Forthcoming in *Operations Research*

The Gatekeeper's Dilemma: "When Should I Transfer This Customer?"

Brett Hathaway
Carey Business School, Johns Hopkins University, Baltimore, MD 21202, bhathaw2@jhu.edu

Evgeny Kagan
Carey Business School, Johns Hopkins University, Baltimore, MD 21202, ekagan@jhu.edu

Maqbool Dada
Carey Business School, Johns Hopkins University, Baltimore, MD 21202, dada@jhu.edu

In many service encounters front-line workers (often referred to as gatekeepers) have the discretion to attempt to resolve a customer request or to transfer the customer to an expert service provider. Motivated by an incentive redesign at a call center of a mid-size US-based bank, we formulate and solve an analytical model of the gatekeeper's transfer response to different incentive schemes and to different congestion levels. We then test several model predictions experimentally. Our experiments show that human behavior matches the predictions qualitatively, but not always in magnitude. Specifically, transfer rates are disproportionately low in the presence of monetary penalties for transferring, even after controlling for the economic (dis)incentive to transfer, suggesting an overreaction to transfer cost. In contrast, the transfer response to congestion information shows no systematic bias. Taken together, these results advance our understanding of cognitive capabilities and rationality limits on human server behavior in queueing systems.

Key words: decision-making, behavior in queueing systems, service operations, incentive design

History: **This version: October 5, 2021**

1. Introduction

The delivery of services often consists of a complex series of service encounters, during which the customer may go through a number of steps, interacting with several specialized workers (Gans et al. 2003, Sampson and Froehle 2006, Heineke and Davis 2007). In many of these encounters front-line workers (often referred to as gatekeepers) have the discretion to attempt to resolve a customer request single-handedly, or to transfer the customer to an expert. This attempt-or-transfer

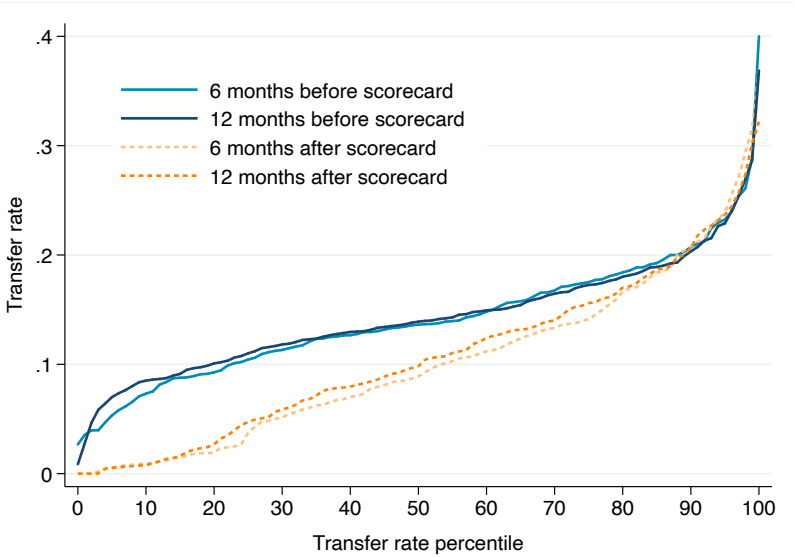
decision is at the core of many customer/worker interactions in call centers, ticket-based work in Information Technology (IT), patient screenings in health care, and in other settings where the solution to the problem cannot be identified prior to the encounter.

Transfers can have important service quality and cost implications. On the one hand, transfers may detract from customer satisfaction, and may increase staffing costs. On the other hand, transfers may be necessary for more complicated requests that cannot be solved within a reasonable timeframe, for example, due to a lack of technical proficiency of the gatekeeper. As more requests accumulate in the queue, waiting to be resolved, failure to make timely transfers can further exacerbate congestion, extending the time spent in the system.

Consider the case of the ABC (name redacted) Bank call center, which provides customer service support to approximately 3 million consumers and small business owners. In late 2017 the call center management team decided to re-examine several service performance indicators, including discretionary transfers, i.e., transfers that call center agents initiate after determining that they have spent sufficient time (or expended sufficient effort) working on a request. As part of their investigation, the management team discovered what they perceived as high variation in transfer rates. For example, for Online Banking access requests, individual rates ranged between 3.8% and 36.9% – a range wide enough to suggest that agents could change their transfer behavior in response to an appropriately-chosen incentive system. (See Hathaway et al. (2021) for more details of this field study.)

On 4/1/2018, the management team implemented such an incentive system (referred to internally as the new "agent scorecard"). The scorecard was used to measure agents' performance and link it to their monetary bonuses and shift assignments. Among other components (mainly related to customer satisfaction), the scorecard included a *productivity score*, with transfer decisions contributing substantially to it. In essence, the management team designed this productivity score to encourage agents to quickly transfer issues requiring above-average service time, and discourage them from transferring issues requiring short service time.

Figure 1 Transfer Rates Before and After the Scorecard Rollout



Note. Transfer rate percentiles for the 201 (234) agents who handled at least 100 Online Banking access calls six (twelve) months prior to the scorecard change, and the 216 (266) agents who handled at least 100 Online Banking access calls six (twelve) months following the scorecard change.

To explore the effects of the new scorecard on transfer behavior we examined agent transfer rates (the share of incoming calls that an agent transfers) for Online Banking access requests before and after its implementation. Figure 1 plots the transfer rate percentiles for agents who handled at least 100 Online Banking access calls prior to the scorecard rollout (solid lines), and agents who handled at least 100 Online Banking access calls following the scorecard rollout (dashed lines). With the exception of the transfer rates above the 85th percentile, transfer rates dropped substantially. For example, the median transfer rate dropped from 13.6% (13.9%) to 8.9% (9.8%) in the six (twelve) months following the scorecard rollout.

The ABC bank call center case demonstrates that transfers can be an important concern for managers of service systems and that they are willing to explore ways to align transfer behavior with what they perceive as appropriate. However, designing an incentive system that effectively regulates transfers in a queueing system requires a thorough understanding of gatekeeper responses to different components of the incentive system under different congestion levels, and potential biases affecting those responses. Do gatekeepers transfer more when the time for resolving a request

goes up? Are they able to incorporate into their behavior a reward for solving a customer request and a cost for transferring? Finally, are they able to combine congestion information with incentive and resolution time information to arrive at the right transfer decision?

In this paper we study these questions using analytical modeling and experiments. We use modeling to develop benchmarks for optimal gatekeeper transfer behavior under different incentive systems and congestion levels. We then use experiments to test model predictions in a controlled setting. Both the model and the experiment design are inspired by our interactions with the ABC bank managers and call center agents, and by the insights from the data they shared with us. Thus, our approach draws on the distinct strengths of field and lab data to develop queueing models that incorporate human behavior (Schultz et al. 1998, Ülkü et al. 2019).

We model the work shift of a gatekeeper as a series of customer requests. Each request may require several attempts to be resolved, making the exact number of attempts until successful resolution uncertain. Within a request, each attempt may result in successful resolution, but also depletes the time budget (a proxy for shift duration). Further, the gatekeeper receives a reward for each successfully resolved request and may be penalized with a cost for transferring. Hence, each request resembles a stopping problem, where the gatekeeper faces a trade-off between immediate losses due to a transfer and future losses due to fewer requests handled. We obtain the gatekeeper's optimal decision policy by solving the above finite-horizon problem using dynamic programming.

To develop testable hypotheses regarding human gatekeeper behavior, we embed the dynamic program solutions into a random utility framework, which allows the possibility of random errors, particularly for decisions with small payoff consequences. We then test these hypotheses in two experiments. In the first experiment, participants solve the problem under one of two different incentive schemes: one that rewards each resolved request with a fixed bonus, and one that additionally imposes a transfer cost. In the second experiment, participants solve the problem in a system with a variable queue state (i.e., sometimes empty and sometimes nonempty).

Our experimental design and econometric approach carefully control for the payoff difference between the optimal and non-optimal actions across treatment conditions. The higher the payoff

difference, the more subjects benefit from being able to identify the correct transfer decision; thus, holding the payoff difference constant helps separate behaviors caused by the *strength* of the incentives from behaviors caused by the *structure* of the incentives, exposing robust biases in decision-making (Harrison 1989, Smith and Walker 1993). To be able to use this approach in a dynamic setting, we carefully tailor terminal conditions such that our model and experiments have a finite horizon and a stationary optimal policy. In addition to providing a clean benchmark for hypothesis tests, stationarity and finiteness remove the need for time discounting, facilitating participant comprehension and convergence to stable strategies.

Our experimental results are as follows: 1) Consistent with our hypotheses, transfer rates are lower when a transfer cost is present. However, contrary to our hypotheses, transfer rates in the presence of a cost are lower even after controlling for the incentive differences between the pure bonus and the bonus+cost incentive regimes. That is, gatekeepers appear to overreact to cost by disproportionately reducing transfer rates. 2) Consistent with our hypotheses, transfer rates increase when the queue is nonempty. However, the queue state does *not* lead to systematic over/under transferring, suggesting no systematic biases in response to varying congestion levels.

Taken together, these results advance our understanding of cognitive capabilities and rationality limits on human server behavior in queueing systems. With simple bonus-based systems, transfer decisions are noisy but exhibit no bias towards or away from transferring, even with variable congestion. In contrast, with bonus+cost systems, decisions are biased away from transferring. The implications of these results in practice depend on the desirability of transfers by the customers and by the organization operating the gatekeeper system.

2. Literature

Our investigation draws on and contributes to two streams of literature: (1) gatekeeper literature in service and healthcare operations, and (2) work on queueing systems with human servers.

Gatekeepers The term *gatekeeper* has been used in two types of operational environments. One is a system where the gatekeeper's job is to sort people or tasks based on their characteristics,

e.g., a triage nurse screening patients in an emergency room. The other is a system in which there is an overlap between the tasks completed by the gatekeeper and by an expert, with the gatekeeper deciding to either handle each task single-handedly or transfer it to the expert. Our investigation focuses on the latter system.

Gatekeeper-expert systems have been first studied analytically using the principal-agent framework (Shumsky and Pinker 2003, Hasija et al. 2005, Lee et al. 2012). Recent empirical studies have examined the effects of congestion on gatekeeper decisions in health care systems. Freeman et al. (2017) show that workload (induced by congestion) affects midwives' referral (transfer) behavior to a physician, which creates variation in the patients' quality of care. Batt and Terwiesch (2017) show that under high congestion, nurses (gatekeepers) take over a portion of the physician's work to reduce waiting times. We contribute to this stream by conducting the first (to our knowledge) experimental test of an analytical model of transfer decisions.

Queueing Systems with Human Servers Although some early work in queueing theory has considered discretionary server decisions (for example, see Edie 1954), there is a renewed interest in research that explicitly incorporates server behavior. Schultz et al. (1998) show experimentally that workers change their service speed in response to the amount of work to be done. Other studies in this vein include analytical studies of servers with discretionary speed (George and Harrison 2001), quality (Hopp et al. 2007), speed and quality (Zhan and Ward 2019), and field studies of servers with discretionary speed (Oliva and Sterman 2001, Berry Jaeker and Tucker 2017), and task selection (KC et al. 2020, Ibanez et al. 2018). Finally, an integrative review of the literature on the effects of load on service times is offered in Delasay et al. (2019).

While the extant literature documents a variety of settings in which servers have discretion over service provision, few studies examine the internal decision trade-offs facing the server. Indeed, a recent review of the behavioral queueing literature (Allon and Kremer 2018) suggests that "[...] most of the empirical evidence comes from field settings, which leaves substantial room for deeper investigations of individual-level server behavior under controlled laboratory conditions (page 357)".

Two recent exceptions are found in Shunko et al. (2018) and Rosokha and Wei (2020), who study effort provision in multi-server systems.

Finally, we represent the gatekeeper's work shift as a series of stopping problems, where each transfer is equivalent to a stop. Hence, our work is related to optimal-stopping experiments (Rapoport and Tversky 1970, Cox and Oaxaca 1989, Seale and Rapoport 1997, Bearden et al. 2006, Long et al. 2019, Kremer and de Vericourt 2020), to sequential allocation experiments (Bearden et al. 2008, Leider and Şahin 2014), and more generally to dynamic programming experiments (Hey and Dardanoni 1998, Noussair and Matheny 2000, and references on page 5 of Duffy (2016)), as well as other experiments that use terminal conditions to induce stationarity (Ball and Holt 1998, Noussair et al. 2001, Kirchler et al. 2012). We note that server decisions in our queueing context have both time and payoff implications resulting in decision cycles of varying length, which significantly complicates the construction of terminal conditions (see Hathaway et al. 2021 for details).

3. Model of Transfer Behavior

To provide optimality benchmarks for transfer decision-making we will next develop a discrete-time, finite-horizon model of a single gatekeeper who makes transfer decisions in response to an incentive system over a series of service requests received within a work shift. We first describe the gatekeeper's transfer decision problem and our approach for deriving properties of the gatekeeper's optimal policy. We then focus on a basic (two-attempt) version of this problem that retains the key decision trade-offs, and that is used in our subsequent laboratory experiments. We then embed the model into a random utility framework to account for random errors of human decision-makers. This section has been abridged. See Hathaway et al. (2021) for a longer version appropriate for readers interested in the model analysis, proofs, and extensions.

3.1. The Gatekeeper's Decision Problem

We model the gatekeeper's work shift as a sequence of discrete periods indexed by $t \in \{1, \cdots, T\}$, where T is the duration of the work shift. We model congestion by positing that at time t there

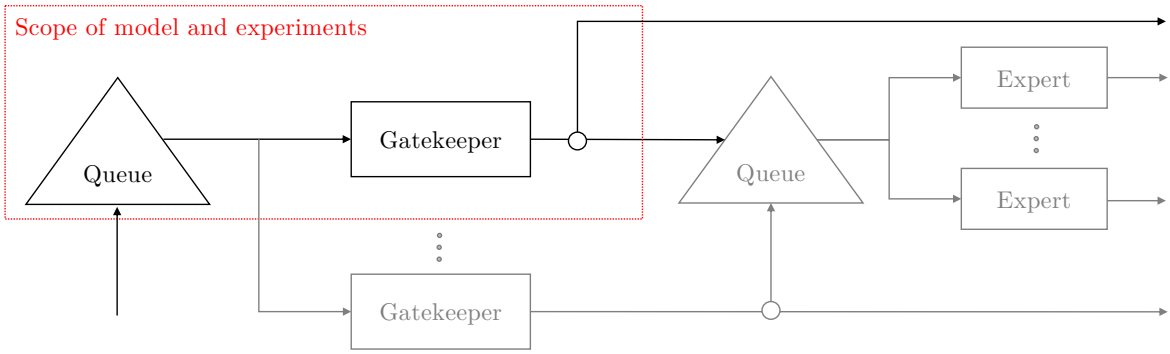
is a stationary probability, q, that there is at least one service request waiting in queue to be handled, and a complementary probability, $1-q$, that the queue is empty. Consequently, at time t, if the gatekeeper is idle or becomes available and the queue is nonempty, then the gatekeeper immediately begins handling a request; otherwise, the gatekeeper is idle until at least period $t+1$.

Each request is characterized by an ordered list of potential solutions $s=\{1,2,...,S\}$, by probabilities, p_s, that solution s succeeds in resolving the request, and by the handling times, τ_s, that the gatekeeper expends to attempt each potential solution. The gatekeeper knows ex ante that one of the solutions is guaranteed to resolve the issue, such that $\sum_{s=1}^{S} p_s = 1$. The choice to guarantee that one of the attempts will resolve the request was made to simplify the problem for experimental participants, but is not required for our analytical results. While the gatekeeper does not know whether potential solution s will resolve the request, the gatekeeper knows the probability of each potential solution resolving the request (p_s), and its respective handling time (τ_s).

At the beginning of each encounter, the gatekeeper attempts the first potential solution ($s=1$), expending τ_1 time units. If the attempt resolves the request, the gatekeeper receives a reward r. If not, then the gatekeeper chooses whether to attempt the second potential solution ($s=2$), expending an additional τ_2 time units, or transfer the request at cost c. This attempt-or-transfer decision is repeated until the request is resolved and the gatekeeper receives the reward, or the request is transferred and the gatekeeper incurs the cost. Upon resolving or transferring a service request, the gatekeeper begins handling the next request, immediately if the queue is nonempty, or as soon as one becomes available if the queue is empty. (The gatekeeper knows the state of the queue when making transfer decisions).

The scope of the studied problem is summarized in Figure 2. We focus on a single gatekeeper whose transfer decisions are informed by current congestion. We model the queue as having a binary state (empty/nonempty) to make the problem more amenable to a laboratory study of behavioral responses to congestion. Further, the stationary queue-state distribution characterized by the fixed, exogenous q parameter means that the gatekeeper is not concerned with how current

Figure 2 Scope of Study



transfer decisions impact future congestion. This is consistent with large-scale systems such as call centers (including the ABC Bank call center described in §1), technical support, and other settings in which individual server decisions have a negligible impact on load (See for example Garnett et al. 2002, Dong et al. 2015, for steady-state analyses of such large-scale service systems). Lastly, the sequential nature of the resolution process is reflective of more standardized service processes for which the problem categories and the attendant resolution procedures are mapped out and reasonably well understood by the organization (τ_s and p_s are fixed and known). In contrast, smaller service systems, systems in which gatekeepers may need to collaborate (with other gatekeepers, or with experts) to resolve a problem, or systems in which problems are not easily categorized, would be excluded.

3.2. Two-Attempt Problem

In Hathaway et al. (2021) we formulate the above problem as a finite-horizon dynamic program. Using inductive arguments we show the following for a general S-attempt problem: 1) for any given period, if it is optimal for the gatekeeper to transfer when the queue is empty, then it is optimal to transfer when it is nonempty; and that 2) under an intuitively appealing set of terminal conditions a stationary policy is optimal. The optimality of a stationary policy in conjunction with a finite horizon and intuitive terminal conditions removes the need for discounting or other time-based adjustments to the decision. This significantly simplifies the decision situation and makes the model more amenable to experimentation with human subjects.

Our experimental investigation will focus on the two-attempt version of the gatekeeper's decision problem ($S = 2$). While muting certain aspects of the decision, the two-attempt version still captures the key determinants of the transfer decision, specifically the effects of resolution times, incentive structures, and current congestion levels on transfer behavior.

When $S = 2$, a direct analysis yields an analytic expression for the optimal expected profit per unit time. Consider the gatekeeper's problem after attempting potential solution 1. If the problem has been resolved, the gatekeeper has no decision to make: if the queue is empty, the gatekeeper waits until the next request arrives; otherwise, the queue is nonempty and the gatekeeper begins handling the next request. If potential solution 1 does not resolve the request, then the gatekeeper must decide whether to attempt potential solution 2 or transfer and move on to the next request. This decision is simplified by the stationarity of the optimal policy and can thus be made through a direct comparison of the expressions corresponding to each stationary policy. Specifically, the optimal policy is one of three admissible policies: 1) "Always Transfer", i.e., transfer irrespective of the queue state, 2) "Always Continue", i.e., attempt potential solution 2 irrespective of the queue state, and 3) "Transfer When Nonempty", i.e., only transfer when the queue is nonempty. Note that the fourth possible stationary policy in which the gatekeeper only transfers when the queue is empty is dominated by either "Always Transfer" or "Always Continue".

We will denote by $\boldsymbol{\theta} = \{r, c, \tau_1, \tau_2, p_1, p_2, q\}$ the vector of model parameters, and by $R(\cdot, \cdot, \boldsymbol{\theta})$ the expected profit per unit time given $\boldsymbol{\theta}$, where the first (second) argument is the transfer decision when the queue is empty (nonempty). We use "1" to denote "Transfer" and "2" to denote "Continue". Thus, $R(1, 1, \boldsymbol{\theta})$ denotes the expected profit per unit time of following the "Always Transfer" policy, $R(2, 2, \boldsymbol{\theta})$ denotes that of the "Always Continue" policy, and $R(2, 1, \boldsymbol{\theta})$ denotes that of the "Transfer When Nonempty" policy. Solving the recursions of the dynamic program under each of these three policies yields the following result:

PROPOSITION 1. *Under terminal conditions that lead to a stationary threshold policy being optimal, when $S = 2$, the optimal profit per unit time is given by the maximum of $R(1, 1, \boldsymbol{\theta})$, $R(2, 2, \boldsymbol{\theta})$, and $R(2, 1, \boldsymbol{\theta})$, where*

$$R(1,1,\boldsymbol{\theta}) = \frac{p_1 r - (1-p_1)c}{\tau_1 + (1-q)/q}, \qquad R(2,2,\boldsymbol{\theta}) = \frac{r}{\tau_1 + (1-p_1)\tau_2 + (1-q)/q} \quad \text{and}$$

$$R(2,1,\boldsymbol{\theta}) = \frac{r(1-q+p_1 q) \; - \; (1-p_1)qc}{\tau_1 + (1-p_1)(1-q)\tau_2 + [1-q(1-p_1)](1-q)/q}.$$

While the formal proof of Proposition 1, and the derivation of the terminal conditions are presented in Hathaway et al. (2021), we note that the $R(\cdot,\cdot,\boldsymbol{\theta})$ expressions have a more intuitive interpretation. Specifically, the numerator of each expression represents the expected reward per "cycle", and the denominator represents the expected "cycle time". Therefore, the $R(\cdot,\cdot,\boldsymbol{\theta})$ expressions can be interpreted as the "expected profits per unit time" resulting from each stationary policy.

3.3. Random Utility Model

Proposition 1 states that the optimal transfer decision depends on the relative returns from transferring (continuing), which can be evaluated by correctly calculating the rewards per unit time. Given the random, dynamic nature of the problem, we can reasonably expect that humans will deviate from the optimum at least some of the time. To control for these deviations and expose potential systematic biases in transfer decision-making we incorporate the relative returns from transferring into a random utility model (McFadden 1973, Ben-Akiva 1973; see also Hyndman and Embrey 2018, for the usage of such models in behavioral operations research).

The basic premise of random utility models is that the choice between alternatives depends on the utility difference between them, and on a random error. When decisions are one-shot, the total utility of a decision-maker is separable into the utility increments resulting from each decision. But, when decisions are linked dynamically, as in our model, we need to make an assumption on what constitutes the utility of a decision alternative. To resolve this we will assume that decision-makers compare the expected profits per unit time under each stationary policy as derived in Proposition 1. (Alternative specifications are discussed in §6.) Specifically, we use the following ratio to assign a utility measure to the transfer decision:

DEFINITION 1. When $q = 1$, the *transfer return* $\pi(\boldsymbol{\theta})$ of a condition $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}) = \frac{R(1,1,\boldsymbol{\theta})}{R(2,2,\boldsymbol{\theta})} - 1 \tag{1}$$

In other words, the transfer return $\pi(\boldsymbol{\theta})$ measures the percentage gain (loss) in payoff per unit time incurred by transferring over continuing. The transfer return contains the following information. First, its sign indicates the optimal policy: transfer return is positive (negative) when transferring (continuing) is optimal. Second, its magnitude measures the gains (losses) of choosing the optimal over the non-optimal policy. (The construction of $\pi(\boldsymbol{\theta})$ for the $q < 1$ case is deferred to §5.)

Using the transfer return measure (eq. 1), we can formulate a random utility model, which will allow us to expose features of the decision environment that systematically affect behavior towards or away from transferring. Specifically, we will examine the effects on transfer decisions of the presence of a transfer cost (Experiment 1), and of queue state (Experiment 2).

4. Experiment 1: Different Incentive Systems

The remainder of this paper uses the two-attempt model as a benchmark for studying gatekeeper transfer decisions in controlled behavioral experiments. In Experiment 1 we set $q = 1$, and study the effects of two different incentive systems: one with a transfer cost, and one without. In Experiment 2, we set $q < 1$ and study the effects of variable congestion levels.

4.1. Hypotheses

In Experiment 1 we vary τ_2 and/or c while holding the remaining parameters constant. Comparative statics on $R(\cdot,\cdot,\boldsymbol{\theta})$ from Proposition 1 in §3.2 reveal that increasing τ_2 decreases $R(2,2,\boldsymbol{\theta})$ while $R(1,1,\boldsymbol{\theta})$ remains constant, which increases the return to transferring. Conversely, increasing c decreases $R(1,1,\boldsymbol{\theta})$ while $R(2,2,\boldsymbol{\theta})$ remains constant, which increases the return to continuing. Therefore, increasing τ_2 (c) increases (decreases) transfer return ($\pi(\boldsymbol{\theta})$), which should then make decision-makers more (less) likely to transfer, i.e., their *transfer rates* (the proportion of transferable issues they transfer) should be higher (lower). Hence:

H1a: *Holding $q = 1$ and all other parameters constant, transfer rates are increasing in τ_2.*

H1b: *Holding $q = 1$ and all other parameters constant, transfer rates are decreasing in c.*

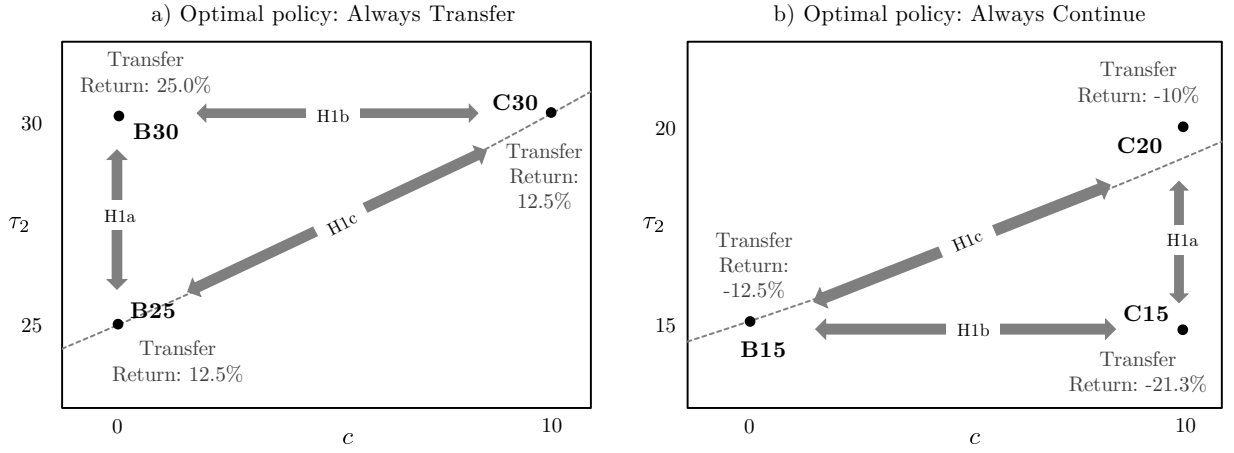
In addition to examining the behavioral response to a single parameter change (τ_2 and c), we are also interested in the effects on behavior of changing the structure of the incentive system (specifically the presence of a transfer cost). To do so, we can use our random utility framework, which suggests that any two conditions with the same transfer return should lead to the same transfer rates (since the expected utilities are the same under each parameterization). Hence:

H1c: *Holding the transfer return constant, the presence of a transfer cost has no effect on transfer rates.*

Put differently, while H1a and H1b state the null that transfer response is qualitatively consistent with the model, H1c states the null that transfer response is well-calibrated under a system with and without a transfer cost. Consequently, the null would be rejected if there is a systematic bias towards or away from transferring in the presence of a transfer cost.

Figure 3 illustrates our approach by overlaying the hypotheses with the six conditions that will be used in the experiment. In panel a) the optimal policy is "Always Transfer" and the dashed line represents various combinations of τ_2 and c that result in a transfer return of 12.5%. In panel b) the optimal policy is "Always Continue" and the dashed line represents a transfer return of -12.5%. According to H1a, going from B25 to B30 (C15 to C20) should increase transfer rates, since transfer return increases in τ_2. According to H1b, going from B30 to C30 (B15 to C15) should decrease transfer rates, since transfer return decreases in c. Finally, according to H1c, going from B25 to C30 should have no effect on transfer rates, since transfer return is the same in those conditions. Similarly, going from B15 to C20 should have only minimal effect on transfer rates. (Condition C20 features a slightly different transfer return than condition B15; the reason is that we chose all parameters to be multiples of 5 to facilitate computation for experimental subjects.)

Figure 3 Experiment 1: Summary of Hypotheses



Note. Parameters: $p_1 = p_2 = 0.5, r = 100, \tau_1 = 10$. Dashed lines represent transfer return of 12.5% in panel (a) and -12.5% in panel (b). Points correspond to the six conditions in Table 1.

4.2. Experiment Design

In our experiments, participants work on the two-attempt version of the dynamic program introduced in §3. This means that incoming requests are either resolved on the first attempt (which takes τ_1 time units) or on the second attempt (which takes τ_2 time units). Whenever the request is unresolved after the first attempt, the participant is asked to choose between transferring the request or continuing to the second attempt. For each resolved request, participants receive r points, and for each transfer they lose c points. Each round of the experiment continues until the participant runs out of the allocated T time units. At the end of the experiment the accumulated point earnings are paid out in US Dollars at the rate of 10 points = 1 cent. Instructions and screen shots of the experiment are reproduced in Appendix A.

Treatments Experiment 1 treatments and conditions are summarized in Table 1 (also see Figure 3). All conditions are between-subject. In each condition, participants receive a reward of $r = 100$ for each completed request, and each of the two attempts is equally likely to resolve the request ($p_1 = p_2 = 0.5$). However, the treatments vary in the presence of a transfer cost (c): the *Baseline* treatment has no transfer cost while the *Cost* treatment has a transfer cost of 10. Within each treatment we explore (again, using a between-subjects design) several parametrizations

Table 1 Summary of Experiment 1 Treatments and Conditions

Treatment/ Condition	# Subjects	Parametrization ($\boldsymbol{\theta}$)							Optimal Policy	Transfer Return
		r	p_1	p_2	τ_1	τ_2	c	q		
Baseline	124									
B30	43	100	0.5	0.5	10	30	0	1	Always Transfer	25.0%
B25	41	100	0.5	0.5	10	25	0	1	Always Transfer	12.5%
B15	40	100	0.5	0.5	10	15	0	1	Always Continue	-12.5%
Cost	111									
C30	34	100	0.5	0.5	10	30	10	1	Always Transfer	12.5%
C20	34	100	0.5	0.5	10	20	10	1	Always Continue	-10.0%
C15	43	100	0.5	0.5	10	15	10	1	Always Continue	-21.3%

Note: Subject numbers in column 2 exclude 49 participants who failed comprehension tests.

(conditions) by varying τ_2 while holding the remaining parameters constant. The specific τ_2 values are chosen to allow pairwise comparisons at similar levels of transfer return, and to provide a balanced treatment-average magnitude of transfer returns (see right panel of Table 1).

Participant Pool, Comprehension Checks, and Additional Measures The experiment was programmed in oTree (Chen et al. 2016) and conducted on the Amazon MTurk platform in February and September 2020. A total of 322 participants were recruited for Experiment 1 and 49 participants were excluded from the data (participants were excluded if they made more than 3 errors in the comprehension test and spent less than 10 seconds per page on the instruction pages). Upon completion of the experiment, we elicited participants' risk preferences (separately in the gain and in the loss domain) using the Eckel and Grossman (2002) measure, to control for individual differences in the regression analysis.

Time Budget and End of Horizon After completing an unincentivized training round (with $T = 100$), participants played two incentivized rounds of the experiment with a time budget of $T = 200$. The time budget was chosen such that we could collect at least 6 data points (decisions) for each participant, regardless of their transfer strategy and treatment. All participants received an initial endowment of 200 points to prevent the possibility of negative payoffs. Average time to complete the experiment was 15 minutes; average earnings was $4, including the show-up fee.

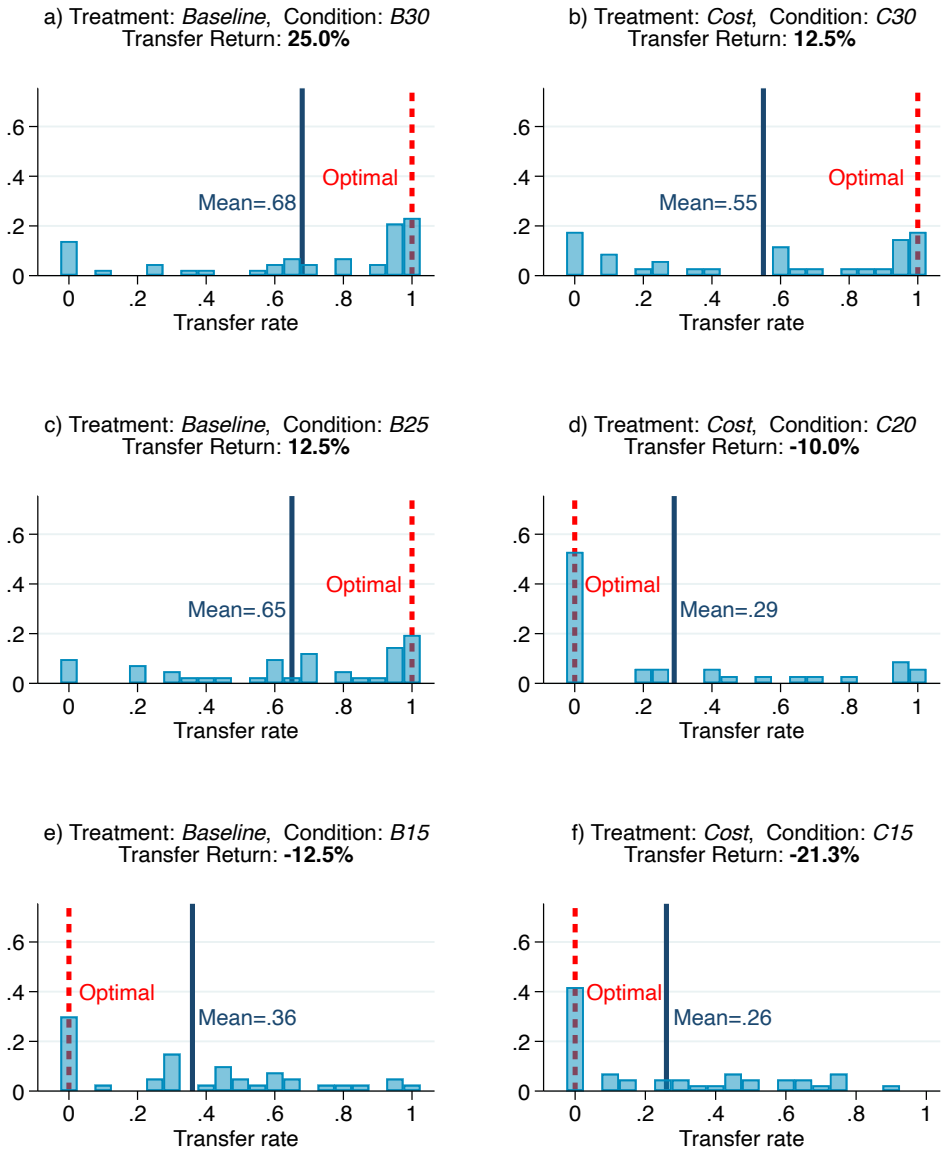
Due to the finite time budget, participants faced end-of-horizon situations in which they were left with an insufficient number of time units to resolve a new request. To ensure that end-of-horizon effects did not interact with the participants' transfer strategy, we adapted the terminal conditions developed in Hathaway et al. (2021) to the experimental setting. In particular, if at the end of the time horizon, participants had less than $\tau_1 + \tau_2$ time units remaining, they were compensated for those remaining time units at a fixed rate. Consistent with the terminal conditions presented in the proof of Proposition 1, the compensation rate was chosen to be equal to the expected return from following the optimal strategy indefinitely.

4.3. Experiment Results

Descriptive Statistics Figure 4 shows the distribution of transfer rates in each treatment and condition of Experiment 1. Each data point in the histograms is a subject average, with 0 indicating that the subject never transferred and 1 indicating that the subject always transferred. Although the data reveal a wide range of transfer behaviors, the mode in five of the six conditions is identical to the theoretical optimum, which is either 0 or 1. Further, the mean transfer rates in these five conditions are significantly different from 0.5 (two-sided $t-$tests, $p < 0.01$), providing initial evidence that participant behavior is directionally consistent with the theoretical optimum. In contrast, the mean transfer rate in the *Cost* condition in panel b) is not statistically distinguishable from 0.5 ($p = 0.43$), suggesting that participants may have difficulties incorporating a transfer cost into their decisions.

Figure 4 also reveals that mean transfer rates vary between treatments and between conditions within each treatment. In particular, mean transfer rates decrease from 0.68 to 0.36 as we vary τ_2 in the *Baseline* treatment and from 0.55 to 0.26 in the *Cost* treatment. Transfer rates also appear to be affected by the presence of a transfer cost. Indeed, when τ_2 is set to 30 in panel a) and panel b) (set to 15 in panel e) and panel f)), the mean transfer rate is 13 (10) points lower under the *Cost* treatment relative to *Baseline*.

Figure 4 *Baseline* and *Cost* Treatments: Transfer Rates



Comparing conditions with similar transfer return, cost still appears to decrease transfer rates. Specifically, mean transfer rates are 10 percentage points lower in the *Cost* treatment with a transfer return of 12.5% (panel b)) than in the *Baseline* treatment with the same transfer return (panel c), and 7 points lower in the *Cost* treatment with a transfer return of -10.0% (panel d) than in the *Baseline* treatment with a comparable transfer return of -12.5% (panel e).

Additional Condition: *C35* Our exploratory analysis thus far suggests that transfer rates drop by 7 to 10 percentage points in the presence of transfer costs, even after holding constant

the economic incentives to transfer. Before formally testing this effect, we run one additional *Cost* condition with $\tau_2 = 35$ (labeled *C35*). In the *C35* condition, participants faced similar economic incentives as in condition *B30*: both conditions have a similar magnitude of the return measure ($\pi(\boldsymbol{\theta}) = 23.8\%$ and $\pi(\boldsymbol{\theta}) = 25\%$), i.e., a relatively high payoff for following the optimal policy.

A direct comparison of transfer rates in these two conditions again shows a reduction of transfer rates in the presence of a transfer cost. However, the treatment gap decreases: the mean transfer rate in *C35* is 66%, which is only two percentage points less than in *B30*. This implies that as the return for following the optimal policy increases, identifying the correct policy becomes easier, helping participants find the optimal policy even under more complex incentive systems (such as our *Cost* treatment). Further implications of this result will be discussed in §6.

Hypothesis Tests We next test our hypotheses by examining the output of several random coefficient Logit regressions. As discussed in §3.3, our random utility model is based on decision-makers maximizing the monetary utility of transferring vs. continuing. The specific form of the utility of transferring is as follows (the utility of continuing is normalized to 0):

$$u_{ij}^{tr}(\boldsymbol{\theta}) = \alpha_i + \beta \cdot \pi(\boldsymbol{\theta}) + \boldsymbol{\gamma} \cdot \boldsymbol{x}_j + \epsilon_{ij}^{tr}, \tag{2}$$

where $u_{ij}^{tr}(\boldsymbol{\theta})$ is the utility of transferring received by gatekeeper i working on request j, α_i is gatekeeper i's individual tendency towards or away from transferring, β is the rationality parameter that captures the gatekeeper's response to the transfer return $\pi(\boldsymbol{\theta})$, $\boldsymbol{x}_j$ is the vector of the remaining characteristics of request j, $\boldsymbol{\gamma}$ is the vector of the effects of $\boldsymbol{x}_j$, and ϵ_{ij}^{tr} is the error term for gatekeeper i's jth decision.

The regression output in Table 2 is based on all seven conditions used in Experiment 1 (the six treatments in Table 1 and the additional condition *C35*). In (1) we test H1a and H1b by dropping transfer return from the model and regressing transfer decisions directly on τ_2 and c. Consistent with H1a and H1b, increasing τ_2 significantly increases transfer rates, while increasing c significantly reduces transfer rates (both $p < 0.01$). Further, in (2) we find that the effects persist

Table 2 Random Logit Regressions: Experiment 1 Treatments (*Baseline* and *Cost*)

	(1)	(2)	(3)	(4)
τ_2	0.150***	0.143***		
	(0.025)	(0.025)		
c	-0.150***	-0.141***		
	(0.037)	(0.037)		
Transfer Return ($\pi(\boldsymbol{\theta})$)			6.417***	6.133***
			(1.093)	(1.064)
Cost Treatment			-0.787**	-0.730**
			(0.378)	(0.367)
Controls: demographics, quiz errors risk/loss preferences, time remaining	no	yes	no	yes
Observations	3618	3618	3618	3618
Participants	273	273	273	273
Log Likelihood	-1572.053	-1559.645	-1572.128	-1559.614
AIC	3152.107	3139.290	3152.255	3139.228

Note. Random effects Logit regression coefficients are reported. Dependent variable is transfer decision (1: transfer, 0: continue). Intercept term not displayed. Observations are weighted by the inverse of the number of decisions faced by the participant, scaled to add up to 3618, the number of observations. The weighting is done using the ratio of the number of decisions made by a participant to the average number of decisions. Participant demographics (age, gender, call center work experience), time remaining, quiz errors, and elicited risk preferences in the gain and mixed (losses and gains) domain are controlled for in (2) and (4). ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

even after controlling for demographics, risk and loss preferences, quiz errors, and time budget remaining. We discuss the role of personal characteristics in transfer decisions in §6.

To test H1c (that transfer rates are unaffected by the presence of a transfer cost after controlling for transfer return), in (3) and (4) we replace the τ_2 and c regressors by the transfer return measure ($\pi(\boldsymbol{\theta})$), and include a *Cost* treatment indicator. The estimates confirm that participants respond to the transfer return, with the transfer return coefficient being significant in both specifications ($p < 0.01$). This is unsurprising given that each unit increase in transfer return increases the gains from transferring. However, contrary to H1c, transfer rates are lower in the *Cost* conditions (both $p < 0.05$), even after controlling for transfer return. The effect is not only statistically but also economically significant. For example, measured at transfer return $= 0$, the marginal effect of the *Cost* treatment is 14.4 percentage points ($p < 0.05$), suggesting a strong overreaction to the presence of a transfer cost. Hence:

Result 1: *H1a is supported: transfer rates are increasing in resolution time (τ_2). H1b is supported: transfer rates are decreasing in transfer cost (c). H1c is not supported: after controlling for the economic incentive to transfer, the presence of a transfer cost reduces transfer rates.*

The result regarding H1c has important implications for models of discretionary server behavior and for incentive design in practice. In §6 we further explore drivers of this result. We next describe our second experiment, in which we vary congestion.

5. Experiment 2: Variable Congestion

The remainder of our experimental investigation focuses on understanding how variable congestion ($q < 1$) impacts gatekeeper behavior. Similar to Experiment 1, we use the random utility model in §3.3 to develop and test hypotheses.

To estimate the random utility model for the variable congestion case we need to modify the transfer return measure $\pi(\boldsymbol{\theta})$. When $q = 1$ there are only two stationary policies and the transfer return measure is fully characterized by the parameter vector $\boldsymbol{\theta}$ (equation 1). In contrast, when $q < 1$ the construction of $\pi(\boldsymbol{\theta})$ is more complicated. This is because there are now three admissible stationary policies, and we must choose which policies factor into the utility calculations of the decision-maker. We resolve this by assuming: 1) A gatekeeper who transfers when the queue is empty is following the "Always Transfer" policy, because if the gatekeeper was willing to transfer when the queue is empty and be idle before starting a new request, then the gatekeeper should also transfer when the queue is nonempty, 2) A gatekeeper who continues when the queue is nonempty is following the "Always Continue" policy, because if the gatekeeper was willing to continue even though the gatekeeper could transfer and immediately beginning handling a new request, then the gatekeeper would also be willing to continue if transferring requires additional idle time, and 3) in all other cases, the gatekeeper is following the policy with the highest profit per unit time implied by the gatekeeper's action.

Denoting by R^* the maximum of $R(1,1,\boldsymbol{\theta})$, $R(2,2,\boldsymbol{\theta})$, and $R(2,1,\boldsymbol{\theta})$, we can formalize this as follows:

DEFINITION 2. When $q < 1$, the *transfer return* $\pi(\boldsymbol{\theta})$ of a condition $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}) = \begin{cases} \frac{R(1,1,\boldsymbol{\theta})}{R(2,2,\boldsymbol{\theta})} - 1, & \text{if } \left[Q = 0 \text{ and } R^* = R(2,2,\boldsymbol{\theta})\right] \text{ or } \left[Q = 1 \text{ and } R^* = R(1,1,\boldsymbol{\theta})\right] \\ \frac{R(1,1,\boldsymbol{\theta})}{R(2,1,\boldsymbol{\theta})} - 1, & \text{if } \left[Q = 0 \text{ and } R^* = R(1,1,\boldsymbol{\theta})\right] \text{ or } \left[Q = 0 \text{ and } R^* = R(2,1,\boldsymbol{\theta})\right] \\ \frac{R(2,1,\boldsymbol{\theta})}{R(2,2,\boldsymbol{\theta})} - 1, & \text{if } \left[Q = 1 \text{ and } R^* = R(2,1,\boldsymbol{\theta})\right] \text{ or } \left[Q = 1 \text{ and } R^* = R(2,2,\boldsymbol{\theta})\right] \end{cases} \tag{3}$$

5.1. Hypotheses

With variable congestion ($q < 1$), the gatekeeper experiences periods with an empty queue and periods with a nonempty queue and may therefore be idle while waiting for a request to arrive. When the queue is nonempty, the gatekeeper can transfer the current request and immediately begin handling the next request. But, when the queue is empty, the gatekeeper must wait for a new request to arrive. Hence, transfer return and, consequently, transfer rates should be lower when the queue is empty than when it is nonempty.

H2a: *Holding all other parameters constant, transfer rates are lower when the queue is empty.*

Similar to H1a and H1b in Experiment 1, H2a states the null that the direction of the response to a single change in the decision environment (here: queue state) is consistent with what the random utility model would predict. Further, analogous to H1c in Experiment 1, we can again examine the magnitude of that response by testing the null hypothesis that transfer rates should remain unaffected by queue state after controlling for transfer return.

H2b: *Holding transfer return constant, the queue state has no effect on transfer rates.*

To test H2b we include the queue state in the random utility model, and test the null that the response to the queue state is fully captured by its economic implications ($\pi(\boldsymbol{\theta})$). Rejecting H2b would then suggest that human gatekeepers over/underreact to the queue state.

5.2. Experiment Design

Experiment 2 was conducted using the same subject pool, recruitment method, and protocols as Experiment 1. A total of 137 participants were recruited (16 participants were excluded from the

Table 3 Summary of Experiment 2 Conditions

Treatment/ Condition	**# Subjects**	**Parametrization ($\boldsymbol{\theta}$)**							**Optimal Policy**	**Transfer Return**	
		r	p_1	p_2	τ_1	τ_2	c	q		**Queue Empty**	**Queue Nonempty**
VarQ	121										
V45	42	100	0.5	0.5	10	45	0	0.5	Always Transfer	11.1%	25.0%
V30	37	100	0.5	0.5	10	30	0	0.5	Transfer When Nonempty	-5.6%	5.9%
V15	42	100	0.5	0.5	10	15	0	0.5	Always Continue	-25.0%	-3.6%

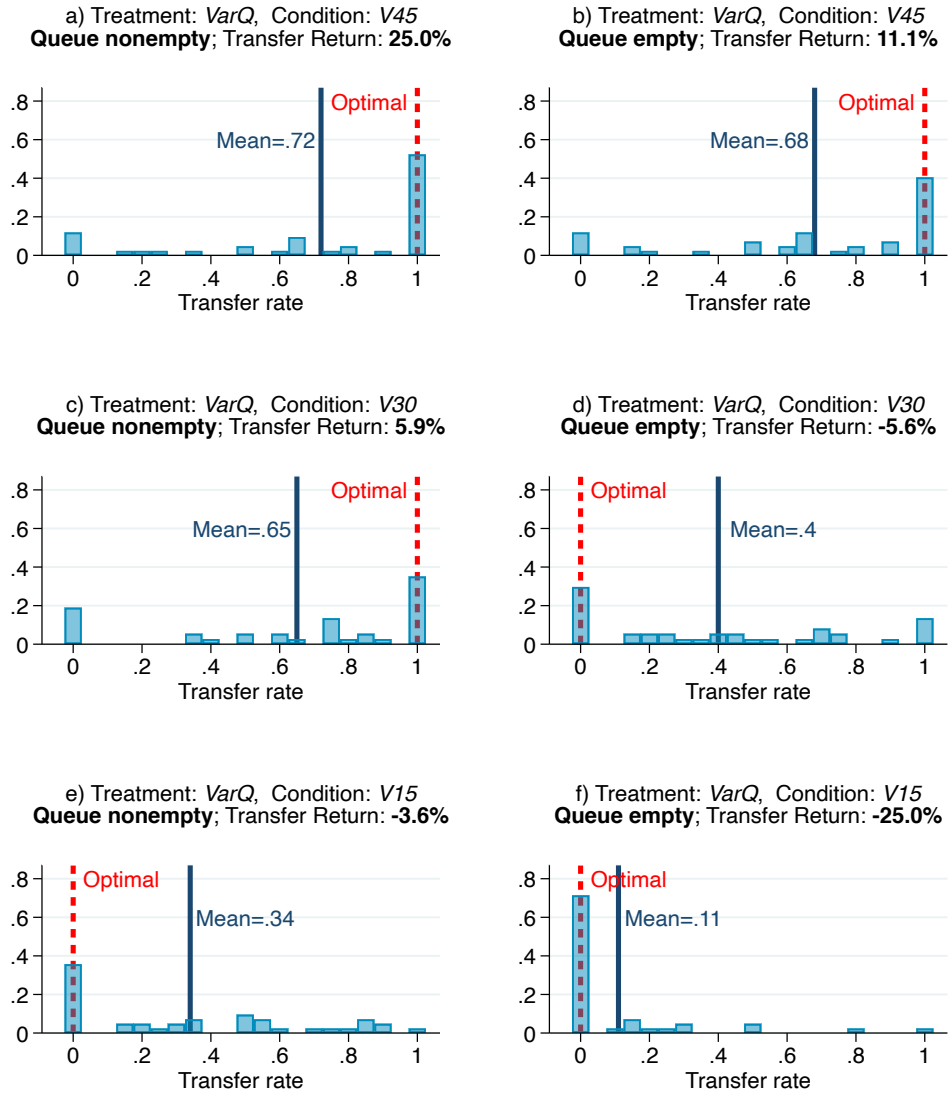
Note. Subject numbers in column 2 exclude 16 participants who failed comprehension tests. The unit of time is set to five periods (since all resolution times are multiples of five). It can be shown that the payoffs under each admissible stationary policy are equivalent to the payoffs when the unit of time is set to one period, and the reward (r) and resolution times (τ_1, τ_2) above are divided by five.

data based on comprehension tests). In each condition, we set q to 0.5. Similar to Experiment 1, we choose all time parameters to be multiples of five. Further, to mitigate the effects of uncertainty on behavior we replaced stochastic waiting times with a deterministic waiting time (10 time units). To focus on the effect of variable congestion on transfer behavior, we set $c = 0$ and $\tau_2 = \{15, 30, 45\}$, such that each of the three possible optimal policies emerge: "Always Transfer", "Always Continue", and the state-contingent "Transfer When Nonempty" policy. The remaining parameters are unchanged relative to Experiment 1. Similar to Experiment 1, the average participant earnings was approximately \$4 and the average duration was 18 minutes. The experimental conditions and the transfer return measures are summarized in Table 3. (When comparing Experiment 1 and Experiment 2 conditions, we will sometimes refer to the pooled Experiment 2 conditions as the *VarQ* treatment.)

5.3. Experiment Results

Descriptive Statistics Figure 5 shows the transfer rate distribution in the Experiment 2 conditions, where each row pertains to one of the three conditions. Within each row, the left column displays transfer rates when the queue was nonempty, while the right column displays transfer rates when the queue was empty. Our first observation concerns the proximity of the transfer rates to their theoretical optimum. Indeed, similar to the Experiment 1 conditions, modal transfer rates in all six scenarios coincide with the model predictions. Further, mean transfer rates in five out of six scenarios are significantly different from 0.5 at the 0.05 significance level (one sample $t-$tests;

Figure 5 *VarQ* Treatment: Transfer Rates



in the $\tau_2 = 30$, queue empty case, the $p-$value is 0.094). Our second observation concerns the effect that the queue state has on participant transfer rates. Recall that transfer return is higher when the queue is nonempty since the gatekeeper can immediately handle a new request. It appears that participants understand this distinction as transfer rates are higher when the queue is nonempty (left column of Figure 5) than when the queue is empty (right column of Figure 5).

Hypothesis Tests Table 4 presents Random Logit regressions with the decision to transfer (0-1) as the dependent variable. In (1) and (2) we regress transfer decisions on τ_2 and on an indicator that the queue is empty, with and without controls. We find that, similarly to Experiment

Table 4 Random Logit Regressions: Experiment 2 Treatment (*VarQ*)

	(1)	(2)	(3)	(4)
τ_2	0.104***	0.110***		
	(0.019)	(0.019)		
Queue Empty	-1.368***	-1.375***	0.109	0.176
	(0.161)	(0.161)	(0.281)	(0.279)
Transfer Return ($\pi(\boldsymbol{\theta})$)			9.515***	9.994***
			(1.625)	(1.630)
Controls: demographics, quiz errors, risk/loss preferences, time remaining	no	yes	no	yes
Observations	1529	1529	1529	1529
Participants	121	121	121	121
Log Likelihood	-711.487	-703.517	-709.060	-700.971
AIC	1430.975	1427.034	1426.120	1421.941

Note. Random effects Logit regression coefficients are reported. Dependent variable is transfer decision (1: transfer, 0: continue). Intercept term not displayed. Observations are weighted by the inverse of the number of decisions faced by the participant, scaled to add up to 1529, the number of observations. The weighting is done using the ratio of the number of decisions made by a participant to the average number of decisions. Participant demographics (age, gender, call center work experience), time remaining, quiz errors, and elicited risk preferences in the gain and mixed (gains and losses) domain are controlled for in (2) and (4). ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

1, increasing τ_2 significantly increases transfer rates. Also, consistent with H2a, transfer rates are significantly lower when the queue is empty. In (3) and (4) we explicitly control for the impact that τ_2 and the queue state have on the gatekeeper's economic incentive to transfer by regressing transfer decisions on transfer return. Additionally, to determine whether the queue state influences gatekeeper behavior beyond its effect on transfer return, we continue to include the "queue empty" indicator. We find, unsurprisingly, that transfer return has a significant positive effect on transfer rates. Further, after controlling for transfer return, consistent with H2b, queue state does not have a significant effect on transfer rates ($p = 0.527$). Hence:

Result 2: *H2a is supported: transfer rates are lower when the queue is empty. H2b is supported: holding transfer return constant, the queue state has no effect on transfer rates.*

Optimality We have so far examined how incentive systems and congestion influence transfer decisions. Alternatively, we can examine average optimality of transfer decision-making in each treatment (i.e., average deviations between observed and optimal transfer decisions). This comparison is possible because average transfer return, and hence average difficulty of finding the right

solution, is similar in the *Baseline*, *Cost* and *VarQ* treatments. The comparison shows that participants make the optimal transfer decision at a similar rate across the three treatments, with the differences not being statistically significant (65.6, 67.1, and 69.2 percent optimality in the *Baseline*, *Cost* and *VarQ* treatments, respectively; rank sum tests: $p > 0.494$). This suggests that, on average, the presence of a transfer cost or variable congestion does not make participants significantly better or poorer decision-makers.

6. Mechanisms and Discussion

Our experimental results suggest that gatekeepers exhibit transfer behaviors that are directionally consistent with the analytical benchmarks: transfer rates increase with the handling time of a request and with the congestion level in the queue. Further, behaviors appeared consistent with a random utility model, in that optimal policies were followed more closely when there was a stronger incentive to do so. However, there were some differences between the incentive systems. With a pure bonus-based incentive system, deviations from optimality were equally strong in both directions (transferring when it is optimal to continue and vice versa). However, when we introduced an explicit transfer cost into the incentive system, decision-makers transferred more even when controlling for the relative payoff differences between policies (transfer return). This suggests that gatekeepers may overreact to the cost component in the incentive system.

To better understand the reduction of transfers in the presence of a cost, we examined several potential explanations. One possible explanation suggested in the sequential decision-making literature is loss aversion (Gans and Croson 2008, Long et al. 2019). If gatekeepers experience disproportionate pain from the monetary losses following a transfer, relative to the gain following a successful resolution of a request, transfer costs may reduce transfer rates relative to a pure bonus-based system with equivalent incentive strength. Indeed, in the *Cost* treatment, loss aversion was correlated with transfer rates ($\rho = -0.22, p = 0.067$). However, as shown in column (4) of Table 2 the result persisted even after controlling for loss aversion (that is, loss aversion alone cannot explain the result).

Table 5 Random Logit Regressions with Exponential Discounting (Experiment 1)

		Discount factor						
	Long-Term	0.99	0.95	0.9	0.8	0.7	0.6	0.5
Transfer Return ($\pi(\boldsymbol{\theta})$)	4.128***							
Cost Treatment	-0.883**							
Discounted Transfer Return		3.887***	3.157***	2.418***	1.390***	0.769***	0.414***	0.212***
Log likelihood	-1493.31	-1495.38	-1494.87	-1494.74	-1495.19	-1496.18	-1496.98	-1498.13
AIC	3008.62	3010.76	3009.74	3009.48	3010.38	3012.37	3013.96	3016.25

Note. Dependent variable is Transfer decision. Data include *Baseline* and *Cost* treatments (Experiment 1). Long-term specification uses transfer return measure ($\pi(\boldsymbol{\theta})$) from Definition 1 in §3.3. Remaining specifications use discounted measure of transfer return instead of original measure, with discount factors reported in second row of table. Discounting occurs every 5 periods. To control for extreme values of discounted transfer returns at the end of the horizon, all specifications include transfer return for all but the final decision in each round, and a dummy for the final decision. All specifications use same controls as Table 2, column (4) specification. Standard errors are omitted for brevity. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Another possible explanation is narrow bracketing (Rabin and Weizsäcker 2009) or limited-look-ahead behavior (Johnson et al. 2002, Gabaix et al. 2006), which assumes a partially myopic consideration of the value function. These mechanisms would suggest that a transfer cost is treated differently, because it leads to an *immediate* cost accrual while delaying earnings. To test this explanation we examined whether a model in which the more distant payoffs are weighted less than the immediate ones would fit our data better. Specifically, we replicated the analysis in Table 2, col. (4), removing the *Cost* Treatment dummy, and replacing the time-invariant $R(\cdot,\cdot,\boldsymbol{\theta})$ measure by the expectation of the sum of discounted payoffs, with discounting factors of 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99.

The log likelihoods and AIC measures for our original specification and for the models with discounting are summarized in Table 5. Examining both of these criteria suggests that all models with discounting result in a poorer fit relative to the standard model; however, among these models, discount factors of 0.9 and 0.95 perform the best (AICs of 3009.74 and 30009.48), and the most myopic model (discount factor of 0.5) performs the worst (AIC of 3016.25). Thus, overall subjects appear quite forward-looking, i.e., myopia alone would not produce our results.

Lastly, theories of information processing would suggest that more complex incentive schemes may lead to the adoption of simplified decision heuristics (Sweller 2010). In other words, when the problem is complex, the mere presence of a transfer cost may prompt a mental shortcut (or a more visceral reaction) towards avoiding transfers. This explanation appears to be most consistent

with our data. Indeed, when the penalty for errors is low (transfer return close to -0.1 or 0.1), the transfer rate gap between *Baseline* and *Cost* treatments is between 7 and 10 percentage points. In contrast, when the penalty for errors is high (transfer return above 0.2), the transfer rate gap is reduced to 2 percentage points. Moreover, the overreaction to transfer cost appears to be stronger for participants who have difficulties combining the time and reward information available to them: transfer rates in the *Cost* treatment are negatively correlated with the participant's performance on the comprehension questions ($\rho = -0.37, p < 0.01$). Taken together, these results are consistent with incorrect information processing as the main behavioral mechanism.

Our second result is that under variable congestion subjects made random errors but did not over/undertransfer under an empty/nonempty queue state. That is, different from the overreaction to cost, there were no systematic deviations in response to congestion information. One reconciliation of this result is that processing monetary information has a different affect than processing operational information, such as queue states. Taken together, these results suggest that information processing deficits can be nuanced, especially in problems that involve trading off money and time, and that further research is needed to better understand these problems.

7. Concluding Remarks

In this study we modeled gatekeeper transfer decisions as a dynamic finite-horizon problem in which a random number of stopping decisions are made sequentially. We solved this problem analytically and developed comparative statics to formulate hypotheses regarding human gatekeeper behavior. We then tested these hypotheses in two experiments, in which we examined how an explicit transfer penalty, and variable congestion levels affect transfer decisions.

To expose potential biases in transfer decision-making, we adopted a research design that uses payoff space (as opposed to action space) to make all-else-equal comparisons between experimental conditions (Harrison 1989, Smith and Walker 1993). To extend this idea to a dynamic setting, we developed a dynamic model with a stationary optimal policy and a finite horizon, so that the payoff consequences of decisions are time-invariant and can be collapsed into a single number

("transfer return"). We then examined behaviors using pairwise comparisons of conditions with similar transfer return, and a random utility model that controls for transfer return econometrically.

Our findings expose the features of the incentive system and queueing environment that human decision-makers can respond to, and the ones they have difficulties incorporating. Specifically, we find that (1) decision-makers tend to overreact to a transfer cost and (2) decision-makers do not over/underreact to the queue being full or empty.

The strong reduction of transfers in response to cost aligns with the ABC Bank (our industry partner) data discussed in §1. Indeed, while the management team of the ABC bank may have been pleased with the drop in transfer rates after the scorecard rollout, our experimental results suggest that the rates may have fallen too much, as agents struggled to correctly calibrate their response to incentives. While we cannot make any definitive conclusions about the performance effects in the field, our findings suggest that managers should be careful when introducing loss incentives (monetary penalties). Such incentives may send unintentional signals about which behaviors are desirable and which are not, making it more difficult for gatekeepers to make rational decisions.

The mismatch between desired and realized transfer behaviors can be costly for the organization operating the gatekeeper system. It is therefore natural to ask what drives this behavior and how it can be remedied. While our investigation provides some initial evidence that the reduction of transfers is a more psychological response to the presence of the cost, a deeper investigation of the mechanisms and moderators of that response may be worthwhile. For example, can the deviations be reduced through more careful rollout and communication of the incentive system? A broader question to consider would be to determine the pervasiveness of such behaviors: do they happen independent of the operating environment, or will they be more pronounced as task or queue system complexity is added to the operating system?

While decision-makers struggled to correctly respond to bonus+cost incentives, we did not find any systematic deviations in response to queue state. This suggests an opportunity to introduce congestion information into the call management system used by call centers like the one at ABC

Bank. To the extent that encouraging transfers during periods of high congestion can help reduce waiting times, managers may consider making congestion levels more visible or salient to gatekeepers who appear to (at least directionally) be capable of responding to simple congestion information.

Given that the main objective of our model was to develop analytical benchmarks for behavioral experiments, our needs were served by a single gatekeeper subsystem within the larger system, as shown in Figure 2. However, the general gatekeeper-expert framework may prove helpful in examining broader questions around service system design, including optimal gatekeeper/expert staffing levels, as well as the implications of transfers on perceived and actual quality of service delivered to consumers. An even more comprehensive investigation would examine the transfer dynamics in various service settings, for example, call centers, IT Support, healthcare delivery, and other forms of collaborative service production.

References

Allon G, Kremer M (2018) Behavioral foundations of queueing systems. The Handbook of Behavioral Operations 9325.

Ball SB, Holt CA (1998) Speculation and bubbles in an asset market. Journal of Economic Perspectives 12(1):207–218.

Batt RJ, Terwiesch C (2017) Early task initiation and other load-adaptive mechanisms in the emergency department. Management Science 63(11):3531–3551.

Bearden JN, Murphy RO, Rapoport A (2008) Decision biases in revenue management: Some behavioral evidence. Manufacturing & Service Operations Management 10(4):625–636.

Bearden JN, Rapoport A, Murphy RO (2006) Sequential observation and selection with rank-dependent payoffs: An experimental study. Management Science 52(9):1437–1449.

Ben-Akiva ME (1973) Structure of passenger travel demand models. Ph.D. thesis, Massachusetts Institute of Technology.

Berry Jaeker JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. Management Science 63(4):1042–1062.

Chen DL, Schonger M, Wickens C (2016) otree—an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9:88–97.

Cox JC, Oaxaca RL (1989) Laboratory experiments with a finite-horizon job-search model. Journal of Risk and Uncertainty 2(3):301–329.

Delasay M, Ingolfsson A, Kolfal B, Schultz K (2019) Load effect on service times. European Journal of Operational Research 279(3):673–686.

Dong J, Feldman P, Yom-Tov GB (2015) Service systems with slowdowns: Potential failures and proposed solutions. Operations Research 63(2):305–324.

Duffy J (2016) Macroeconomics: a survey of laboratory research. Handbook of experimental economics 2:1–90.

Eckel CC, Grossman PJ (2002) Sex differences and statistical stereotyping in attitudes toward financial risk. Evolution and human behavior 23(4):281–295.

Edie LC (1954) Traffic delays at toll booths. Journal of the operations research society of America 2(2):107–138.

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: An empirical analysis of a maternity unit. Management Science 63(10):3147–3167.

Gabaix X, Laibson D, Moloche G, Weinberg S (2006) Costly information acquisition: Experimental analysis of a boundedly rational model. American Economic Review 96(4):1043–1068.

Gans N, Croson R (2008) Introduction to the special issue on behavioral operations. Manufacturing & Service Operations Management 10(4):563–565.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management 5(2):79–141.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. Manufacturing & Service Operations Management 4(3):208–227.

George JM, Harrison JM (2001) Dynamic control of a queue with adjustable service rate. Operations research 49(5):720–731.

Harrison GW (1989) Theory and misbehavior of first-price auctions. The American Economic Review 749–762.

Hasija S, Pinker EJ, Shumsky RA (2005) Staffing and routing in a two-tier call centre. International Journal of Operational Research 1(1/2):8–29.

Hathaway B, Kagan E, Dada M (2021) Transfer decisions in services: A multi-method study. SSRN Working Paper URL `http://ssrn.com/abstract=3771633`.

Heineke J, Davis MM (2007) The emergence of service operations management as an academic discipline. Journal of operations management 25(2):364–374.

Hopp WJ, Iravani SM, Yuen GY (2007) Operations systems with discretionary task completion. Management Science 53(1):61–77.

Hyndman K, Embrey M (2018) Econometrics for experiments. The Handbook of Behavioral Operations, 35–88 (John Wiley & Sons, Inc. Hoboken, NJ, USA).

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. Management Science 64(9):4389–4407.

Johnson EJ, Camerer C, Sen S, Rymon T (2002) Detecting failures of backward induction: Monitoring information search in sequential bargaining. Journal of Economic Theory 104(1):16–47.

KC DS, Staats BR, Kouchaki M, Gino F (2020) Task selection and workload: A focus on completing easy tasks hurts performance. Management Science .

Kirchler M, Huber J, Stöckl T (2012) Thar she bursts: Reducing confusion reduces bubbles. American Economic Review 102(2):865–83.

Kremer M, de Vericourt F (2020) Ticket queues with regular and strategic customers. Working paper .

Lee HH, Pinker EJ, Shumsky RA (2012) Outsourcing a two-level service process. Management Science 58(8):1569–1584.

Leider S, Şahin Ö (2014) Contracts, biases, and consumption of access services. Management Science 60(9):2198–2222.

Long X, Nasiry J, Wu Y (2019) A behavioral study on abandonment decisions in multistage projects. Management Science .

McFadden D (1973) Conditional logit analysis of qualitative choice behavior .

Noussair C, Robin S, Ruffieux B (2001) Price bubbles in laboratory asset markets with constant fundamental values. Experimental Economics 4(1):87–105.

Oliva R, Sterman JD (2001) Cutting corners and working overtime: Quality erosion in the service industry. Management Science 47(7):894–914.

Rabin M, Weizsäcker G (2009) Narrow bracketing and dominated choices. American Economic Review 99(4):1508–43.

Rapoport A, Tversky A (1970) Choice behavior in an optional stopping task. Organizational Behavior and Human Performance 5(2):105–120.

Rosokha Y, Wei C (2020) Cooperation in queueing systems. Available at SSRN 3526505 .

Sampson SE, Froehle CM (2006) Foundations and implications of a proposed unified services theory. Production and operations management 15(2):329–343.

Schultz KL, Juran DC, Boudreau JW, McClain JO, Thomas LJ (1998) Modeling and worker motivation in jit production systems. Management Science 44(12-part-1):1595–1607.

Seale DA, Rapoport A (1997) Sequential decision making with relative ranks: An experimental investigation of the" secretary problem". Organizational behavior and human decision processes 69(3):221–236.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. Management Science 49(7):839–856.

Shunko M, Niederhoff J, Rosokha Y (2018) Humans are not machines: The behavioral impact of queueing design on service time. Management Science 64(1):453–473.

Smith VL, Walker JM (1993) Monetary rewards and decision cost in experimental economics. Economic Inquiry 31(2):245–261.

Sweller J (2010) Cognitive load theory: Recent theoretical advances. .

Ülkü S, Hydock C, Cui S (2019) Making the wait worthwhile: Experiments on the effect of queueing on consumption. Management Science .

Zhan D, Ward AR (2019) Staffing, routing, and payment to trade off speed and quality in large service systems. Operations Research 67(6):1738–1751.

Appendix A: Experimental Instructions and Screen Shots (Baseline treatment, abridged)

This study consists of two rounds of a virtual task. In both rounds you will be playing the role of a customer service representative resolving a series of customer issues. Customer issues will arrive one by one, and you will be launching several attempts to resolve them. Before starting the first round, you will complete a training round. In the training round you will play a shorter version of the task, and your points will not count towards your bonus. The training round will help you familiarize with the task.

As you resolve customer issues, you will be accumulating points: each resolved issue earns you 100 points. However, you do not know how many attempts are needed to resolve each issue. For each issue, the number of attempts required to resolve it will be determined by luck (a random number generated by the computer, similar to a lottery). In particular, each issue can take either one or two attempts to resolve, and it is equally likely that the issue will be resolved after one attempt, or after two attempts. In other words, there is a 50% chance that the issue will be resolved after just one attempt, and a 50% chance that it will require two attempts to be resolved. Each issue is unrelated to the previous ones. This means the number of attempts needed to resolve any given issue does not depend on the number of attempts needed for the previous issues.

In the training round you will have a total of 150 virtual time units. You will earn 100 points for each resolved issue. The times required for each attempt are listed below.

attempt #1: 10 time units; attempt #2: 25 time units

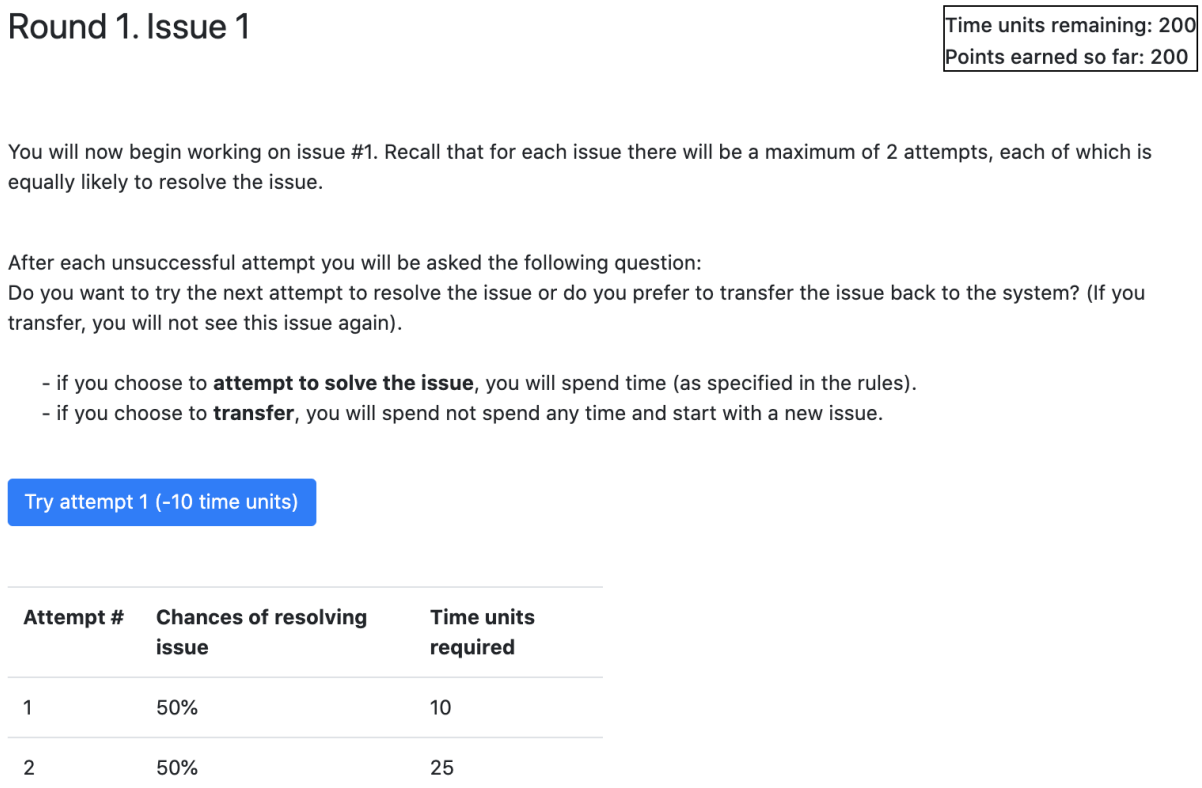
For example, suppose you start the game and the first issue takes one attempt to resolve. Then, your time allowance will drop by 10 units, from 150 to 140 time units. If, in contrast, you take two attempts to solve the issue, then your time allowance will drop by 10+25 = 35 units, from 150 to 115 time units.

Note: you cannot change the order of the attempts. This means that you must start every issue with attempt #1. Then, if the first attempt does not resolve the issue, you decide whether to launch attempt #2.

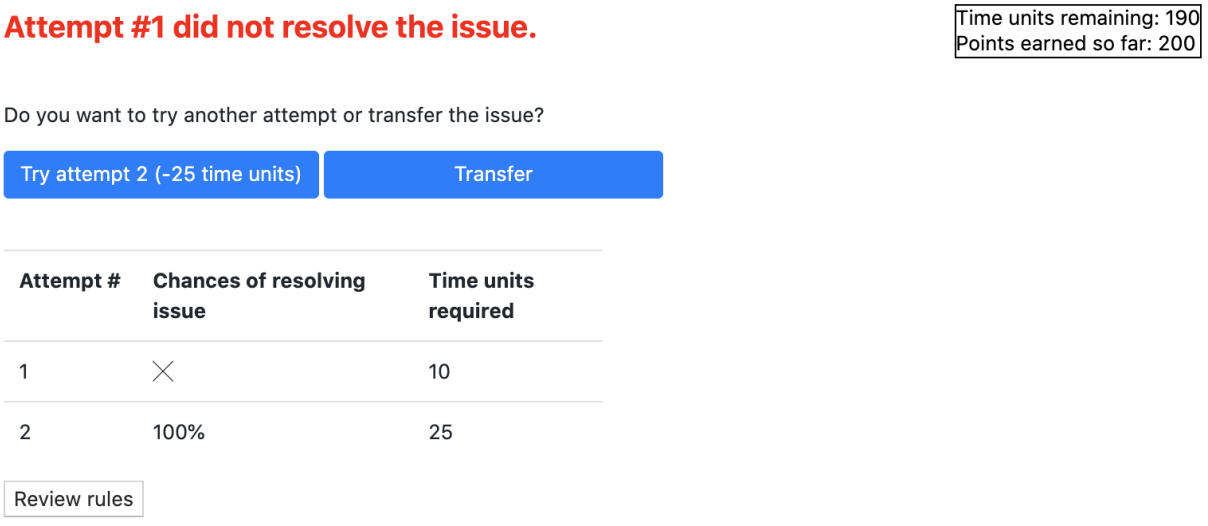
At any point you can decide to stop the attempts, and instead transfer the issue back to the system. After that, you will not see the issue again and instead you will begin working on a new issue. This process will repeat until you run out of the 150 time units allocated to you at the beginning of the experiment.

How long will I be working on the task? Recall that you will have a time budget of 150 virtual time units in this round. This means at some point your time budget will drop to a level that is not sufficient to resolve an issue. When that happens, the round will end and we will pay the last few time units at a per unit price to make you whole. In particular, you will then be compensated with 5 point(s) per remaining time unit. For example, suppose you have resolved 4 issues, transferred 2 issues and have 10 time units remaining in your budget. You will earn 4*100 = 400 points from resolving the issues. In addition, we pay you 10 * 5 = 50 point(s) for the remaining time units. This means that your total earnings will be 400 + 50 = 450 points.

Figure A.1 Screenshots of the main decision screens.



(a) Screenshot of decision screen: participant is about to attempt the first potential solution.



(b) Screenshot of decision screen: participant is deciding whether to transfer the issue.