). We deal

	Optimal solution and deviations	h=1	h=2	h=3	exp pay	exp cost	% of optimal
	optimal thresholds	14	28	35	8.25	0.00	100%
row							
1	average empirical thresholds (AET)	12	22	27	7.76	0.48	94%
2	AET each lowered by 1 st.dev.	8	18	23	6.69	1.56	81%
3	AET each raised by 1 st.dev.	16	26	31	8.10	0.14	98%
4	only first AET lowered by 1 st.dev.	8	22	27	7.11	1.13	86%
5	only first two AET lowered by 1 st.dev.	8	18	27	6.86	1.39	83%
6	only last two AET lowered by 1 st.dev.	12	18	23	7.09	1.15	86%
7	0-1 person (no thresholds 2 and 3)	21			6.42	1.83	78%
	"0-2" jumpers						
8	example 1		25	29	6.52	1.73	79%
9	example 2		21	23	6.23	2.02	76%
10	example 3		6	34	3.31	4.94	40%
11	example 4		19	24	6.53	1.72	79%
12	example 5		15	32	6.63	1.62	80%
	"0-3" jumpers						
13	example 1			21	5.77	2.48	70%
14	example 2			23	5.91	2.34	72%
15	example 3			27	5.80	2.45	70%
16	example 4			18	5.34	2.91	65%
	"1-3" jumpers						
17	example 1	11		21	7.11	1.14	86%
18	example 2	14		22	7.25	1.00	88%
19	example 3	17		31	7.97	0.28	97%
	non-monotonic						
20	0…0 32132111111222223…321213211				4.45	3.80	54%
21	imputed	9	20	25	7.16	1.08	87%
22	reduced expected payoff difference	5	9	27	4.45	3.80	54%
23	15 x 0 333221…1222				5.81	2.43	70%
24	imputed	16	35		7.80	0.45	95%
25	reduced expected payoff difference	8	12		5.91	2.34	72%
26	0001110…0111 27				2.70	5.55	33%
27	imputed	29		32	5.63	2.62	68%
28	reduced expected payoff difference	2		27	2.79	5.46	34%
29	000010101010221113...3				5.38	2.87	65%
30	imputed	9		19	6.58	1.67	80%
31	reduced expected payoff difference	5		19	5.37	2.88	65%
32	11 x 0 3 000 3 111 3 ...3				5.36	2.89	65%
33	imputed	16		21	6.80	1.45	82%
34	reduced expected payoff difference	5		21	5.53	2.71	67%
	h>3	h=1	h=3	h=4			
35	example 1	6	10	13	3.68	4.57	45%
36	imputed	6	10		3.94	4.31	48%

Figure 3: Expected costs of deviating from the optimal strategy

with all these cases in accordance with the above described prediction/imputation procedure and do not exclude any of them when computing threshold 1.

Jumping behavior affects the computation of threshold 1 to a much larger extent. Recall that we interpret "0-2" and "0-3" jumpers (subject-rounds) as missing threshold 1. For rounds 11-20, 8 subjects (2 from the Short treatment and 6 from the Long treatment) have no or only one threshold 1. We exclude these subjects from computing threshold 1 since most statistics cannot be computed for them. Another 13 subjects have threshold 1 in 5-9 rounds. We use these subjects (subject-rounds) when computing various between-subjects statistics for threshold 1, so the sample composition changes slightly over rounds. For within-subjects

statistics, we further exclude 4 of these subjects since their (5 or 6) threshold 1 observations are insufficiently balanced over rounds to permit an "early-late" comparison of behavior.

Figure 4 summarizes the sample sizes (number of subjects) used for computing threshold 1 in each treatment and overall, though this summary hides the between-rounds changes in sample composition arising from the exclusions described above. Figure 5 shows the mean and median threshold 1 for the pooled sample, accom-

round	treatment Short	Long	Total
11	27	21	48
12	29	20	49
13	29	22	51
14	28	22	50
15	27	22	49
16	26	23	49
17	26	23	49
18	25	23	48
19	26	23	49
20	27	22	49

Figure 4: Sample sizes used for computing threshold 1 by treatment and overall

panied by the dashed mean absolute deviation from the mean (mdev) and median absolute deviation from the median (mad). Our mean threshold 1 is close to Bearden et al.'s mean estimate of 13 periods and hence also close to the optimal threshold 1 of 14 periods; our median is slightly lower. In terms of both the mean and median absolute deviations, between-subject heterogeneity is quite substantial and does not decrease over time. We show below how much within-subjects heterogeneity (over rounds) this graph conceals. Figures 6 and 7 show separate

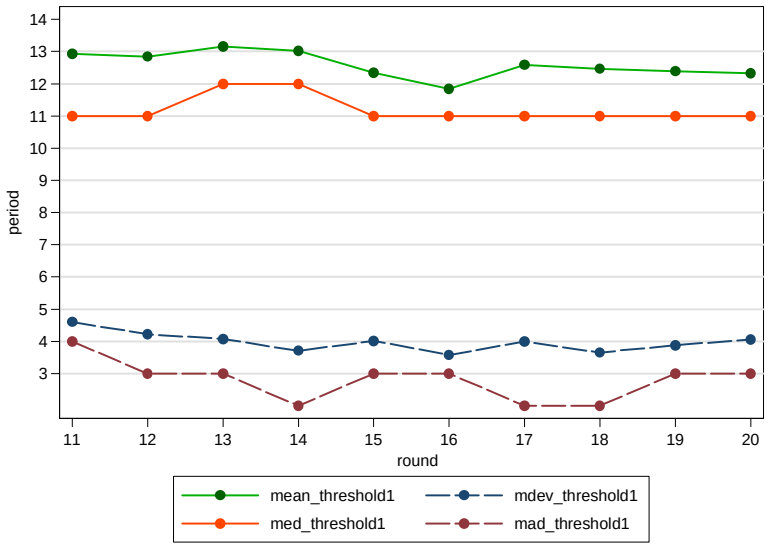

Figure 5: Between-subject summary statistics for threshold 1 for the pooled sample

statistics for the Short and Long treatments, namely, averages and standard devi-

ations in the first graph, and medians and 10th and 90th percentiles in the second graph (the various statistics should be viewed with the relatively small sample sizes per treatment in mind). Both central tendency and variability are similar across treatments, so our treatment design does not seem to induce a treatment effect for threshold 1. This could be expected, given that the Short and Long treatments differ relatively little in the frequency of positive-payoff applicants in (very) early search periods; hence the treatment design is a priori more likely to "work" for the higher-order thresholds examined below.

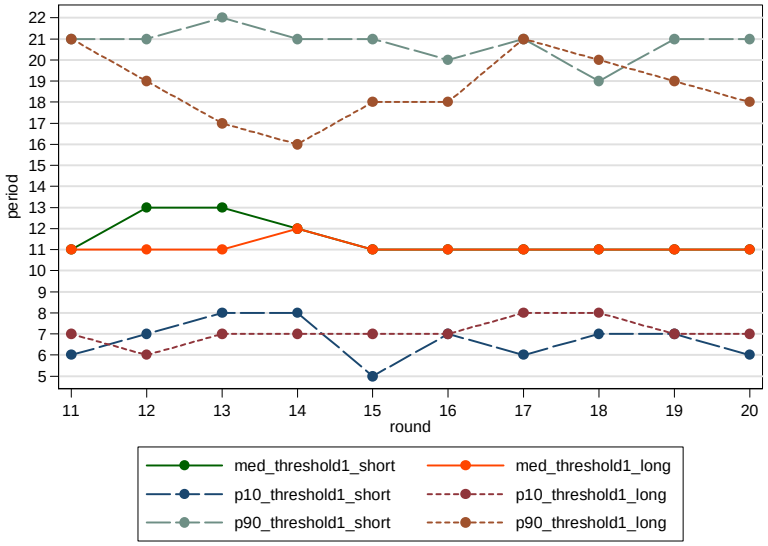

Figure 6: Between-subject summary statistics for threshold 1 by treatment [a]

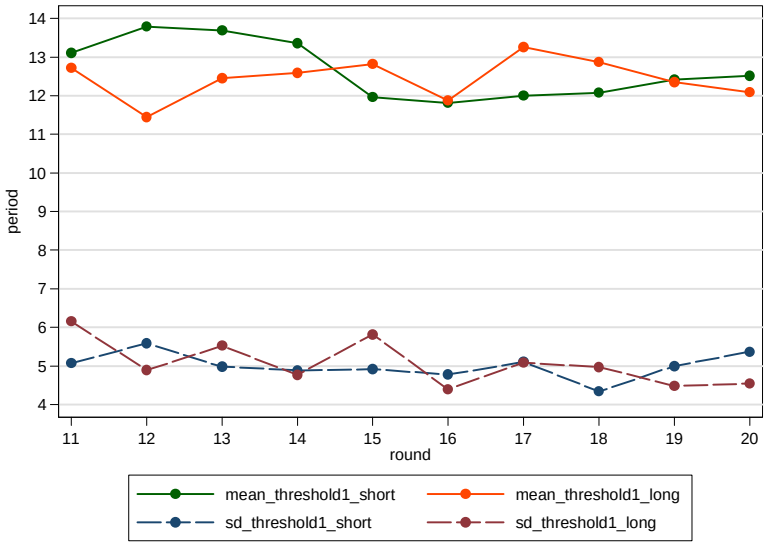

Figure 7: Between-subject summary statistics for threshold 1 by treatment [b]

We next look at within-subjects variation over rounds. The general patterns do not differ much across the two treatments, so we pool the data. We compute various statistics for all rounds and separately for "early" rounds (11-15) and "late" rounds (16-20). One should not take the "early" and "late" labels literally

since subjects already complete 13 search rounds (counting also the three warm-up rounds) before the "early" rounds. As one can see in figure 2 above, the stream-type composition of the early- and late-round segments is quite similar within each treatment. We could instead focus only on the Assessment rounds, some of which even use exactly the same applicant streams, but we decided to utilize as much data as possible and focus on the Assessment-rounds comparisons only for the higher-order thresholds below.

Figure 22 in the appendix shows the distribution of the mean absolute deviation from the mean for each subject, first across all rounds and then across the early and late rounds (the pattern of distributions of the median absolute deviation from the median are similar). There is quite substantial within-subjects variation in threshold 1 - about 1 period per round on average and at the median - which decreases only slightly over time (comparing the early and late figures). Also, the within-subjects variation varies considerably across subjects, i.e., some subjects adjust threshold 1 markedly while others make no adjustments at all.

We next look closer at the direction in which subjects on average adjust their threshold 1 in early versus late rounds. In particular, figure 23 in the appendix display the distribution of within-subjects differences in the mean and median threshold 1, respectively, calculated for the (available) early and late rounds. The figure shows that while the early-late adjustment of threshold 1 is on average close to zero, some subjects tend to adjust their threshold 1 upwards over rounds while others do the opposite. The magnitude of the adjustments seems non-negligible, especially taking into account that it occurs in late stages of the experiment where subjects already have had plenty of experience with setting threshold 1 (they must have done so in every previous round, unlike for the higher-order thresholds).

3.2.2 MTR threshold 2

For threshold 2 and threshold 3, we only analyze the six Assessment rounds (3, 4, 11, 12, 19 and 20) for which we generally observe almost complete h streams. As for threshold 1, we exclude few subject-rounds where subjects (perhaps accidentally) stop in initial periods because of entering an h equal to the period number - this concerns two subjects in round 3, three subjects in round 4, and one subject in round 12. As explained above, we also exclude the 2.5% of Assessment subject-rounds with extensive non-monotonicity, while minor non-monotonic cases (outlined above) are included and dealt with according to our imputation procedure.

As to jumping behavior and its effect on the computation of threshold 2, recall that "0-3" and "1-3" jumpers (subject-rounds) are interpreted as missing threshold 2. For 34 subjects, threshold 2 is missing in one or more Assessment rounds due to jumping (or for reasons explained in the previous paragraph). We exclude 10 subjects (5 from the Short treatment and 5 from the Long treatment) who have no or only one threshold 2, since most statistics cannot be computed for them. The remaining 24 subjects have threshold 2 in 2-5 Assessment rounds. We use these subjects (subject-rounds) when computing various between-subjects statistics for threshold 2, so the sample composition changes slightly over rounds, but 4 of these subjects with less than four observations for threshold 2 are excluded from the computation of within-subjects statistics (i.e., "early-late" comparisons). Figure 8 summarizes the sample sizes (number of subjects) used for computing

threshold 2 in each treatment and overall. The sample sizes are stable or increase slightly over rounds, but sample composition changes as discussed above. Out

round	treatment Short	Long	Total
3	21	17	38
4	22	16	38
11	23	16	39
12	22	19	41
19	24	22	46
20	25	21	46

Figure 8: Sample sizes used for computing threshold 2 by treatment and overall

of the subjects included in figure 8, the ones in figure 9 are not "0-2" jumpers; the remaining few subjects are thus "0-2" jumpers whom we both include and exclude when computing threshold 2. More specifically, we observe that three subjects are always "0-2" jumpers while the remainder jumps only occasionally. The occasional "0-2" jumpers sometimes (though not always) have quite different threshold 2 when they jump compared to when they do not. This is an indication that setting h=2 may not have the same meaning for "0-2" jumpers as it has for non-jumpers. Figures 10 and 11 below show the mean and median threshold 2

round	treatment Short	Long	Total
3	17	17	34
4	22	14	36
11	22	15	37
12	21	15	36
19	21	20	41
20	23	18	41

Figure 9: Sample sizes used for computing threshold 2 by treatment and overall, excluding 0-2 jumpers

for the pooled sample, accompanied by the dashed mean absolute deviation from the mean (mdev) and median absolute deviation from the median (mad). The first graph contains "0-2" jumpers while the second graph does not. Comparing the graphs, one can see that including the few "0-2" jumpers slightly decreases the means and the medians. The mean and median threshold 2 is generally highest in rounds 11 and 12 and eventually "stabilizes" at 22-23 periods, which is slightly higher than Bearden et al.'s mean estimate of 22 periods (but recall that our optimal threshold 2 is 28 periods while Bearden et al.'s is 29 periods). The between-subject heterogeneity is again quite substantial and decreases only slightly over time. Figures 12 and 13 show separate statistics for the Short and Long treatments, namely, averages and standard deviations in the first graph, and medians and 10th and 90th percentiles in the second graph (one should again bear in mind the relatively small sample sizes per treatment). Both graphs include "0-2"

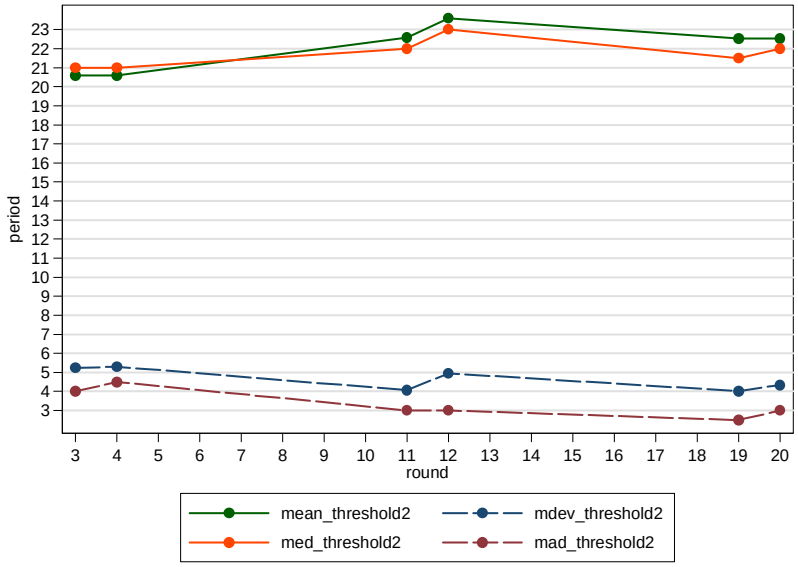

Figure 10: Between-subject summary statistics for threshold 2 for the pooled sample

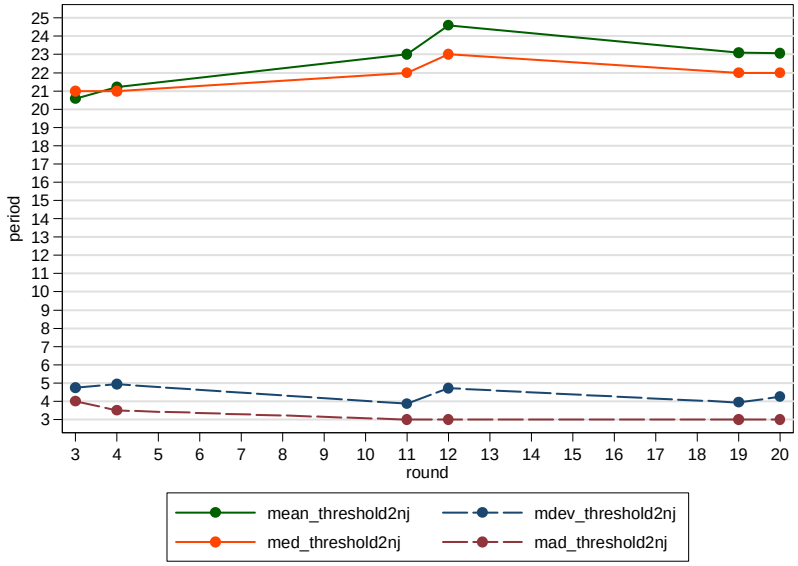

Figure 11: Between-subject summary statistics for threshold 2 for the pooled sample, excluding 0-2 jumpers

jumpers since the general patterns are very similar with or without them. Both the graphs mostly (though not always) show a treatment effect in the expected direction in later Assessment rounds. Looking only at the means and the medians, the size of the treatment effect fluctuates between 0 and 4 periods, being largest in the upper part of the distribution (90th percentile). An interesting phenomenon (already partly observed for threshold 1) is the non-negligible fluctuation of the various statistics (and hence of the treatment effect) in the adjacent Assessment rounds; more on this below. We next look at within-subjects variation over rounds. For this analysis, we find it "cleaner" to exclude the few "0-2" jumpers

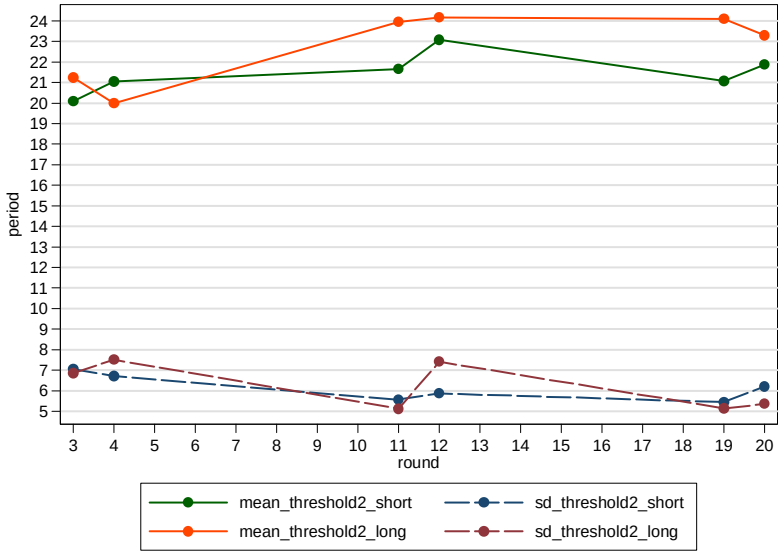

Figure 12: Between-subject summary statistics for threshold 2 by treatment [a]

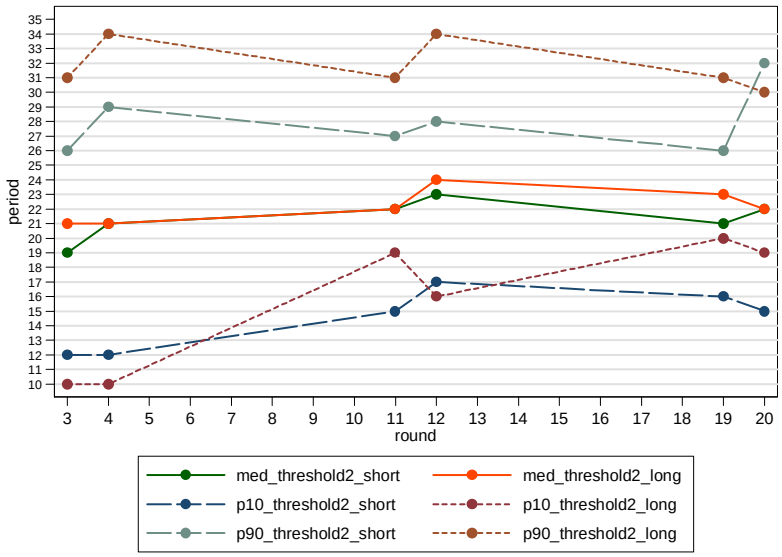

Figure 13: Between-subject summary statistics for threshold 2 by treatment [b]

(subject-rounds) because their within-subject variation partly arises from switching between jumping and non-jumping. Subjects with less than four observations for threshold 2 are also dropped. This leaves us with the following number of subjects in each treatment and in total that is shown in figure 14, with sample composition changing slightly over rounds as explained above. Figure 24 in the appendix shows the distribution of the mean (median) absolute deviation from the mean (median) for each subject. The general patterns do not differ much across the two treatments, so we pool the data. On average, there is substantial within-subjects variation in threshold 2 over the six Assessment rounds (recall that this variation is not inflated by "0-2" jumpers who generally have even higher variation in threshold 2). The within-subjects variation also varies considerably across subjects, i.e., some subjects adjust threshold 2 considerably while others make no

round	treatment		
	Short	Long	Total
3	16	13	29
4	19	13	32
11	21	15	36
12	19	13	32
19	20	15	35
20	22	15	37

Figure 14: Sample sizes used for computing within-subjects statistics for threshold 2 by treatment and overall

adjustments at all.

In figure 25 in the appendix we present within-subjects comparisons between various pairs of Assessment rounds. Each comparison naturally includes only those subjects who have both of the threshold 2 observations; the number of subjects involved in the various statistics can be inspected in each figure (variable "Obs"). Figure 25 shows the distribution of threshold 2 adjustments between Assessment rounds 3 and 11 which feature identical applicant streams (see the green table above), for the pooled sample and then separately for each treatment. On average, there is an upward adjustment in threshold 2, which is more pronounced in the Long treatment as expected. The between-subjects variability of the adjustment is large.

Analogously, figure 26 in the appendix shows the distribution of threshold 2 adjustments between Assessment rounds 4 and 20 which again feature identical applicant streams, for the pooled sample and then separately for each treatment. Here the Short treatment features little or no upward adjustment on average, contrary to the Long treatment. The between-subjects variability of the adjustment is again substantial.

We next look at within-subjects comparison of adjacent Assessment rounds (leaving out rounds 3 and 4). Figure 27 in the appendix shows the distribution of threshold 2 adjustments between the adjacent Assessment rounds 11 and 12 (which feature different applicant streams). The adjustments are small on average but rather large for a non-negligible proportion of subjects, especially taking into account that these are adjacent search rounds (though the adjustments are generally smaller than for nonadjacent Assessment rounds).

Last, figure 28 in the appendix shows the distribution of threshold 2 adjustments between the last pair of adjacent Assessment rounds 19 and 20 (which again feature different applicant streams). The adjustments are small on average and for most subjects, and much smaller compared to those observed for rounds 11 and 12 (though this comparison is also subject to changing sample composition and sample size).

3.2.3 MTR threshold 3

Here we again analyze only the six Assessment rounds for which we generally observe almost complete h streams. In addition to the subjects excluded from the

calculation of threshold 2 due to very early stopping or extensive non-monotonicity (see the first paragraph of the previous section), we exclude two subjects in round 11 for whom no threshold 3 exists because they enter $h = 3$ only in the last period (we interpret this as a nonexistent threshold 3 since one can enter any $h \geq 3$ in the last period with no payoff consequences).

As already discussed in early sections, the computation of threshold 3 is partly affected by $h > 3$ behavior, which usually occurs in late periods of a subject's search in a given round. In general, its occurrence is quite frequent for a subset of 6 subjects (in 3 or more rounds 3-21). Importantly, $h > 3$ behavior predominantly occurs only after a subject has entered a stream of $h = 3$, which we interpret as having no influence on the position of threshold 3. There are rare cases (8 subject-rounds) where a stream of $h > 3$ directly follows a stream of $h < 3$ (we interpret this switching point as threshold 3, or threshold 3+ if you wish) or where subjects enter $h > 3$ in a one-off manner within a stream of $h < 3$ (which is taken care of by our imputation procedure). There are two odd cases of $h > 3$ behavior (in rounds 3 and 4) which are excluded from the analysis of threshold 3.

Besides the above exclusions, there is non-random attrition of subjects due to stopping search before reaching threshold 3. There are in total 24 subjects for whom threshold 3 is missing in one or more Assessment rounds due to attrition (or for reasons explained above). We exclude four of these subjects (two from each treatment) who have only one threshold 3, since most statistics cannot be computed for them. The remaining 20 subjects have threshold 3 in 2-5 Assessment rounds. We use these subjects (subject-rounds) when computing various between-subjects statistics for threshold 3, so the sample composition changes slightly over rounds, but 10 of these subjects with less than four observations for threshold 3 are excluded from the computation of within-subjects statistics (i.e., "early-late" comparisons). Figure 15 summarizes the sample sizes (number of subjects) used for computing threshold 3 in each treatment and overall. The sample sizes are quite stable over rounds, but sample composition changes as documented right above. Out of the subjects included in figure 15, those in figure 16 are not jumpers;

round	treatment Short	Long	Total
3	26	24	50
4	23	21	44
11	25	23	48
12	23	23	46
19	26	25	51
20	24	25	49

Figure 15: Sample sizes used for computing threshold 3 by treatment and overall

the remaining subjects (about a third in each round) are "0-2" or "1-3" or "0-3" jumpers who we both include and exclude when computing threshold 3. More specifically, there are 34.7% subject-rounds affected by jumping behavior: 6.6% are "0-2" jumpers, 14.6% are "1-3" jumpers and 13.5% are "0-3" jumpers. Some subjects jump always or almost always while others jump occasionally or only once. Some subjects even switch between the three jumping types. Figures 17 and 18 show the mean and median threshold 3 for the pooled sample, accompanied by

round	treat 1	2	Total
3	18	15	33
4	17	10	27
11	19	11	30
12	17	13	30
19	17	17	34
20	18	15	33

Figure 16: Sample sizes used for computing threshold 3 by treatment and overall, excluding jumpers

the dashed mean absolute deviation from the mean (mdev) and median absolute deviation from the median (mad). The first graph contains jumpers while the second graph does not. Comparing the graphs, one can see that including jumpers decreases the means and the medians by about 2 periods. The mean and median threshold 3 figures start of at around 25-26 periods and eventually "stabilize" at around 27-30 periods, which is slightly lower than Bearden et al.'s mean estimate of 30 periods (but recall that our optimal threshold 3 is 35 periods while Bearden et al.'s is 37 periods). The between-subject heterogeneity is again quite substantial and decreases only slightly over time. Figures 19 and 20 show separate statistics

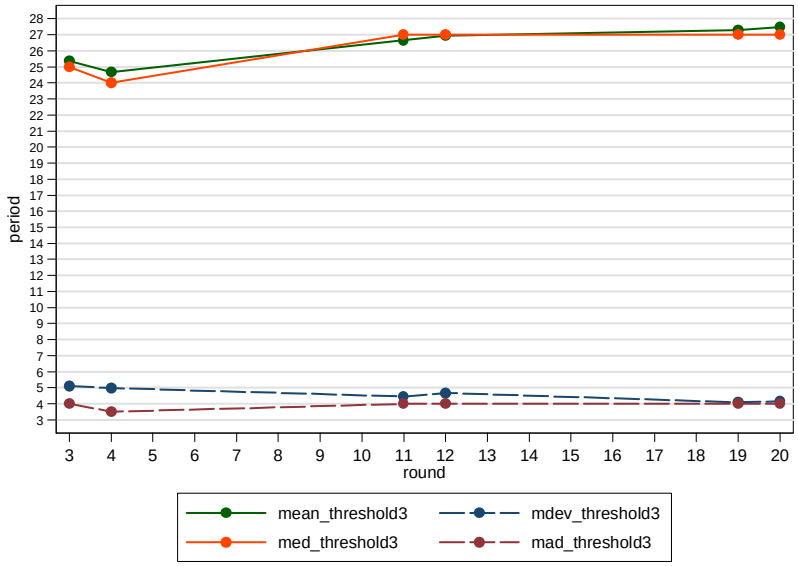

Figure 17: Between-subject summary statistics for threshold 3 for the pooled sample

for the Short and Long treatments, namely, averages and standard deviations in the first graph, and medians and 10th and 90th percentiles in the second graph (one should again bear in mind the relatively small sample sizes per treatment). Both graphs include jumpers since the patterns are very similar with or without them. Both graphs generally show a treatment effect in the expected direction in later Assessment rounds. Looking only at the means and medians, the size of the treatment effect fluctuates between 2 and 4 periods, being largest in the upper

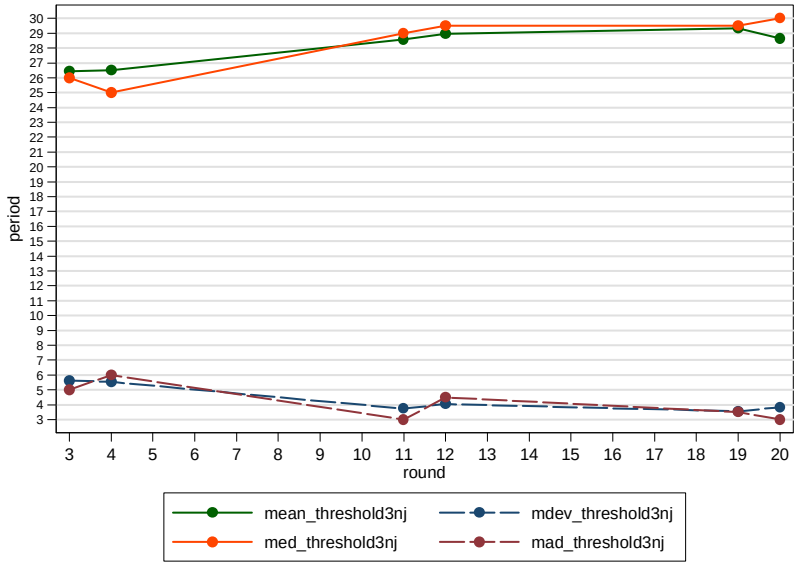

Figure 18: Between-subject summary statistics for threshold 3 for the pooled sample, excluding jumpers

part of the distribution (90th percentile). Early on, the various statistics (and hence of the treatment effect) fluctuate between the adjacent Assessment rounds, but these fluctuations disappear in the last two adjacent Assessment round; more on this below. We next look at within-subjects variation over rounds. We again

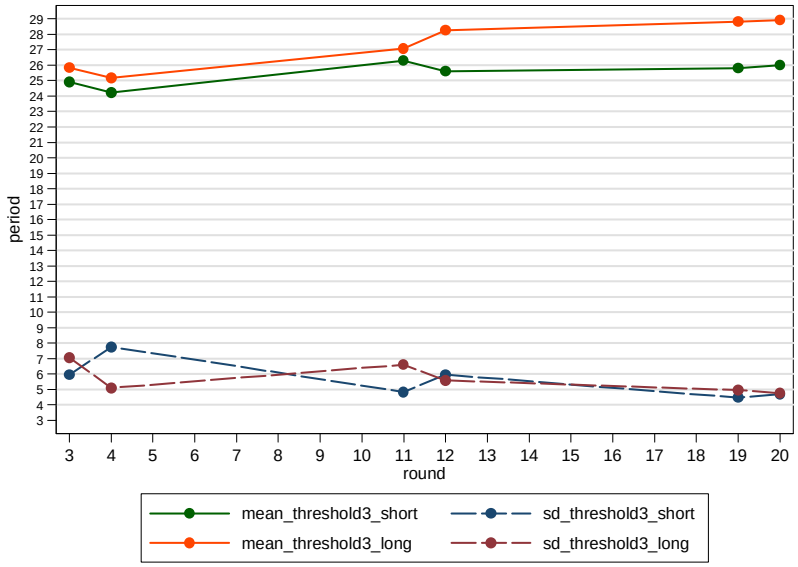

Figure 19: Between-subject summary statistics for threshold 3 by treatment [a]

find it "cleaner" to exclude all jumpers (subject-rounds) because their within-subject variation partly arises from switching between jumping and non-jumping or between jumping types. Subjects with less than four observations for threshold 3 are also dropped. This leaves us with the following relatively small number of subjects in each treatment and in total, with sample composition changing slightly over rounds as explained above. Figure 29 in the appendix shows the distribution

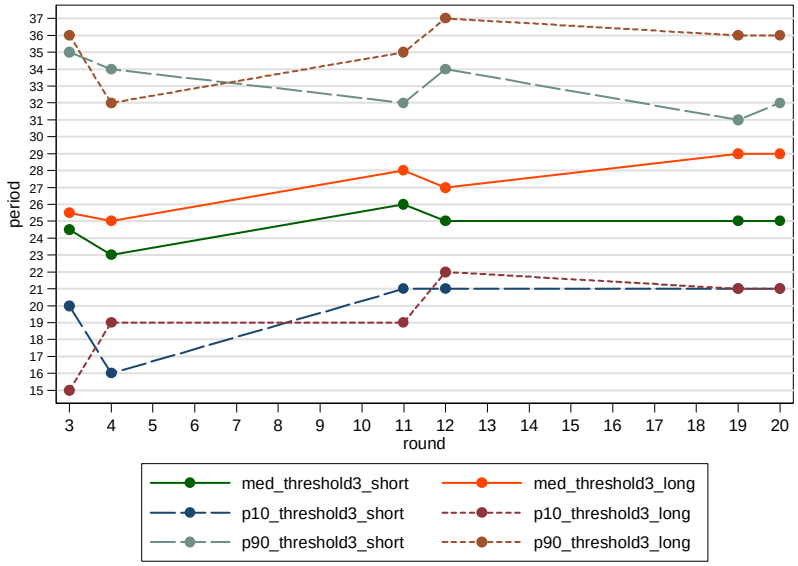

Figure 20: Between-subject summary statistics for threshold 3 by treatment [b]

round	treat 1	2	Total
3	12	10	22
4	13	9	22
11	15	9	24
12	15	10	25
19	14	11	25
20	16	11	27

Figure 21: Sample sizes used for computing within-subjects statistics for threshold 3 by treatment and overall

of the mean (median) absolute deviation from the mean (median) for each subject. The figures do not differ much across the two treatments, so we pool the data. On average, there is substantial within-subjects variation in threshold 3 over the six Assessment rounds, quite similar to the variation observed for threshold 2. The within-subjects variation also varies considerably across subjects, i.e., some subjects adjust threshold 3 markedly while others make no adjustments at all. Overall, the size of the variation is similar to that observed for threshold 2.

As for threshold 2, we next present within-subjects comparisons between various pairs of Assessment rounds. Each comparison includes only those subjects who have both of the threshold 3 observations; the number of subjects involved in the various statistics can be inspected in each figure (variable "Obs"). Figure 30 in the appendix shows the distribution of threshold 3 adjustments between Assessment rounds 3 and 11 which feature identical applicant streams, for the pooled sample and then separately for each treatment. On average, there is an upward adjustment in threshold 3 (much larger than for threshold 2), which is more pronounced in the Long treatment as expected. Similar to threshold 2, the between-subjects variability of the adjustment is very large.

Analogously, figure 31 in the appendix shows the distribution of threshold 3 adjustments between Assessment rounds 4 and 20 which again feature identical applicant streams, for the pooled sample and then separately for each treatment. Here the Short treatment features little or no upward adjustment on average, in contrast to the Long treatment. The between-subjects variability of the adjustment is again substantial.

Figure 32 in the appendix shows the distribution of threshold 3 adjustments between the adjacent Assessment rounds 11 and 12 (which feature different applicant streams). The adjustments are zero on average but again rather large for a non-negligible proportion of subjects, taking into account that these are adjacent search rounds (though the adjustments are generally smaller than for nonadjacent Assessment rounds, and smaller than the corresponding threshold 2 adjustments).

Last, figure 33 in the appendix shows the distribution of threshold 3 adjustments between the last pair of adjacent Assessment rounds 19 and 20 (which again feature different applicant streams). The adjustments are again small on average and for most subjects, and much smaller compared to those observed for rounds 11 and 12 (though this comparison is also subject to changing sample composition and sample size).

4 Conclusion

In this experiment we investigate behavior in a sequential search task known as the secretary search problem. Using a procedure that makes search strategies explicit, we are able to analyze aspects of individual behavior that could not be analyzed before. The results are ambivalent. On one hand, they confirm previous studies showing that subjects search too short on average. On the other hand, they suggest that search behavior is more complex and less stable than assumed. First, we find that not all subjects use multi-threshold rules (MTRs). Second, those subjects who do use MTRs use very different thresholds, both across and within subjects. Third, subjects' thresholds are influenced by the characteristics of the applicant streams they face, although they know of the random nature of these streams.

References

Bearden, J.N., Rapoport, A., and R.O. Murphy (2006). Sequential observation and selection with rank-dependent payoffs: An experimental test. Management Science, 52(9), 1437-1449.

Fischbacher, U. (2007).z-Tree: Zurich Toolbox for Ready-made Economic Experiments. Experimental Economics, 10(2), 171-178.

Harrison, G.W. (1992). Theory and misbehavior of first-price auctions: Reply. American Economic Review, 82(5), 1426-1443.

Harrison, G.W. (1989). Theory and misbehavior of first-price auctions. American Economic Review, 79(4), 749-762.

Seale, D. A., and A. Rapoport (1997). Sequential decision making with relative ranks: An experimental investigation of the secretary problem. Organizational Behavior and Human Decision Processes, 69(3), 221-236.

Seale, D. A., and A. Rapoport (2000). Optimal stopping behavior with relative ranks: The secretary problem with unknown population size. Journal of Behavioral Decision Making, 13(4), 391-411.

Sonnemans, J. (1998). Strategies of Search. Journal of Economic Behavior & Organization, 35, 309-332.

Zwick, R., Rapoport, A. Lo, A. K. C., and A. V. Muthukrishnan (2003). Consumer sequential search: Not enough or too much? Marketing Science, 22(4), 503-519.

Appendix

mdev_threshold1

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	48
25%	.32	0	Sum of Wgt.	48
50%	1.11		Mean	1.237554
		Largest	Std. Dev.	1.102893
75%	2.09	2.91358		
90%	2.86	3.555556	Variance	1.216374
95%	3.555556	3.68	Skewness	.7280064
99%	3.9	3.9	Kurtosis	2.492652

mdev_threshold1_early

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	48
25%	.32	0	Sum of Wgt.	48
50%	.8		Mean	1.198426
		Largest	Std. Dev.	1.149484
75%	2.24	3.12		
90%	3	3.28	Variance	1.321314
95%	3.28	3.76	Skewness	.8766525
99%	4.24	4.24	Kurtosis	2.709503

mdev_threshold1_late

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	48
25%	.32	0	Sum of Wgt.	48
50%	.76		Mean	.9688657
		Largest	Std. Dev.	.9576947
75%	1.44	2.56		
90%	2.5	2.88	Variance	.917179
95%	2.88	2.88	Skewness	1.292239
99%	4.32	4.32	Kurtosis	4.668546

Figure 22: The distribution of the within-subject mean absolute deviation of threshold 1 for rounds 11-20, 11-15, and 16-20

meandif_threshold1

Percentiles		Smallest		
1%	-6.8	-6.8		
5%	-4.549999	-4.6		
10%	-3.400001	-4.549999	Obs	48
25%	-1.3	-4	Sum of Wgt.	48
50%	-.1499996		Mean	-.5333333
		Largest	Std. Dev.	1.899854
75%	.1999998	1.400001		
90%	1.4	1.8	Variance	3.609444
95%	1.8	3.6	Skewness	-.8514491
99%	3.8	3.8	Kurtosis	5.074171

meddif_threshold1

Percentiles		Smallest		
1%	-8	-8		
5%	-5.5	-6		
10%	-3	-5.5	Obs	48
25%	-1	-4	Sum of Wgt.	48
50%	0		Mean	-.5416667
		Largest	Std. Dev.	2.118393
75%	0	2		
90%	1	3	Variance	4.487589
95%	3	3	Skewness	-1.274823
99%	4	4	Kurtosis	6.067852

Figure 23: The distribution of the within-subject differences in mean and median threshold 1 between rounds 11-15 and 16-20

mdev_threshold2

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	.32	0	Obs	37
25%	1.333333	.32	Sum of Wgt.	37
50%	2.444444		Mean	2.371441
		Largest	Std. Dev.	1.522136
75%	3.333333	4.75		
90%	4.75	4.88	Variance	2.316899
95%	4.888889	4.888889	Skewness	.1377973
99%	5.666667	5.666667	Kurtosis	2.211356

mad_threshold2

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	37
25%	0	0	Sum of Wgt.	37
50%	1.5		Mean	1.405405
		Largest	Std. Dev.	1.25741
75%	2	3.5		
90%	3.5	3.5	Variance	1.581081
95%	4	4	Skewness	.6099296
99%	4.5	4.5	Kurtosis	2.601567

Figure 24: The distribution of the within-subject mean and median absolute deviation of threshold 2, excluding 0-2 jumpers

dif pooled

Percentiles		Smallest		
1%	-10	-10		
5%	-4	-4		
10%	-3	-3	Obs	28
25%	-.5	-3	Sum of Wgt.	28
50%	3		Mean	3.142857
		Largest	Std. Dev.	5.835487
75%	6.5	11		
90%	12	12	Variance	34.05291
95%	13	13	Skewness	.140051
99%	15	15	Kurtosis	2.646407

dif Short

Percentiles		Smallest		
1%	-10	-10		
5%	-10	-3		
10%	-3	-2	Obs	15
25%	-1	-1	Sum of Wgt.	15
50%	0		Mean	2.8
		Largest	Std. Dev.	6.909621
75%	10	10		
90%	13	12	Variance	47.74286
95%	15	13	Skewness	.3299484
99%	15	15	Kurtosis	2.413192

dif Long

Percentiles		Smallest		
1%	-4	-4		
5%	-4	-3		
10%	-3	-2	Obs	13
25%	0	0	Sum of Wgt.	13
50%	5		Mean	3.538462
		Largest	Std. Dev.	4.539005
75%	6	6		
90%	8	7	Variance	20.60256
95%	11	8	Skewness	-.3157261
99%	11	11	Kurtosis	2.100694

Figure 25: The distribution of the within-subject differences in threshold 2 between rounds 3 and 11 for the pooled sample, Short treatment and Long treatment, excluding 0-2 jumpers

dif pooled

Percentiles		Smallest		
1%	-12	-12		
5%	-9	-9		
10%	-6	-6	Obs	32
25%	-2.5	-6	Sum of Wgt.	32
50%	1		Mean	1.03125
		Largest	Std. Dev.	6.056105
75%	3.5	7		
90%	7	8	Variance	36.67641
95%	11	11	Skewness	.6360197
99%	20	20	Kurtosis	4.754423

dif Short

Percentiles		Smallest		
1%	-12	-12		
5%	-12	-6		
10%	-6	-5	Obs	19
25%	-3	-4	Sum of Wgt.	19
50%	0		Mean	.6842105
		Largest	Std. Dev.	6.749919
75%	2	4		
90%	11	7	Variance	45.5614
95%	20	11	Skewness	1.093085
99%	20	20	Kurtosis	5.138699

dif Long

Percentiles		Smallest		
1%	-9	-9		
5%	-9	-6		
10%	-6	-4	Obs	13
25%	0	0	Sum of Wgt.	13
50%	3		Mean	1.538462
		Largest	Std. Dev.	5.09273
75%	5	5		
90%	7	6	Variance	25.9359
95%	8	7	Skewness	-.7878654
99%	8	8	Kurtosis	2.643445

Figure 26: The distribution of the within-subject differences in threshold 2 between rounds 4 and 20 for the pooled sample, Short treatment and Long treatment, excluding 0-2 jumpers

dif pooled

Percentiles		Smallest		
1%	-9	-9		
5%	-8	-8		
10%	-5	-7	Obs	31
25%	-2	-5	Sum of Wgt.	31
50%	0		Mean	.7096774
		Largest	Std. Dev.	5.12636
75%	3	5		
90%	5	11	Variance	26.27957
95%	12	12	Skewness	.6671348
99%	14	14	Kurtosis	3.963479

dif Short

Percentiles		Smallest		
1%	-9	-9		
5%	-9	-7		
10%	-7	-5	Obs	18
25%	-2	-4	Sum of Wgt.	18
50%	0		Mean	.1666667
		Largest	Std. Dev.	5.447611
75%	2	2		
90%	11	2	Variance	29.67647
95%	14	11	Skewness	.9647232
99%	14	14	Kurtosis	4.325535

dif Long

Percentiles		Smallest		
1%	-8	-8		
5%	-8	-3		
10%	-3	-2	Obs	13
25%	0	0	Sum of Wgt.	13
50%	1		Mean	1.461538
		Largest	Std. Dev.	4.754215
75%	3	3		
90%	5	5	Variance	22.60256
95%	12	5	Skewness	.2285667
99%	12	12	Kurtosis	3.751939

Figure 27: The distribution of the within-subject differences in threshold 2 between rounds 11 and 12 for the pooled sample, Short treatment and Long treatment, excluding 0-2 jumpers

dif pooled

Percentiles		Smallest		
1%	-6	-6		
5%	-5	-5		
10%	-2	-4	Obs	39
25%	-1	-2	Sum of Wgt.	39
50%	0		Mean	.3333333
		Largest	Std. Dev.	2.765705
75%	1	4		
90%	4	6	Variance	7.649123
95%	7	7	Skewness	.8731785
99%	9	9	Kurtosis	5.414159

dif Short

Percentiles		Smallest		
1%	-2	-2		
5%	-1	-1		
10%	-1	-1	Obs	21
25%	0	-1	Sum of Wgt.	21
50%	0		Mean	.952381
		Largest	Std. Dev.	2.578298
75%	1	2		
90%	2	2	Variance	6.647619
95%	7	7	Skewness	2.062584
99%	9	9	Kurtosis	6.727355

dif Long

Percentiles		Smallest		
1%	-6	-6		
5%	-6	-5		
10%	-5	-4	Obs	18
25%	-1	-1	Sum of Wgt.	18
50%	0		Mean	-.3888889
		Largest	Std. Dev.	2.872566
75%	0	0		
90%	4	3	Variance	8.251634
95%	6	4	Skewness	.1575981
99%	6	6	Kurtosis	3.460714

Figure 28: The distribution of the within-subject differences in threshold 2 between rounds 19 and 20 for the pooled sample, Short treatment and Long treatment, excluding 0-2 jumpers

```
                    mdev_threshold3
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0             0
 5%          .48           .48
10%           .5            .5       Obs                  27
25%     .8333333      .5555556       Sum of Wgt.          27

50%         2.125                    Mean           2.375185
                        Largest      Std. Dev.      1.694292
75%         3.625             4
90%           4.8           4.8      Variance       2.870627
95%      5.888889      5.888889      Skewness       .7535293
99%      6.555555      6.555555      Kurtosis       2.942491


                    mad_threshold3
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0             0
 5%            0             0
10%            0             0       Obs                  27
25%           .5             0       Sum of Wgt.          27

50%            1                     Mean           1.314815
                        Largest      Std. Dev.      1.287127
75%          2.5           2.5
90%          2.5           2.5      Variance       1.656695
95%            4             4      Skewness       1.136863
99%            5             5      Kurtosis       3.884319
```

Figure 29: The distribution of the within-subject mean and median absolute deviation of threshold 3, excluding jumpers

```
                          dif pooled
-------------------------------------------------------------
      Percentiles      Smallest
 1%           -7             -7
 5%           -3             -3
10%           -2             -2       Obs                  22
25%            0             -2       Sum of Wgt.          22

50%            5                      Mean           4.136364
                        Largest       Std. Dev.      5.800619
75%            7             10
90%           12             12       Variance       33.64719
95%           14             14       Skewness       .2417034
99%           16             16       Kurtosis       2.510263

                          dif Short
-------------------------------------------------------------
      Percentiles      Smallest
 1%           -7             -7
 5%           -7             -3
10%           -3              0       Obs                  13
25%            0              0       Sum of Wgt.          13

50%            2                      Mean           3.923077
                        Largest       Std. Dev.      6.550592
75%            9              9
90%           12             10       Variance       42.91026
95%           16             12       Skewness       .2400188
99%           16             16       Kurtosis       2.210651

                          dif Long
-------------------------------------------------------------
      Percentiles      Smallest
 1%           -2             -2
 5%           -2             -2
10%           -2              2       Obs                   9
25%            2              5       Sum of Wgt.           9

50%            5                      Mean           4.444444
                        Largest       Std. Dev.      4.876246
75%            6              5
90%           14              6       Variance       23.77778
95%           14              7       Skewness       .3928779
99%           14             14       Kurtosis       2.952716
```

Figure 30: The distribution of the within-subject differences in threshold 3 between rounds 3 and 11 for the pooled sample, Short treatment and Long treatment, excluding jumpers

dif pooled

Percentiles		Smallest		
1%	-10	-10		
5%	-7	-7		
10%	-6	-6	Obs	22
25%	-3	-3	Sum of Wgt.	22
50%	.5		Mean	1.954545
		Largest	Std. Dev.	6.69383
75%	6	9		
90%	11	11	Variance	44.80736
95%	14	14	Skewness	.4301935
99%	16	16	Kurtosis	2.6041

dif Short

Percentiles		Smallest		
1%	-10	-10		
5%	-10	-7		
10%	-7	-6	Obs	13
25%	-3	-3	Sum of Wgt.	13
50%	-1		Mean	.3846154
		Largest	Std. Dev.	7.467022
75%	2	2		
90%	14	3	Variance	55.75641
95%	16	14	Skewness	.9350493
99%	16	16	Kurtosis	3.206321

dif Long

Percentiles		Smallest		
1%	-3	-3		
5%	-3	-1		
10%	-3	0	Obs	9
25%	0	3	Sum of Wgt.	9
50%	4		Mean	4.222222
		Largest	Std. Dev.	4.918785
75%	9	6		
90%	11	9	Variance	24.19444
95%	11	9	Skewness	-.0681096
99%	11	11	Kurtosis	1.668886

Figure 31: The distribution of the within-subject differences in threshold 3 between rounds 4 and 20 for the pooled sample, Short treatment and Long treatment, excluding jumpers

dif pooled

Percentiles		Smallest		
1%	-10	-10		
5%	-6	-6		
10%	-3	-3	Obs	23
25%	-1	-2	Sum of Wgt.	23
50%	0		Mean	.173913
		Largest	Std. Dev.	3.53749
75%	2	3		
90%	3	3	Variance	12.51383
95%	4	4	Skewness	-.7577674
99%	8	8	Kurtosis	5.009554

dif Short

Percentiles		Smallest		
1%	-10	-10		
5%	-10	-3		
10%	-3	-1	Obs	15
25%	-1	-1	Sum of Wgt.	15
50%	0		Mean	.4
		Largest	Std. Dev.	3.850788
75%	2	2		
90%	4	3	Variance	14.82857
95%	8	4	Skewness	-.8520857
99%	8	8	Kurtosis	5.361038

dif Long

Percentiles		Smallest		
1%	-6	-6		
5%	-6	-2		
10%	-6	-2	Obs	8
25%	-2	0	Sum of Wgt.	8
50%	0		Mean	-.25
		Largest	Std. Dev.	3.058945
75%	2.5	0		
90%	3	2	Variance	9.357143
95%	3	3	Skewness	-.6442734
99%	3	3	Kurtosis	2.537206

Figure 32: The distribution of the within-subject differences in threshold 3 between rounds 11 and 12 for the pooled sample, Short treatment and Long treatment, excluding jumpers

dif pooled

Percentiles		Smallest		
1%	-8	-8		
5%	-6	-6		
10%	-2.5	-3	Obs	30
25%	-1	-2	Sum of Wgt.	30
50%	0		Mean	-.0333333
		Largest	Std. Dev.	2.592873
75%	2	3		
90%	3	3	Variance	6.722989
95%	4	4	Skewness	-.9647409
99%	5	5	Kurtosis	5.169236

dif Short

Percentiles		Smallest		
1%	-8	-8		
5%	-8	-6		
10%	-6	0	Obs	15
25%	0	0	Sum of Wgt.	15
50%	0		Mean	-.2666667
		Largest	Std. Dev.	3.011091
75%	2	2		
90%	2	2	Variance	9.066667
95%	4	2	Skewness	-1.457474
99%	4	4	Kurtosis	4.71794

dif Long

Percentiles		Smallest		
1%	-3	-3		
5%	-3	-2		
10%	-2	-2	Obs	15
25%	-1	-1	Sum of Wgt.	15
50%	0		Mean	.2
		Largest	Std. Dev.	2.17781
75%	2	2		
90%	3	3	Variance	4.742857
95%	5	3	Skewness	.7232473
99%	5	5	Kurtosis	2.797286

Figure 33: The distribution of the within-subject differences in threshold 3 between rounds 19 and 20 for the pooled sample, Short treatment and Long treatment, excluding jumpers

APPENDIX: EXPERIMENTAL INSTRUCTIONS

Instructions

Welcome and thanks for participating in this experiment. In this experiment you can earn a certain amount of money, which depends on your decisions. **So it is very important that you read these instructions carefully. You should not be irritated by the length of the instructions – most of them consist of examples.** If you have a question, please raise your hand. We will then come to you and answer your question individually.

Please note that these instructions are only meant for you, and that you are not allowed to share any information with other participants. Similarly, it is not allowed to talk to other participants during the entire experiment. If you have a question, please raise your hand. We will then come to you and answer your question individually. Please do not ask your questions aloud. If you do not follow these rules, we have to stop the experiment. Please also turn off your mobile phones now.

1. Procedure

1.1. Your task

Imagine you have a deck of 40 cards. A number is printed on each card. This means that overall there are cards with numbers 1 to 40. The cards are well shuffled, i.e., any order of their lying on the deck is possible.

You draw the cards one after another. This means that you always see only the one card you are drawing from the deck. Cards are put on the deck with the numbers facing down. You cannot turn the cards to see their numbers. You only see the current rank of each card. The concept of the "current rank" is explained below.

It is your task to find a card with a number as small as possible. The smaller the number, the higher is your payoff. How you determine your card is also explained below.

The example of the cards serves only as illustration. In the experiment, you don't draw real cards from a deck. However, the real procedure in the experiment, which is played on the computer, is very similar to the example. You can refer to it for clarification.

Before the experiment starts, the computer "shuffles" the cards via its random generator. This means that any order of the cards is possible. With each draw you see exactly one card. This procedure continues until you choose a card. This card will then be your card and the search ends. In the experiment, a draw is called "period". For how many periods you proceed therefore depends on when you terminate your search.

When you draw a card, the computer does not show you its number. It only shows you the current rank of this card, relative to the cards that were drawn before.

In the following, we first explain in detail how the current rank of a card is determined. Afterwards, we explain how you choose your card.

1.2. Current rank of a card

Imagine that so far you have drawn five cards. These cards had the numbers 30, 18, 26, 7 and 33. According to these numbers, the cards can be assigned ranks. We start with the smallest number: The card with the number 7 ("card 7" in what follows) has rank 1, card 18 has rank 2, card 26 has rank 3, card 30 has rank 4 and card 33 has rank 5. These ranks will be called "current ranks", since they refer only to the comparison with the other four cards that were drawn so far, i.e., the "current" inventory of cards.

The following table shows the screen that would appear if the cards were indeed drawn in the order of the above example. On the left you see the table with the full information, i.e., including the numbers of the cards (shaded column). However, in the experiment you only see the table on the right. The numbers of the cards are only shown once you have chosen your card, i.e., once the round has ended. (Note that the table in the experiment includes 40 periods, which are filled with ranks as you are searching. The five periods listed here only serve as illustration.)

Period	Current rank	Card number
1	4	30
2	2	18
3	3	26
4	1	7
5	5	33

This would be the complete table.

Period	Current Rank	Card number
1	4	
2	2	
3	3	
4	1	
5	5	

This is the table you are shown.

The first column of the table shows the period, i.e., how many cards you have drawn already. This means that in the example you are in period 5. The column in the middle shows the current ranks in of all cards drawn so far, for a particular period (period 5 here). The right column shows the numbers of the cards as soon as the round has ended (numbers of 40 cards in the experiment). This column is empty as long as you haven't chosen a card yet.

The current ranks change depending on which cards have already been drawn. For illustration, consider again the above example.

Imagine you are in period 1 and draw the first card. Only card 30 is visible on the screen. Since there are no other cards to compare to, the current rank of this card is 1. This is shown in the following tables:

Period	Current rank	Card number
1	1	30
2		
3		
4		
5		

This would be the complete table.

Periode	Current rank	Card number
1	1	
2		
3		
4		
5		

This is the table you are shown.

.

Imagine now that you are in period 2. This means that cards 30 and 18 have been drawn. Since 18 is lower than 30, card 18 has the current rank 1, while card 30 has the current rank 2. This is shown in the next tables:

Periode	Current rank	Card number
1	2	30
2	1	18
3		
4		
5		

This would be the complete table.

Periode	Current rank	Card number
1	2	
2	1	
3		
4		
5		

This is the table you are shown.

Now imagine that you are in period 3. This means that cards 30, 18 and 26 have been drawn. Since 18 is lower than 26, and 26 is lower than 30, card 18 has the current rank 1, card 26 has the current rank 2 and card 30 has the current rank 3. This is shown in the next tables:

Periode	Current rank	Card number
1	3	30
2	1	18
3	2	26
4		
5		

This would be the complete table.

Periode	Current rank	Card number
1	3	
2	1	
3	2	
4		
5		

This is the table you are shown.

As you can see, in period 3 the current rank of card 30 worsened by one. The current rank of card 18 remained the same. This is due to card 18 keeping its current rank of 1, since 18 is lower than 26. In contrast, card 30 is higher than card 26, which made it lose one rank.

Please fill in the tables of periods 4 and 5 yourself. If you have questions or feel uncertain about what to write, please raise your hand and we will come to you.
When you have filled in the tables, please continue with the instructions. Before the experiment starts we will come and check whether you filled in the current ranks correctly.

Period	Current rank	Card number
1	…	30
2	…	18
3	…	26
4	…	7

Complete table in period 4.

Period	Current rank	Card number
1	…	30
2	…	18
3	…	26
4	…	7
5	…	33

Complete table in period 5.

1.3. Choosing your card

„Your“ card, that is, the card that determines your payoff, is chosen as follows. At the beginning of each period, i.e., before a card is drawn for this period, you see the following item on the screen:

I want to choose the next card if its current rank is not higher than: 

The current rank that you state there is called “maximum rank”. It has the following meaning, which is also explained in the example below. If your maximum rank is **higher or as high** as the current rank of the card that is drawn in this period, this card becomes your card. The search ends at this point, no further cards are drawn.
If your maximum rank is **lower** than the current rank of the card that is drawn in this period, this card does **not** become your card. Instead, the search continues, that is, a new card is drawn. If you haven’t chosen a card until the end of period 40, i.e., your maximum rank was always lower than the current rank of the next card, card 40 automatically becomes your card.

Again the procedure in each period as an overview:

1. You state your maximum rank. (“I want to choose the next card if its current rank is not higher than: “)
2. The computer draws the next card, i.e., the card for this period.

3a. If your maximum rank is **higher or as high** as the current rank of this card, this card becomes your card. **The search ends**.

3b. If your maximum rank is lower than the current rank of this card, this period’s card does not become your card. You cannot choose it later on either. **The search continues**.

The way your maximum rank works can be illustrated using the above example. Imagine you are at the beginning of period 3. On the screen you see the following table. Since the current rank of the card of this period is not yet known, you see a question mark in place of it.

Period	Current rank	Card number
1	2	
2	1	
3	?	
4		
5		

At this point you will be asked for you maximum rank for this period. If, for example, you state “2“ as your maximum rank, this means that you choose the card of period 3 if its current rank is either 1 or 2. You do not choose this card if its current rank is 3.
If you state “1“ as your maximum rank, this means that you choose the card of period 3 only if its current rank is 1. Finally, you can state “0“ as your maximum rank. With this you ensure that you do not choose the card of period 3, no matter which current rank it has. This guarantees that the search continues.

After you have stated your maximum rank the card for this period is drawn and the current ranks of all cards drawn so far are updated in the table:

Period	Current rank	Card number
1	3	
2	1	
3	2	
4		
5		

Since in the example the current rank of the card drawn in period 3 is 2, you would have chosen this card if you stated either 2 or 3 as your maximum rank. In the case, the search would end here. The number of this card would be 26 (see above). The number 26 would then be relevant for your payoff. How this payoff is determined is explained in detail below. If, in contrast, you stated 0 or 1 as your maximum rank, the search would continue in period 4.

Summarizing, the following facts hold for your maximum rank:

1. Your maximum rank always refers to the **next card**, that is, the card thats current rank you **do not see yet**.
2. If the card that is drawn in this period has a current rank that is lower or as high as your maximum rank, the **search ends irrevocably**. This card then becomes your card, i.e., it determines your payoff.
3. If the card that is drawn in this period has a current rank that is higher than your maximum rank, the **search continues irrevocably**. You can not reverse your decision afterwards and return to a card you missed. Therefore, in each period you should carefully consider which maximum rank to state.
4. You can set your **maximum rank to 0** and with this **ensure that the search does not end** in this period.
5. You can set a different maximum rank in each period.
6. If you did not choose a card until period 40, the 40th card automatically becomes your card.

1.4. Display of card numbers

Once you have chosen your card and the search has ended, the numbers of all cards are displayed. This includes the cards that you draw before your card as well as those that would have come after your card.

2. Your payment from the experiment

2.1. General procedure

The search for the card with the lowest number that was explained in section 1 is repeated multiple times during the experiment. First you play three practice rounds, which are not relevant for your payment from the experiment. Then you play 21 rounds that are relevant for your payment from the experiment.

1. First you play three practice rounds.
2. Then you play 21 "real" rounds. In sum there are 3 + 21 rounds.
3. In each round there are 40 cards. Depending on how long you search, you therefore play up to 40 periods (40 cards) in each of the 3 + 21 rounds.
4. The procedure is the same in all of the 3 + 21 rounds (see above). It is always your task to find a card with a number as low as possible. The order in which the cards are drawn (in which they lie on the desk) is **determined randomly for each period**.

2.3. Payoff per round

Your payoff per round depends on the number of your card. The lower the number of your card, the higher your payoff. The details of the allocation of payoffs to card numbers are shown in the table:

The number of your card	**Your payoff in this round**
1	15 Euro
2	10 Euro
3	5 Euro
4-40	0 Euro

2.2. Payment

After you have played all the 21 rounds that are relevant for your payment, your payoffs from all rounds are shown on the screen. Afterwards you twice draw a number from a bowl with the numbers 1-21. The sum of the payoffs from the two rounds that you draw forms your payment. In addition, you receive 2.50 Euro show-up fee.

2.3. Questionnaire and end of the experiment

After filling in a questionnaire the experiment ends and you receive your payment. If you wish, you can afterwards check the order of the 40 cards for each of the 3 + 21 rounds.

3. Questions (The examples included here should not be understood as hints!):

1. If in a certain period you state a maximum rank of 2 ...
   - o … the search stops if the next card has a current rank of 1 or 2.
   - o … the search stops only if the next card is card 2.
   - o … the search stops if the previous card had the current rank 1 or 2.
   - o … the search continues for sure.

2. If in a certain period you state a maximum rank of 2 ...
   - o … the search stops immediately and the first card becomes your card.
   - o … the search stops only if the first card is card 1.
   - o … the search continues for sure.
   - o … one cannot know what happens.

3. If the maximum rank that you stated in a certain period is below the current rank of this period's card…
   - o .. I can choose this card afterwards.
   - o .. I have the chance that the same card is drawn again later on.
   - o .. I can be sure that a card with a smaller number will be drawn later on.
   - o .. I cannot choose this card afterwards.

4. If the maximum rank that you stated in a certain period is above the current rank of this period's card…
   - o .. the search stops and this period's card becomes my card.
   - o .. I can choose whether the search stops or continues.
   - o .. this period's card becomes my card and the search continues.
   - o .. the search continues automatically.

5. In each round you receive a payoff according to your card. How is your final payment from the experiment determined?
   - o … the average of the payoffs of all rounds.
   - o … the sum of the payoffs of all rounds
   - o … the sum of the payoffs of two rounds which I can choose.
   - o … the sum of the payoffs of two rounds which I draw randomly.

6. If in period 20 you draw a card of current rank 1, this means that …
   - o … this card has the number 1.
   - o … this card has a number smaller than or equal to 21.
   - o … for this card you get a payoff above zero for sure.
   - o … this card has a number smaller than or equal to 11.

7. If in period 10 you draw a card of current rank 3, this means that …
   - o … this card has the number 3.
   - o … this card has a number smaller than or equal to 33.
   - o … for this card you get a payoff above zero for sure.
   - o … this card has a number smaller than or equal to 27.