Sponsor Preferences and Inventory Optimization in Child Sponsorship Programs

Jun Ho Choi

Thesis for the MA in Computational Social Science Program at the University of Chicago

Abstract

While few studies underline the potential impact of child sponsorship programs (CSPs), the literature on how to maximize sponsorship (or attain other goals of child sponsorship organizations [CSOs]) is almost non-existent. Focusing on the online sponsorship marketing platforms administered by CSOs, this paper addresses the said lack of discussion in the following manner. Firstly, using a simple economic framework, I find partial characterizations of the CSO's optimal strategies in ordering the different types of potential recipients (PRs). This effort is targeted at reducing the computational burdens to find an optimal "inventory" to the CSO's problem. Further, the study conducts empirical analyses to find specific features of PRs that hinder or promote probabilities of yielding a match. Finally, I conduct additional statistical investigations to detect whether there exists a discrepancy between the proposed strategy of the CSO and what the economic framework predicts. The study finds that geography- and urgency-related variables are significant to match statuses and that the CSO's behavior is within the scope of theoretical prediction. All empirical analyses in this study utilize data gathered from Compassion International using web-scraping.

1 Introduction

As non-governmental child sponsorship organizations (henceforth CSO) are continuing to expand their operations around the world, more children – both domestic and international – are receiving sponsorship through direct matches with sponsors, who often are economically better-off than their beneficiaries. For instance, Compassion International (2018) have exceeded 1,800,000 in the number of directly-sponsored children in 2018; World Vision reports that their sponsorship program reaches out to approximately 3,800,000 globally (Huber 2019). Studies investigating the said programs' effectiveness are few, but they commonly underline the positive impacts of the child sponsorship programs (henceforth CSP) to the lives of recipients. Wydick et al. (2013), for example, use a dataset of approximately 10,000 individuals across six countries – part of whom were supported through Compassion International – to conclude that sponsorship has increased the recipients' years of education, school completion rates, and employment outcomes as adults (pp. 418-430). An earlier study by Kremer et al. (2003) also provides evidence for the positive impact of CSP, although the program under scrutiny specifically targeted schools.

Despite the said few discoveries highlighting the positive results of CSP for the sponsorship recipients, there are even fewer items in the literature studying how the sponsor-recipient matches are created, and how their numbers can be maximized. There are two dimensions to this hardly-inspected set of problems. The first is identifying sponsors' preferences over recipient characteristics, as some attributes may hinder or promote the selection of potential recipients (henceforth PRs) by potential sponsors (henceforth PSs). This problem perhaps owes in part to the lack of public data surveying the motives behind PSs' participation in CSP. Another aspect is to analyze whether the current structure of delivering PR information to PSs is optimal in terms of maximizing sponsorship.[1] While this is a complicated standalone question, note that one should account for PS preferences as well in order to more thoroughly assess the optimality of the said structure.

This paper aims to provide answers to the said dimensions in the following manner. Firstly, the study introduces a simple model of inventory design with search costs, in which the CSO serves as the designer for an "inventory" or ordered sequence of different PR types that anticipates different types of PSs to arrive. While many strategies can be identified, this study first analyzes the ones in which a certain type of PRs appear in the inventory as a block (to be denoted as "bunching" strategies). Further, the paper explores deviations from the aforementioned set of strategies and assess whether such deviations can be beneficial given different demand (PS) settings. As closed-form solutions are difficult to derive in this discrete-and-ordered optimization problem, the study assumes computational methods will be used to characterize potential optimal strategies and targets at reducing the number of optimal strategy candidates.

Furthermore, this study employs an empirical analysis of the Compassion International dataset of 9,518 matched and unmatched children collected via web-scraping. The said analysis is conducted as follows. First, I employ a set of parametric and nonparametric strategies including survival analysis and regression discontinuity design (henceforth RDD) to find evidence that there are variables significant in impacting the PSs' decision-making. In addition to identifying the said variables, the evidence will ensure that there are

[1] Another extension for optimality in this dialogue is maximizing the "social welfare" of the CSP environment by jointly considering maximal sponsorship and PSs' utility.

more than one type of PRs (in terms of utility yielded to PSs once matched), which is a necessary condition for a non-trivial optimization as described in Section 2. Following this step, I scrape additional datasets from Compassion International and assess how the average PR types differ in their page-to-page ordering using statistical tests such as MANOVA. This exercise will help evaluate whether the CSO follows a rule for optimal inventory design as identified in the previous steps.

Following this section, Section 2 elaborates on the economic framework of the model of inventory design and key results from the said framework. Section 3 describes the data used in this paper as well as the empirical strategies to be employed. Section 4 explicates the results of the said empirical strategies. Section 5 is a short conclusion, summarizing the findings of this paper. References and the Appendix follow Section 5.[2]

2 Economic Framework: A Model of Inventory Design

2.1 Motivation and Differences from Other Inventory Optimization Models

Many, if not all, of the CSOs use a particular way of revealing the PRs available for sponsorship; that is to show only a portion from the entire set of PRs, and to have the PSs make further efforts if they would like to view more. Such efforts are usually having to simply interact with buttons to load more PRs, but some organizations like Compassion International have further options such as being able to display the PRs by longest waiting time. This commonality in the strategies to display different PRs is contrasted with the variation in the ways of presenting PR-specific information. The variables described in the PR-specific information pages of different organizations are listed in Table 6 of the Appendix.

It is not clear why many of the organizations use “inventory-style” display of different PRs as opposed to having the PSs input their preferences and the CSOs recommending a suitable PR. However, a potential reason might be that there are more PRs with features that can be deemed unfavorable by a majority of PSs. Sections 3 and 4 discuss methods and results of identifying variables that either contribute (or hinder) matching, and their potential impacts on the probabilities of matching; such would indirectly show that certain characteristics are preferred (or less preferred) by PSs, using the logic of revealed preference.

Even if one accepts the scarcity (of favorable PRs) argument for using inventory-style display of PRs, it is uncertain whether organizations are optimizing the sequence in which PRs appear in the inventory, given that the set of PRs is fixed. The main concern in this contemplation is that, if there are multiple types of PSs – some of whom might be more meticulous in their selection of PRs –, then the less picky PSs who arrive early to the inventory may deplete the potential matches with PRs whose qualities are more favorable. A further consideration is that the PSs might have search costs. That is, with respect to the online interactions to view more PRs, a PS might exit the said website after having made a few clicks and being unable to find a suitable PR.

[2] All relevant codes will be uploaded to the following link: https://github.com/jtschoi/MA_Thesis_Codes. The said codes may be subject to changes and updates.

Table 1: Notations Used in the Model of Inventory Design with Search Costs

Notation	Meaning
T	The set of PR types; $T \equiv \{t_l, t_h\}$
t_l, t_h	"Low" and "high" types of PR; also utilities to the PSs once matched
M_l, M_h	The numbers of t_l-type and t_h type PRs; $M \equiv M_l + M_h$, $M_h < M_l$
Θ	The set of PS types; $\Theta \equiv \{L, H\}$
o_L, o_H	Utilities from choosing outside options held by L- and H-type PS
A	The set of per-search actions; $A \equiv \{a_s, a_q, a_r\}$
a_s, a_q, a_r	Actions by the PS; refers to "selecting," "quitting", and "re-searching"
c_l, c_h	Costs of taking care of unmatched t_l- and t_h-type PRs; $c_l > c_h > 0$
N	The number of PSs to arrive
$\mathcal{I}$	Notation for an arbitrary inventory (ordered sequence of PR types)
μ	The belief of a single H-type arriving to the inventory ($\mu \in [0, 1]$)
b	The maximum amount of searches an H-type PS conducts
$EV(\mathcal{I}, N, \mu, b)$	(Ex-ante) expected payoff of the CSO given $\mathcal{I}$, N, μ, and b
ℓ	Arbitrary sequence of PSs
$V(\mathcal{I}, \ell, b)$	(Ex-post) deterministic payoff of the CSO given $\mathcal{I}$, ℓ, and b

The above issues with the order in which PRs appear, different types of PSs and PRs, and the search costs they may face inspire the model of inventory design to be described in this section. The model may be reminiscent of other models of optimizing inventories (or the so-called "newsvendor problems"), such as the seminal work by Arrow, Harris, and Marschak (1951) and articles such as Netessine and Rudi (2003). The commonality between such models and the one inspected in this paper is that it assumes there is a stream of demand-side agents, who come to the inventory and leave with the most favorable product (or, in this paper's circumstances, PR) if there is any. However, a key difference is that while a demand-side agent of the aforementioned other models can view the entire inventory, that of this paper's model can only sequentially and with potential costs browse through the inventory. Further, the model of this paper assumes that the composition of PR types is exogenously determined, to reflect the idea that the CSOs do not selectively decide who should be supported by their program. On the other hand, preexisting models previously mentioned assume that the inventory designer optimizes by choosing the types of products to be in their inventory.

Before proceeding with the model setup, I note that the goal of this paper's model is not to perfectly find the closed-form solution (i.e., the optimal inventory) to the inventory optimization or design problem. Rather, it is to characterize some of the attributes of an optimal inventory strategy and remove the non-conforming strategies from the set of all candidates for the optimal strategy. After acquiring the set of remaining viable strategies, the CSO may use search methods such as grid search or randomized search to acquire the optimal strategy (or an approximate one, if there are too many cases to cover). Having said this, the study will acknowledge certain cases in which optimal strategies are fully identifiable and describe them accordingly.

2.2 Setup

In the model of child sponsorship as an inventory design problem, there are two key sets of players: the PSs and the CSO. Reflecting many of the real-world CSPs, the PRs are passive in that they endow their information to the CSO in the hopes of yielding sponsorship matches. In this part, I first briefly describe the situation of the PRs. Then, I elaborate on the PSs' situation along with the decision-making they face. Finally, I delineate the CSO's problem; however, the attempt to derive optimization results for the said problem is accomplished in the next part.

Potential recipients. — Let $T \equiv \{t_l, t_h\}$ be the set of PR types, with $t_h > t_l > 0$. t_l refers to those who are unlikely to be helped by sponsorship, due to circumstances like disease or vulnerability from exploitation. On the other hand, t_h refers to those whose situations are likely to improve if external support is given. For simplicity, suppose that t_l and t_h are the levels of utility yielded from helping the respective types as well, and that there is no variation within the same type. Additionally, suppose that there are M_l t_l-type PRs and M_h t_h-type PRs, with the total number of PRs being denoted as $M = M_l + M_h$. I assume that $M_h < M_l$, which is to say that there are more PRs in severely desperate situations.

Potential sponsors. — Let $\Theta = \{L, H\}$ be the set of PS types. The two types differ only by the outside options they have, in which utility from the outside option of an L-type PS (denoted by o_L) is lower than that of an H-type PS (denoted by o_H, and $o_H > o_L$). I assume that every PS, regardless of one's type, knows the values of t_l and t_h. Additionally, I assume the following:

$$t_h > o_H > t_l \quad \text{and} \quad t_h > t_l > o_L$$

That is, while an L-type PS is willing to match with any PR that one sees, an H-type PS only seeks a t_h-type PS; this is due to the higher utility from the outside option that the H-type PSs have.

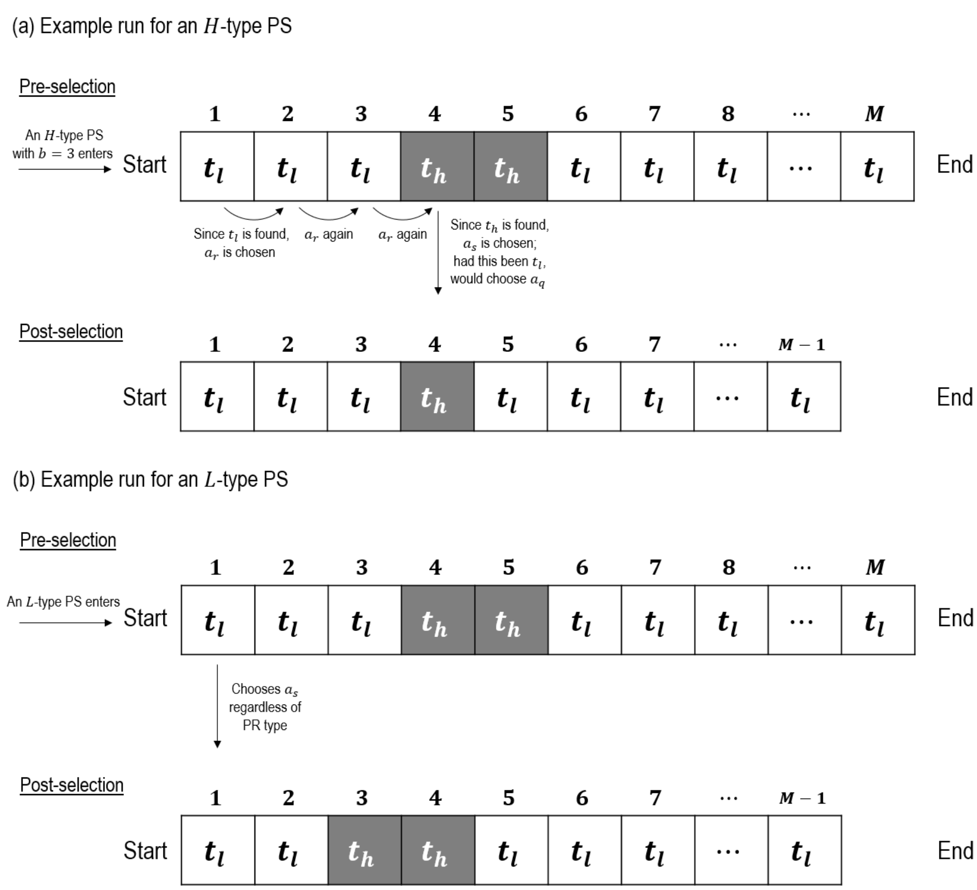
Each PS searches through the list (or inventory) of PRs posted and designed by the CSO, where the said inventory works as follows. A PS enters the inventory by seeing its first element (PR), where there are three possible actions: (1) select the current PR to sponsor and exit the market (denoted as "selecting"); (2) exit the market without selecting any PR, and getting the outside option (denoted as "quitting"); and finally, (3) staying in the inventory and searching for another PR (denoted as "re-searching"). Let $A \equiv \{a_s, a_q, a_r\}$ denote the set of PS actions, in which a_s, a_q, and a_r respectively correspond to selecting, quitting, and re-searching. While the initial entering of the inventory does not incur costs, I assume that the cost of re-searching k times incurs a cumulative cost of $c(k)$, where $c(\cdot)$ is a convex and nonnegative-valued cost function with $c(0) = 0$. Once a PS chooses re-searching and views another PR, the PS once again faces selecting the most optimal action $a \in A$.

While choosing the most optimal (in terms of expected utility) action amongst the three options previously mentioned, I assume for simplicity that an H-type PR is myopic and only considers the next search when it comes to re-searching. To further elaborate on the said myopia, let us define $U_H : A \times \mathbb{N}_0 \to \mathbb{R}$ as the (expected) utility of an H-type PS having made a search. For instance, $U_H(a_s, k)$ refers to an H-type PS selecting the currently-viewing PR after having re-searched k times. The function U_H has the following values:

$$U_H(a,k) = \begin{cases} t(k+1) - c(k) & \text{if } a = a_s \\ o_H - c(k) & \text{if } a = a_q \\ \eta t_h + (1-\eta)t_l - c(k+1) & \text{if } a = a_r \end{cases}$$

where $t(k+1)$ refers to the type of PR in the $(k+1)$th position in the list and $\eta \in [0,1]$ is defined as the belief that the next re-search yields a t_H-type PR. The above-mentioned myopia is captured by the expression for $U_H(a_r, k)$ where the expected utility from re-searching does not consider multiple re-searches but only the next one. Note that while it may be more rigorous to suppose that a PS takes an expectation over all potential re-searches rather than just the immediately subsequent one, the above myopia assumption does not subtract much from the economic dynamics within this model. This is due to the convexity of the cost function $c(\cdot)$, which makes every additional re-search all the more unfavorable.

Figure 1: Example Runs of PS Decision-making



I note that the above simplification of the H-type PSs' behavior makes it as if the said PSs have up to a certain threshold of re-searches that each can conduct. The value of the said threshold is – denoted by b – is dependent on the values of t_h, t_l, o_H, and η and the functional form of $c(\cdot)$ (i.e., $b \equiv b(t_h, t_l, o_H, \eta, c)$). For the sake of convenience in later discussion of CSO optimization, I will often refer to using b in place of η

and $c(\cdot)$; however, this deliberation is made with the consideration that the said values and functional form affect b. Examples of an H-type PS searching through the inventory, which makes use of the threshold b, is provided in panel (a) of Figure 1.

For the L-type PSs, I have previously stated that they are willing to accept any type of PR. For further simplification and distinction between the behaviors of different PS types, let us assume that the L-types do not consider re-searching. For the sake of a more rigorous model, one may consider an L-type PS re-searching when one sees a t_l-type PR and wishes to conduct another set of searches in the expectation that one finds a t_h-type PR. However, once again, the said simplification is not problematic due to the convexity of $c(\cdot)$, which will make re-searching increasingly unfavorable. A trivial example of an L-type PS searching through the inventory is also provided in panel (b) of Figure 1.

After a PS leaves the inventory, had the said PS chosen a PR, the inventory changes by leaving the sub-list of previous PRs (if any) as is, and the sub-list of subsequent PRs advance by one position; this is also illustrated in Figure 1. As a result, the list length shrinks by 1 if a match is yielded.

Child sponsorship organizations. — The CSO's problem, as an "inventory designer," is as follows. Suppose that for each PR (regardless of type) matched with a PS, the CSO yields $v > 0$; this is to indicate that the CSO wants to maximize the number of sponsored children. Without loss of generality, let $v = 1$. However, further suppose that for every t_l-type PR unmatched, a cost of c_l is incurred; similarly, for every t_h-type PR unmatched, the cost incurred is c_h. I assume that $c_l > c_h > 0$, to reflect that it may be more difficult for the CSO to keep supporting those whose situations are more distressed (i.e., the t_l-types). Let m, w_l, and w_h denote the numbers of matched PRs, type-t_l PRs without sponsorship, and type-t_h PRs without sponsorship, respectively, such that $M = m + w_l + w_h$. Then, the ex-post payoff of the CSO, denoted V, can be stated as follows:

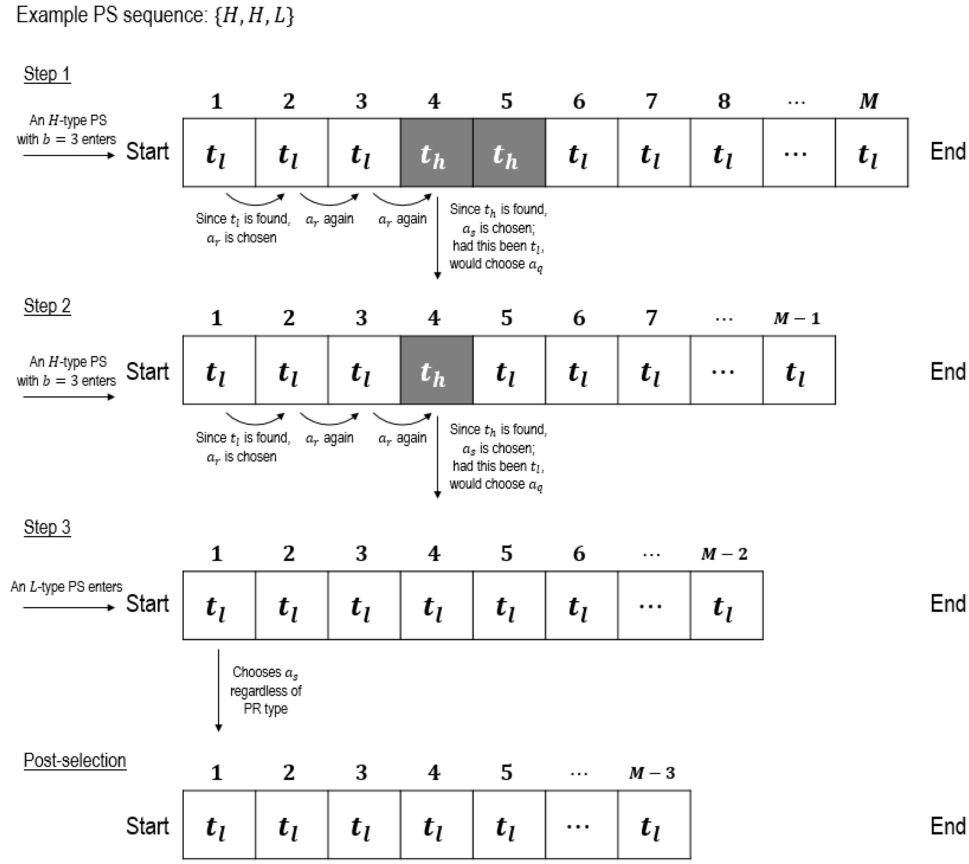
$$V = m - c_l w_l - c_h w_h = M - (1 + c_l)w_l - (1 + c_h)w_h. \tag{1}$$

In this paper, I assume that the CSO seeks to maximize ex-ante expected utility; in order to do so, it designs the inventory $\mathcal{I}$, which is in the form of an ordered sequence or list of PRs (as exemplified in Figure 1) with length M. For simplification, the CSO anticipates that there will be N PSs entering $\mathcal{I}$; further, the CSO expects that an H-type PS arrives with the probability $\mu \in [0, 1]$ (i.e., $\mu = \Pr(H)$ and $1 - \mu = \Pr(L)$). In reality, a CSO might learn of the value of N by assessing the previous numbers of visitors to its website who interact with sponsorship information. Additionally, suppose that the CSO knows of the value of b (i.e., maximum number of voluntary re-searches) of the H-type PSs. Notice that the values of w_l and w_h in are determined by b and the (ordered) sequence of arriving PS types, given $\mathcal{I}$; an example is provided in Figure 2. This means that w_l and w_h, along with V in (1), can be expressed as functions of b, $\mathcal{I}$, and the said sequence of PSs. Using ℓ to denote an arbitrary PS sequence, one can write:

$$V(\mathcal{I}, \ell, b) = M - (1 + c_l)w_l(\mathcal{I}, \ell, b) - (1 + c_h)w_h(\mathcal{I}, \ell, b). \tag{2}$$

With the knowledge of N and μ, the CSO is able to write down all cases of such PS sequences and calculate the probabilities of the said cases happening. Let the set of all possible ordered sequences of PS

Figure 2: Example of a PS Sequence and Their Decision-making



types be denoted by $\mathcal{L}_N$. Also, let $\varphi(\ell, \mu)$ denote the probability that a PS sequence ℓ arrives where the probability of H-type arriving is μ; $\varphi(\ell, \mu)$ can be written out as:

$$\varphi(\ell, \mu) = \prod_{i \in \ell} \Pr(i) \tag{3}$$

This foundation allows the ex-ante expected payoff of the CSO, denoted EV, to be pinned down by $\mathcal{I}$, N, μ, and b given the values of c_h and c_l. With the expressions for $V(\mathcal{I}, \ell, b)$ and $\varphi(\ell, \mu)$ provided in (2) and (3), the expression for EV can be written as follows:

$$EV(\mathcal{I}, N, \mu, b) = \sum_{\ell \in \mathcal{L}_N} \varphi(\ell, \mu) V(\mathcal{I}, \ell, b). \tag{4}$$

Before advancing further, I note that the notations used in this model are organized in Table 1.

2.3 Optimization of the Inventory Designer

A. Characterizing the CSO's Problem

In this paper, I assume that the CSO can only design the order of PR types to appear as $\mathcal{I}$, in which the numbers of t_h-type and t_l-type PRs in $\mathcal{I}$ respectively are M_h and M_l. In reality, a CSO may consider different strategies. For instance, it may invest in creating a more user-friendly website so that search costs are lowered on average and b (i.e., the number of maximum searches) increases. Or, by extensive marketing, it may attract more L-type PSs, thereby increasing N and decreasing μ.

In light of the above discussion and the expression for the expected payoff in (4), one can write the CSO's problem as follows. The CSO chooses an optimal sequence $\mathcal{I}^*$ of PR types where

$$\begin{aligned} &\mathcal{I}^* = \underset{\mathcal{I}}{\operatorname{argmax}}\ EV(\mathcal{I}, N, \mu, b) \\ &\text{s.t. the numbers of } t_h\text{-type and } t_l\text{-type PRs in } \mathcal{I} \text{ are } M_h \text{ and } M_l. \end{aligned} \tag{5}$$

Recursive formula for the function V. — A key difficulty in solving for the CSO's problem is that the organization is not choosing numbers and types of PRs to be displayed in the inventory but rather the order in which they appear (represented by $\mathcal{I}$). More specifically, it would be difficult to characterize a closed-form solution. Alternatively, testing all possible sequences of PR types (with M_h and M_l fixed) would also be extremely costly in terms of computation. For instance, suppose that $M_h = 30$ and $M_l = 70$; in this case, there would be more than 2.93×10^{25} possible sequences to test. In reality, considering that the order of PRs waiting for sponsorship are thousands (if not tens of thousands), this brute-force computational strategy would not be wise.

Despite this, there exist ways to significantly decrease the number of sequences to test. One of the foundations to do so is to express the function V (i.e., the ex-post payoff of the CSO) recursively. That is, given $\mathcal{I}$, ℓ and b, one can express $V(\mathcal{I}, \ell, b)$ using another set of expressions involving V and ordered subsequences of $\mathcal{I}$ and ℓ. In order to do so, let s_{-i} denote the resulting subsequence by removing the ith element from an arbitrary ordered sequence s, in which s_{-i} retains the order in s. Also, let s_i be the ith element from the said sequence s. Then, $V(\mathcal{I}, \ell, b)$ can be written as:

$$V(\mathcal{I}, \ell, b) = \begin{cases} 1 + V(\mathcal{I}_{-1}, \ell_{-1}, b) & \text{if } \ell_1 = L \\ V(\mathcal{I}, \ell_{-1}, b) & \text{if } \ell_1 = H \text{ and } \forall i \in \{1, 2, \cdots, b+1\},\ \mathcal{I}_i = t_l \\ 1 + V(\mathcal{I}_{-j}, \ell_{-1}, b) & \text{if } \ell_1 = H,\ \mathcal{I}_j = t_h, \text{ and } \forall i \in \{1, 2, \cdots, j-1\},\ \mathcal{I}_i = t_l \end{cases} \tag{6}$$

The cases in (6) can be explained as follows: the first case corresponds to having an L-type PS enter the inventory and matching with whoever he or she sees; the second case corresponds to an H-type PS being unable to find a suitable t_h-type PR to be matched with, and quitting; and finally, the third case corresponds to an H-type PS finding a suitable t_h-type PR and matching with the said PR. Expression (6) will help in the theoretical expositions in this part and serve as a building block for the numerical simulations.

Using (6), one may also write a recursive formula for EV, which is as follows:

$$EV(\mathcal{I}, N, \mu, b) = (1 - \mu)\big(1 + EV(\mathcal{I}_{-1}, N - 1, \mu, b)\big) + \mu f(\mathcal{I}, N, \mu, b) \tag{7}$$

where

$$
\begin{aligned}
f(\mathcal{I}, N, \mu, b) &:= \iota(\mathcal{I}, b) \cdot EV(\mathcal{I}_{-1}, N-1, \mu, b) + \sum_{j=1}^{b+1} \hat{\iota}_j(\mathcal{I}, b) \cdot \big(1 + EV(\mathcal{I}_{-j}, N-1, \mu, b)\big), \\
\iota(\mathcal{I}, b) &:= \mathbb{1}(\mathcal{I}_i = t_l \ \forall i \in \{1, \cdots, b+1\}), \\
\text{and} \quad \hat{\iota}_j(\mathcal{I}, b) &:= \mathbb{1}(\mathcal{I}_j = t_h \text{ and } \mathcal{I}_i = t_l \ \forall i \leq j).
\end{aligned}
\tag{8}
$$

where the function f is introduced to denote the expected payoff after the first PS has been found as an H-type, and $\mathbb{1}(\cdot)$ is a characteristic function.

B. "Bunching" Strategies

One set of strategies that the CSO can employ is having all of the t_h-types aggregated as a one big block to appear in the inventory. I will dub this type of strategy as "bunching." A more formal definition is provided in Defintion 2.1.

Definition 2.1. *(Bunching)* Let $\mathcal{I}$ be an ordered sequence of M_h type-t_h PRs and M_l type-t_l PRs, where $M_l > M_h$ and $M_l + M_h = M$. Then $\mathcal{I}$ is *bunched* if, for any $i, j \in \{1, 2, \cdots, M\}$ with $i < j$ and $\mathcal{I}_i = \mathcal{I}_j = t_h$, any $k \in \{i, i+1, \cdots, j\}$ satisfies $\mathcal{I}_k = t_h$.

Further, if $\mathcal{I}$ is bunched and the first t_h to appear is at $\mathcal{I}_i$ with the last at $\mathcal{I}_j$, then $\mathcal{I}$ is noted to be bunched with a *bunching block* from i to j. A length-M ordered sequence with bunching block from i to j is denoted as $\mathcal{B}^M_{i:j}$.

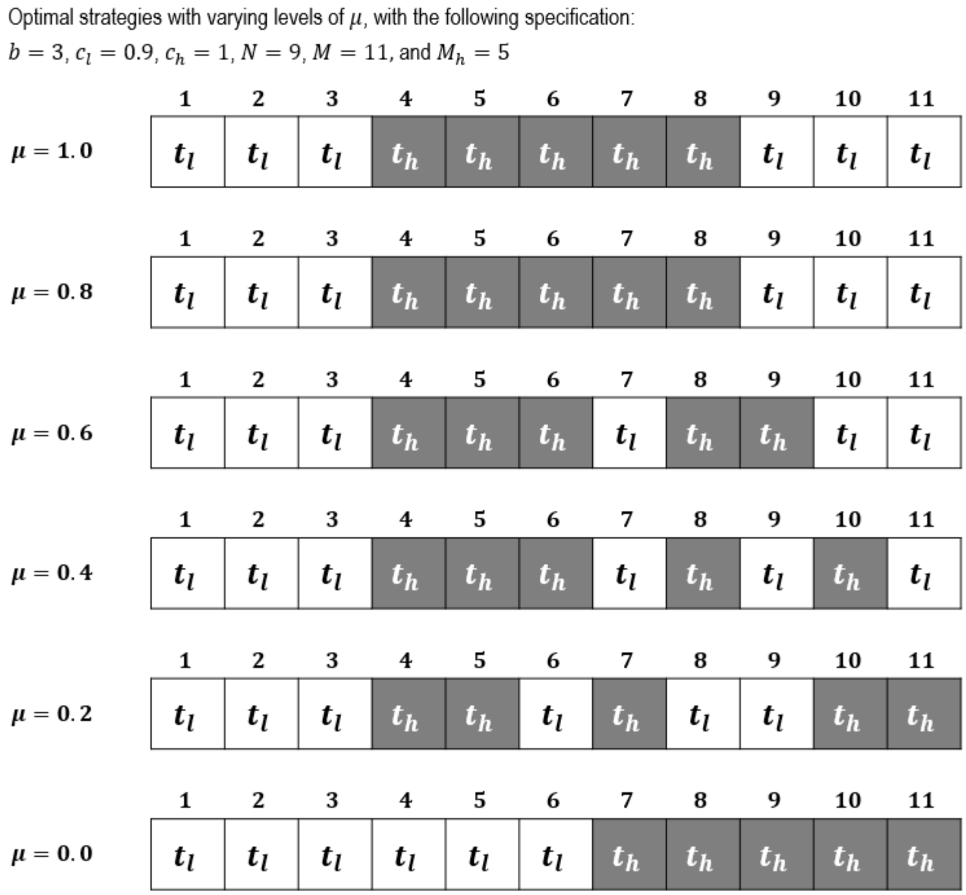
It should be noted that, in many cases, bunching strategies are not always the most optimal ways of designing the inventory. An example can be seen in Figure 3, where it may be necessary to split the bunching block into many parts to attain the maximum expected payoff depending on the levels of μ. However, as it can also be observed from Figure 3, certain bunching strategies – namely, $\mathcal{B}^M_{b+1:b+M_h}$ and $\mathcal{B}^M_{M_l+1:M}$ – become optimal strategies as μ approaches 1 and 0, respectively. This result is stated and proved in Proposition 2.1.

Proposition 2.1. *(Optimal strategies as $\mu \to 0$ and $\mu \to 1$)* Suppose that there are M_h and M_l t_h-type and t_l-type PRs, with $M = M_h + M_l$ and $M_h < M_l$. Further suppose that the probability of an H-type PS arriving is $\mu \in [0, 1]$, and that there are N PSs to arrive and browse the inventory ($N \ll M$); an H-type PS re-searches up to b times. Utility of the CSO from matching is assumed to be 1 per PR; the cost of having unmatched t_h-type (t_l-type) PR is assumed to be c_h (c_l) per PR.

Then, an optimal strategy as $\mu \to 1$ is $\mathcal{B}^M_{b+1:b+M_h}$, and that as $\mu \to 0$ is $\mathcal{B}^M_{M_l+1:M}$.

Proof. Consider first the case where $\mu \to 1$. Then, notice that the probabilities in which any L-type PS out of N PSs appears approaches 0, and that in which N H-type PSs appear tends to 1. In the latter case, a class of optimal strategies that can be used to maximize (deterministic) payoff is bunching, specifically in the form of $\mathcal{B}^M_{k+1:k+M_h}$ where $k \in \{0, 1, \cdots, b\}$. This finding comes from the following set observations: (1) if $M_h > N$, then the said class of strategies produces N matches of t_h-type PRs and H-type PSs, yielding a

Figure 3: Examples of Optimal Strategies with Varying Values of μ



payoff of $N - M_l c_l - (M_h - N)c_h$; (2) if $M_h \leq N$, then the said class produces M_h matches of t_h-type PRs and H-type PSs, yielding a payoff of $M_h - M_l c_l$; (3) payoffs found in observations (1) and (2) are maximum attainable payoff in their respective situations; and (4) any strategy that deviates from the said class by exchanging one of the t_l-types at position $\hat{k} > k + M_h$ (where $k \in \{0, 1, \cdots, b\}$) with a t_h-type at position $\tilde{k} \in \{k+1, k+2, \cdots, k+M_h\}$ is either indifferent or inferior to the original strategy, as it may increase the chance that an H-type PS not being able to find a t_h-type PR before one chooses to quit.

Among the bunching strategies that are optimal when N H-type PSs arrive, notice that $\mathcal{B}_{b+1:b+M_h}$ minimizes the probability that an L-type PS, instead of an H-type, matches with one of the t_h-type PRs. Comparing the first instance of t_h-types, for instance, the probability of an L-type taking the said first instance is $(1-\mu)^k$ if the bunching strategy use is $\mathcal{B}_{k+1:k+M_h}$ (for $k \in \{0, 1, \cdots, b\}$). Therefore, if any probability of an L-type appearing is close to yet not zero (i.e., $\mu < 1$), $\mathcal{B}_{b+1:b+M_h}$ is weakly better off than $\mathcal{B}_{k+1:k+M_h}$ for $k \in \{0, 1, \cdots, b-1\}$.

Next, consider the case where $\mu \to 0$. Notice the probabilities of any case where an H-type appears approach 0, and that of the case where all N appearing PSs are L-types tends to 1. In the case where N L-type PSs appear, note that the maximum payoff attainable is $N - (M-N)c_l$ if $N > M_l$ and $N - M_h c_h - (M_l - N)c_l$

if $N \leq M_l$. The said maximum payoffs are attained by having as many t_l-types match with the PSs, and have the remaining PSs match with t_h-types if there are more PSs than t_l-types than there are PSs (i.e., $M_l > N$). Therefore, as the probability that only L-type PSs appear approaches 1 (and other cases' probabilities to 0), the optimal strategy tends to $\mathcal{B}_{M_l+1:M}$. $\square$

The underlying economic intuition behind Proposition 2.1 is as follows. As seen in the part for model setup, the CSO wants to maximize the number of sponsorship matches, but is afraid of two interlinked situations. Firstly, it may be worried that an L-type PS enters the inventory to find a t_h-type PR and matches with the said PR; as a consequence, an H-type PS later entering the inventory may lose the opportunity to be matched with a PR. Secondly, the CSO may also be concerned about the H-type PSs being unable to find t_h-type PRs due to the placement of the initial t_h-type PR being too far from the start. Considering the two situations, the CSO would want to start far enough from the start so that it is less reachable by an L-type PS (if any exists), but at the same time close enough so that an H-type PS (again, if any exists) can enter the inventory without quitting. This far-away-yet-close starting point for the bunching block is shown to be at $b+1$ as $\mu \to 1$, according to Proposition 2.1. On the other hand, if only L-types are entering the inventory (i.e., $\mu \to 0$), then it would be wiser to have all t_h-types as unreachable as possible; this is embodied in the strategy $\mathcal{B}^M_{M_l+1:M}$.

As exemplified in Figure 3, bunching strategies are not the most optimal ones under various circumstances. Despite this fact, the CSO may still be interested in choosing among bunching strategies (even outside of $\mathcal{B}^M_{b+1:b+M_h}$ and $\mathcal{B}^M_{M_l+1:M}$, mentioned in Proposition 2.1). In addition to their simplicity, there may be reasons like limitations in computing capabilities, as the number of total cases to examine increases quickly with M_h and M_l. In light of this challenge, I identify the optimal strategy among bunching strategies in Proposition 2.2.

Proposition 2.2. *(Optimal bunching strategy)* Continue using the setting of Proposition 2.1 with respect to $M_h, M_l, M, \mu, N, b, c_l$, and c_h. Then, an optimal (in terms of yielding the greatest expected payoff) strategy among the bunching strategies is $\mathcal{B}^M_{b+1:b+M_h}$ if $\mu > 0$. In the case of $\mu = 0$, an optimal bunching strategy is $\mathcal{B}^M_{M_l+1:M}$.

Proof. Firstly, note that if $\mu = 1$ or $\mu = 0$, the results carry on from Proposition 2.1.

For $\mu \in (0, 1)$, I use induction to show that $\mathcal{B}^M_{b+1:b+M_h}$ is the optimal bunching strategy. For the sake of simplicity, let $EV(\cdot, \cdot, \mu, b) \equiv EV(\cdot, \cdot)$ with μ and b fixed. Also, let us denote $\mathcal{B}^M_{b+1:b+M_h}$ as $\hat{\mathcal{B}}$.

Let us first show that, for all $k \in \{1, 2, \cdots, b\}$, $\mathcal{B}^M_{b+1-k:b+M_h-k}$ is an inferior bunching strategy to $\hat{\mathcal{B}}$. Note that, using expression (7),

$$EV(\hat{\mathcal{B}}, N) = (1-\mu)\big(1 + EV(\mathcal{B}^{M-1}_{b:b+M_h-1}, N-1)\big) + \mu\big(1 + EV(\mathcal{B}^{M-1}_{b+1:b+M_h-1}, N-1)\big). \tag{9}$$

But further notice that $EV(\mathcal{B}^{M-1}_{b:b+M_h-1}, N-1)$ is related to $EV(\mathcal{B}^{M}_{b:b+M_h-1}, N)$ in the following manner:

$$
\begin{aligned}
EV(\mathcal{B}^{M-1}_{b:b+M_h-1}, N-1) &= EV(\mathcal{B}^{M}_{b:b+M_h-1}, N-1) + c_l \\
&= EV(\mathcal{B}^{M}_{b:b+M_h-1}, N) + c_l - \alpha
\end{aligned}
\tag{10}
$$

where $\alpha > 0$ is used to reflect that an additional match can be yielded by adding another PS whose type the CSO is uncertain of. Note that $\alpha \geq (1-\mu)(1+c_l)$, since an L-type PS may enter with probability $1-\mu$ and match with a t_l-type PR. However, $\alpha < 1 + c_l$, since an H-type PS may enter and quit due to being unable to find a t_h-type PR.

Therefore, (9) can be rewritten with (10) as follows:

$$
\begin{aligned}
& EV(\hat{\mathcal{B}}, N) = (1-\mu)EV(\mathcal{B}^{M}_{b:b+M_h-1}, N) + \underbrace{\{1 + (1-\mu)(c_l - \alpha)\}}_{=:\alpha_0 \geq \mu} + \mu EV(\mathcal{B}^{M-1}_{b+1:b+M_h-1}, N-1) \\
\Rightarrow\ & EV(\hat{\mathcal{B}}, N) - EV(\mathcal{B}^{M}_{b:b+M_h-1}, N) = \mu\big[EV(\mathcal{B}^{M-1}_{b+1:b+M_h-1}, N-1) - EV(\mathcal{B}^{M}_{b:b+M_h-1}, N)\big] + \alpha_0
\end{aligned}
\tag{11}
$$

Note further that, using the logic in (10), one may write

$$
\begin{aligned}
EV(\mathcal{B}^{M-1}_{b+1:b+M_h-1}, N-1) &= EV(\mathcal{B}^{M}_{b+1:b+M_h-1}, N) + c_l - \alpha \\
&= EV(\mathcal{B}^{M}_{b:b+M_h-1}, N) + c_l - \alpha + \hat{c}
\end{aligned}
\tag{12}
$$

where $\hat{c} \in [c_l, c_h]$ and stands for the expected cost of changing the bth element of $EV(\mathcal{B}^{M}_{b:b+M_h-1}, N)$ to t_l; there may be an H-type entering the inventory to be unable to find a t_h-type due to removing one PR of such type.

With the developments in (11) and (12), I write

$$
\begin{aligned}
EV(\hat{\mathcal{B}}, N) - EV(\mathcal{B}^{M}_{b:b+M_h-1}, N) &= \alpha_0 + \mu c_l - \mu\alpha + \mu\hat{c} \\
&= \underbrace{(1 + c_l) - \alpha}_{>0} + \mu\hat{c} > 0.
\end{aligned}
\tag{13}
$$

Therefore, $EV(\hat{\mathcal{B}}, N) > EV(\mathcal{B}^{M}_{b:b+M_h-1}, N)$; this will serve as the base case for induction.

For the inductive step, suppose that $EV(\mathcal{B}^{M}_{b+1-k:b+M_h-k}, N) > EV(\mathcal{B}^{M}_{b-k:b+M_h-(k+1)}, N)$ for some arbitrary $k \in \{0, 1, \cdots, b-1\}$. For simplicity, let us denote $\hat{b} := b - k$. Similar to the logic in expressions (9) to (13), one may write:

$$
\begin{aligned}
EV(\mathcal{B}^{M}_{\hat{b}:\hat{b}+M_h-1}, N) &= 1 + (1-\mu)EV(\mathcal{B}^{M-1}_{\hat{b}-1:\hat{b}+M_h-2}, N-1) + \mu EV(\mathcal{B}^{M-1}_{\hat{b}:\hat{b}+M_h-2}, N-1) \\
&= EV(\mathcal{B}^{M}_{\hat{b}-1:\hat{b}+M_h-2}, N) + (1 + c_l) - \alpha + \mu\hat{c} > EV(\mathcal{B}^{M}_{\hat{b}-1:\hat{b}+M_h-2}, N).
\end{aligned}
\tag{14}
$$

That is, $EV(\mathcal{B}^{M}_{b-k:b+M_h-(k+1)}, N) > EV(\mathcal{B}^{M}_{b-k-1:b+M_h-(k+2)}, N)$. This inequality completes the inductive step and proves that, for all $k \in \{1, 2, \cdots b\}$, $EV(\hat{\mathcal{B}}, N) > EV(\mathcal{B}^{M}_{b+1-k:b+M_h-k}, N)$.

Next, let us show in similar spirit that for all $k \in \{1, 2, \cdots, M_l - b\}$, $\mathcal{B}^{M}_{b+1+k:b+M_h+k}$ is inferior to $\hat{\mathcal{B}}$. Akin to expression (9), one may write:

$$EV(\mathcal{B}^M_{b+2:b+M_h+1}, N) = (1-\mu) + (1-\mu)EV(\mathcal{B}^{M-1}_{b+1:b+M_h}, N-1) + \mu EV(\mathcal{B}^M_{b+2:b+M_h+1}, N-1) \tag{15}$$

where, as opposed to (9), the case with H-type PS entering first yields no match, therefore not altering the inventory.

Similar to (10), $EV(\mathcal{B}^{M-1}_{b+1:b+M_h}, N-1)$ and $EV(\mathcal{B}^M_{b+2:b+M_h+1}, N-1)$ can be written as:

$$\begin{aligned} EV(\mathcal{B}^{M-1}_{b+1:b+M_h}, N-1) &= EV(\mathcal{B}^M_{b+1:b+M_h}, N) + c_l - \alpha \\ EV(\mathcal{B}^M_{b+2:b+M_h+1}, N-1) &= EV(\mathcal{B}^M_{b+2:b+M_h+1}, N) - \alpha. \end{aligned} \tag{16}$$

and plugging these expressions into (15) yields:

$$\begin{aligned} &EV(\mathcal{B}^M_{b+2:b+M_h+1}, N) = (1-\mu)(1+c_l-\alpha) - \mu\alpha + (1-\mu)EV(\mathcal{B}^M_{b+1:b+M_h}, N) + \mu EV(\mathcal{B}^M_{b+2:b+M_h+1}, N) \\ \Rightarrow\ &(1-\mu)\big\{EV(\mathcal{B}^M_{b+2:b+M_h+1}, N) - EV(\hat{\mathcal{B}}, N)\big\} = (1-\mu)(1+c_l-\alpha) - \mu\alpha \\ \Rightarrow\ &EV(\mathcal{B}^M_{b+2:b+M_h+1}, N) - EV(\hat{\mathcal{B}}, N) = (1+c_l-\alpha) - \frac{\mu}{1-\mu}\alpha < (1-\mu)(1+c_l) - \alpha < 0 \end{aligned} \tag{17}$$

where I use that $\alpha > (1-\mu)(1+c_l)$. Therefore, $EV(\hat{\mathcal{B}}, N) > EV(\mathcal{B}^M_{b+2:b+M_h+1}, N)$ and the base case for induction is established.

For the inductive step, suppose that $EV(\mathcal{B}^M_{b+1+k:b+M_h+k}, N) > EV(\mathcal{B}^M_{b+2+k:b+M_h+k+1}, N)$ for arbitrary $k \in \{0, 1, \cdots, M_l - 1 - b\}$. Again for simplicity, let us denote $\tilde{b} := b + k$. Then, one can express $EV(\mathcal{B}^M_{b+3+k:b+M_h+k+2}, N) \equiv EV(\mathcal{B}^M_{\tilde{b}+3:\tilde{b}+M_h+2}, N)$ as follows, using a similar logic to that applied in the expressions (15) to (17):

$$\begin{aligned} &EV(\mathcal{B}^M_{\tilde{b}+3:\tilde{b}+M_h+2}, N) = (1-\mu)(1+c_l-\alpha) - \mu\alpha + (1-\mu)EV(\mathcal{B}^M_{\tilde{b}+2:\tilde{b}+M_h+1}, N) + \mu EV(\mathcal{B}^M_{\tilde{b}+3:\tilde{b}+M_h+2}, N) \\ \Rightarrow\ &EV(\mathcal{B}^M_{\tilde{b}+3:\tilde{b}+M_h+2}, N) - EV(\mathcal{B}^M_{\tilde{b}+2:\tilde{b}+M_h+1}, N) < (1-\mu)(1+c_l) - \alpha < 0 \end{aligned} \tag{18}$$

and therefore, $EV(\mathcal{B}^M_{b+2+k:b+M_h+k+1}, N) > EV(\mathcal{B}^M_{b+3+k:b+M_h+k+2}, N)$. This completes the induction and proves that, for all $k \in \{1, 2, \cdots, M_l - b\}$, $EV(\hat{\mathcal{B}}, N) > EV(\mathcal{B}^M_{b+1+k:b+M_h+k}, N)$. In summary, out of the bunching strategies, $\hat{\mathcal{B}} \equiv \mathcal{B}^M_{b+1:b+M_h}$ yields the greatest expected payoff provided that N, μ, b, M_h, and M_l are fixed. $\square$

The intuition used in Proposition 2.1 continues to hold in Proposition 2.2. Again, the CSO wants to make the t_h-type PRs close enough so that H-types can access the said PRs while far enough so that they are less likely to be affected by L-types. This far-away-yet-close starting point for the bunching block is shown to be at $b+1$ (given that $\mu \neq 1$), according to Proposition 2.2.

C. Departure from Bunching Strategies

One of the reasons mentioned for using bunching strategies over others was the difficulty and costliness in computation for searching over all the possible combinations of M_l t_l-types and M_h t_h-types. However, it is still possible to reduce number of strategies provided that $b > 0$. Theorem 2.2 helps the CSO in doing so by suggesting that some of the potential strategies are weakly inferior to small alterations to them, and therefore sub-optimal. The following lemma (Lemma 2.1) is integral to proving Theorem 2.2.

Lemma 2.1. *(Substitution of types)* Continue using the setting of Proposition 2.1 with $M_h, M_l, M, \mu, N, b, c_l$, and c_h. Denote the mth element of inventory $\mathcal{I}$ as $\mathcal{I}_m$; denote the ordered subsequence of $\mathcal{I}$ from the mth to the m'th elements as $\mathcal{I}_{m:m'}$ Then, consider the following cases (a)-(c):

(a) Suppose that inventory $\mathcal{I}$ is only composed of M t_l-type PRs. Consider the alteration in which the kth position is changed to a t_h-type, and denote this alteration as $\mathcal{I}^k$.

(b) Suppose that inventory $\mathcal{I}$ is composed of both t_h- and t_l-types, and let $\mathcal{I}_p = t_h$ with $\mathcal{I}_{\hat{p}} = t_l$ for all $\hat{p} < p$. Consider the alteration in which the kth position, where $k < p$ and $\mathcal{I}_k = t_l$, is changed to a t_h-type; denote this alteration as $\mathcal{I}^k$.

(c) Let $\mathcal{I}$ and p be defined as in case (b). Consider the alteration in which the kth position, where $k > p$ and $\mathcal{I}_k = t_l$, is changed to a t_h-type; denote this alteration as $\mathcal{I}^k$.

Then, in all cases (a)-(c), $\mathcal{I}^k$ is weakly better than $\mathcal{I}$ in terms of expected payoff (i.e., the EV function value).

Proof. I provide the proofs for each case as follows:

Case (a). Suppose that $N < k$. Then, in any case, both $(\mathcal{I}^k)_k = t_h$ and $\mathcal{I}_k = t_l$ do not yield a match. This means that the said PR types incur costs of c_h and $c_l (> c_h)$, respectively, meaning that the expected payoff is strictly greater for $\mathcal{I}^k$.

On the other hand, suppose that $N \geq k$. If $\mu > 0$, then the probability of an H-type entering eventually to be matched with $(\mathcal{I}^k)_k = t_h$ is nonzero. This leads to nonnegative contribution to the expected payoff for $\mathcal{I}^k$ as opposed to $\mathcal{I}$. In all other cases (i.e., H-type PSs never entering), the expected payoffs pertaining to such cases for both $\mathcal{I}$ and $\mathcal{I}^k$ are the same. If $\mu = 0$, then expected payoffs for $\mathcal{I}$ and $\mathcal{I}^k$ are exactly the same. Therefore, $\mathcal{I}^k$ yields weakly greater expected payoff than $\mathcal{I}$.

Case (b). Again, similar to case (a), the inventory designer faces the problem of storing one less t_h-type in with $\mathcal{I}$ as opposed to $\mathcal{I}^k$ in the instance where $N < k$. Therefore, expected payoff is greater for $\mathcal{I}^k$ than for $\mathcal{I}$.

On the other hand, suppose $N \geq k$. Notice first that $\mathcal{I}_{1:(k-1)} = (\mathcal{I}^k)_{1:(k-1)}$. and $\mathcal{I}_{(k+1):M} = (\mathcal{I}^k)_{(k+1):M}$ since it was only the kth element of $\mathcal{I}$ being altered to make $\mathcal{I}^k$. There can be three sub-cases in this case. First is where no PS reaches the kth elements; this sub-case, similar to the instance of $N < k$, will incur more cost for $\mathcal{I}$ than $\mathcal{I}^k$ due to having one more t_l-type than t_h-type. Another sub-case is where an L-type PS reaches the kth elements. Due to $\mathcal{I}_{1:(k-1)} = (\mathcal{I}^k)_{1:(k-1)}$, expected payoffs from the first $(k-1)$ elements will be the same for both the original and altered inventory. Further, by an L-type being matched with the kth elements (yielding payoff of 1 for certain) and by $(k+1)$th to Mth elements being the same, the expected payoffs for the remaining $M - (k-1)$ elements would be equal across $\mathcal{I}$ and $\mathcal{I}^k$ as well.

The final sub-case is where the kth element is met by an H-type PS; let us denote the remaining inventories after removing the first $k-1$ elements and after this H-type encounter from $\mathcal{I}$ and $\mathcal{I}^k$ as $\mathcal{I}'$ and $(\mathcal{I}^k)'$. Notice that since $\mathcal{I}_k$ was not matched and $(\mathcal{I}^k)_k$ was, $\mathcal{I}^k$ surely has 1 greater payoff than $\mathcal{I}$ in this sub-case. Also, notice that there are only two possibilities for $\mathcal{I}'$. The first possibility is that $\mathcal{I}'$ is exactly equal to attaching a t_l-type PR in front of $(\mathcal{I}^k)'$ The expected payoff from attaching this t_l-type (or, that from substituting $(\mathcal{I}^k)'$ with $\mathcal{I}'$) is $(1-\mu)\cdot 1 - \mu \cdot c_l = 1-(1+c_l)\mu$. Notice that $1-(1+c_l)\mu$ is a convex combination of a negative value (i.e., $-c_l$) and 1, so 1 is weakly greater than it. This means that

$$EV(\mathcal{I}^k, N) - EV(\mathcal{I}, N) \geq 1 - 1 + (1+c_l)\mu \geq 0$$

and that $\mathcal{I}^k$ is at least weakly better than $\mathcal{I}$.

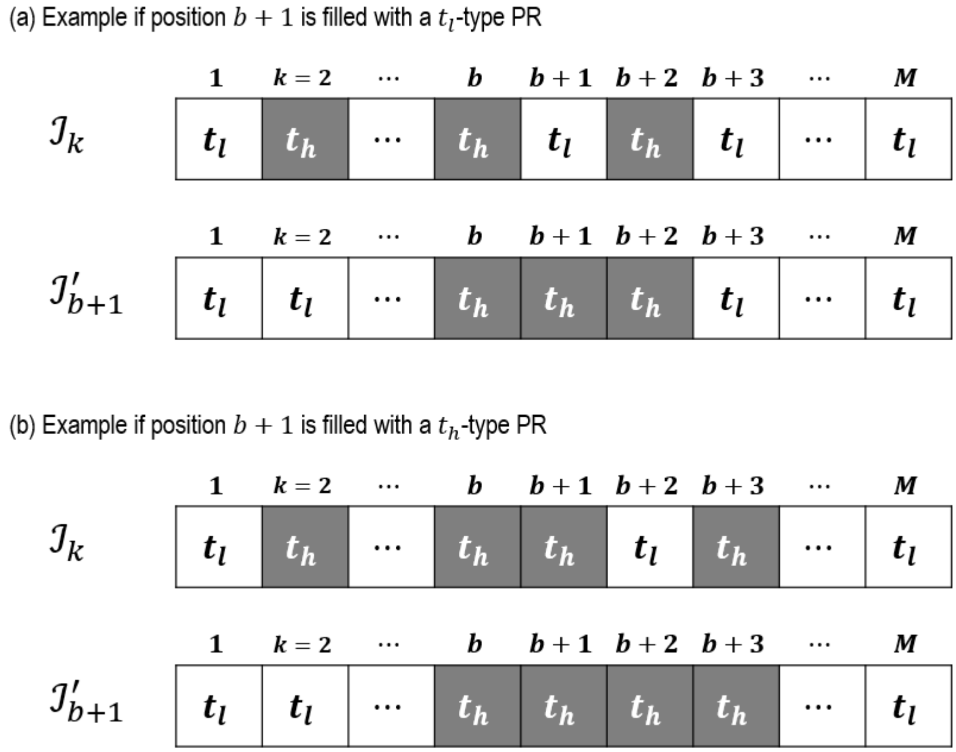
The second possibility is that $(\mathcal{I}^k)'$ is an alteration from $\mathcal{I}'$ in that one of the t_l-types left of the leftmost t_h-type in $\mathcal{I}'$ is altered to a t_h-type PR. Notice that this is exactly the situation that case (c) covers; therefore, assuming that the claim for case (c) also holds, $(\mathcal{I}^k)'$ is weakly better than $\mathcal{I}'$. This leads to $\mathcal{I}^k$ being weakly better that $\mathcal{I}$. Therefore, $\mathcal{I}^k$ yields weakly greater expected payoff than $\mathcal{I}$ in all sub-cases of case (b) (assuming that the claim for case (c) holds).

Case (c). For the instance where $N < k$, one can very similarly show (in reference to cases (a) and (b)) that $\mathcal{I}^k$ is strictly better than $\mathcal{I}$. Therefore, suppose that $N \geq k$. However, notice that case (c)'s logic for $\mathcal{I}^k$ being weakly better than $\mathcal{I}^k$ is much akin to that in case (b). If the kth elements are never reached, the logic used in $N < k$ can be applied; if the kth elements are met by an L-type, the remaining ordered elements are same for $\mathcal{I}^k$ and $\mathcal{I}$. Finally, if the kth elements are met by an H-type, then while $(\mathcal{I}^k)_k$ is matched, $\mathcal{I}_k$ would not; therefore, one can refer to the two possibilities explored in case (b), in which $(\mathcal{I}^k)'$ is either achieved from removing one t_l-type from $\mathcal{I}'$ or by changing one of the t_l-types left of the leftmost t_h-type in $\mathcal{I}'$ to a t_h-type.

As explained in case (b), the second possibility is itself another instance of case (c). Therefore, by comparing $(\mathcal{I}^k)'$ and $\mathcal{I}'$, one of the sub-cases and possibilities previously mentioned will eventually be reached. Notice that this second possibility cannot be encountered forever, since either the number of PSs to visit or that of the t_h-type PRs will run out for the inventories developing from $\mathcal{I}$. In the former instance, one can refer to the situation where $N < k$. In the latter instance, notice that the inventory resulting from $\mathcal{I}$ has the same length as that resulting from $\mathcal{I}^k$ but does not have any t_h-types. This is equivalent to examining case (a), since the inventory resulting from $\mathcal{I}^k$ has exactly one more t_h-type than that from $\mathcal{I}$. Therefore, in all sub-cases and possibilities of case (c), $\mathcal{I}^k$ is weakly better than $\mathcal{I}$. □

Lemma 2.1 states that replacing a t_l-type with a t_h-type is advantageous for the inventory designer regardless of the position of the t_l-type substituted. The intuition behind this lemma is simple: t_h-types incur less cost when unmatched, and is acceptable to all types of PSs provided that they are reachable. The proof process of the said lemma is extensive, just to be rigorous enough in covering all instances of PS arrivals. Further, notice that the results of Lemma 2.1 are not itself useful for inventory optimization of this paper due to the inability of altering the composition of PR types (i.e., changing M_h and M_l given M). Despite this fact, the lemma is crucial for deriving the conclusion in Theorem 2.2.

Figure 4: Example “Alterations” for Theorem 2.2



I now write Theorem 2.2, which states that any strategy with a t_h-type to the left of $(b+1)$th position is either sub-optimal or has another strategy with the same expected payoff even if optimal.

Theorem 2.2. *(Removal of strategies with t_h-types at positions $k < b+1$)* Continue using the setting of Proposition 2.1 with respect to $M_h, M_l, M, \mu, N, b, c_l$, and c_h. Further suppose that $\mathcal{I}^k$ is an arbitrary inventory with the leftmost t_h-type at position $k \in \{1, 2, \cdots, b\}$. Then, there always exists a strategy that is weakly better than $\mathcal{I}^k$. Specifically, if $\mu \in (0,1)$, the said strategy is obtained by exchanging $(\mathcal{I}^k)_k = t_h$ with the closest t_l-type on or beyond position $b+1$.

Proof. For the cases $\mu = 0$ and $\mu = 1$, notice that by Proposition 2.1, optimal strategies are $\mathcal{B}^M_{b+1:b+M_h}$ and $\mathcal{B}^M_{M_l+1:M}$, respectively. The said bunching strategies' leftmost t_h-type PR is at positions $b+1$ and M_l+1, respectively. Therefore, trivially, these strategies weakly dominate any $\mathcal{I}^k$.

Now consider $\mu \in (0,1)$, and also consider the following alteration from $\mathcal{I}^k$. With all PR types the same, exchange the t_h-type at position k with the t_l-type at position $b+1$ or greater that is closest to position k. An illustrative example is provided in Figure 4. Let us denote the said position of the closest t_l-type at or beyond $b+1$ as $\hat{k}$, and denote the altered inventory from $\mathcal{I}^k$ as $\mathcal{I}^{(k,\hat{k})}$.

The goal is to show that $\mathcal{I}^{(k,\hat{k})}$ is weakly better-off than $\mathcal{I}^k$. In order to see this result, consider the following cases:

(i) The kth elements for $\mathcal{I}^k$ and $\mathcal{I}^{(k,\hat{k})}$ are met by an L-type PS.

(ii) The kth elements for $\mathcal{I}^k$ and $\mathcal{I}^{(k,\hat{k})}$ are met by an H-type PS.

First, consider case (i). If the said L-type PS arrives, it will match with the PRs at the kth position regardless of type. Let us denote the remaining ordered sequences after the said matches are made from $\mathcal{I}^k$ and $\mathcal{I}^{(k,\hat{k})}$ as $(\mathcal{I}^k)'$ and $(\mathcal{I}^{(k,\hat{k})})'$, respectively. Then, notice that $(\mathcal{I}^k)'$ has one less t_h-type PS than $(\mathcal{I}^{(k,\hat{k})})'$; by Lemma 2.1, this means that the expected payoff from $(\mathcal{I}^{(k,\hat{k})})'$ is weakly greater than that from $(\mathcal{I}^k)'$. Therefore, with case (i), $\mathcal{I}^k$ is weakly worse-off than $\mathcal{I}^{(k,\hat{k})}$.

With case (ii), notice that the said H-type PS would have matched with $(\mathcal{I}^k)_k = t_h$ for $\mathcal{I}^k$, but would have matched with the leftmost t_h-type *not* at position k but at $k' \leq b+1$ for $\mathcal{I}^{(k,\hat{k})}$. Therefore, even after the arrival of the said H-type, the remaining inventory from $\mathcal{I}^k$ always has its leftmost t_h-type more to the left than that from $\mathcal{I}^{(k,\hat{k})}$. This more-to-the-left positioning of the leftmost t_h-type exposes $\mathcal{I}^k$ to the threat of an L-type PS entering and matching with a t_h-type as opposed to matching with a t_l-type in $\mathcal{I}^(k,\hat{k})$; by Lemma 2.1, the expected payoff is weakly lesser for the remaining inventory from $\mathcal{I}^k$ than for that from $\mathcal{I}^{(k,\hat{k})}$. If none of such cases – where an L-type entering matches with a t_l-type in $\mathcal{I}^{(k,\hat{k})}$ but with a t_h-type in $\mathcal{I}^k$ – happen, then it would mean that all t_h-types on or to left of $\hat{k}$th position are matched with H-type PSs, in which the remaining inventories from both $\mathcal{I}^{(k,\hat{k})}$ and $\mathcal{I}^k$ would be equal to one another. Therefore, in any case, $\mathcal{I}^{(k,\hat{k})}$ is weakly better than $\mathcal{I}^k$ in terms of expected utility.

In conclusion, there always exists a strategy weakly better than $\mathcal{I}^k$. $\square$

The idea behind Theorem 2.2 is straightforward. If $\mu \in (0,1)$, there is uncertainty about the PS types for each PS' arrival. As articulated previously, there are negative consequences to an L-type matching with a t_h-type PR as such an action deprives an H-type to potentially match with the said t_h-type. By moving a t_h-type PR in potential threat of being matched with an L-type to somewhere close enough that an H-type can (eventually) reach (i.e., from $\mathcal{I}^k$ to $\mathcal{I}^{(k,\hat{k})}$ in the proof of Theorem 2.2), an inventory designer can minimize the said negative consequences.

Theorem 2.2 alludes to the idea that there exists an optimal strategy with the positions 1 to b being t_l-types regardless of different parameters. Progressing one step further, the following proposition (Proposition 2.3) dictates that an inventory designer can pin down the $(b+1)$th position provided that a certain condition is met.

Proposition 2.3. *(Condition for optimally placing t_h-type at position $b+1$)* Continue using the setting of Proposition 2.1 with respect to M_h, M_l, M, μ, N, b, c_l, and c_h. Further suppose that $\mathcal{I}^k$ is an arbitrary inventory with the leftmost t_h-type at position $k \in \{b+2, b+3, \cdots, M\}$. Then, if $\mu \neq 0$ and the condition below is met, then there exists a strategy strictly better than $\mathcal{I}^k$; the said strategy is obtained by exchanging $(\mathcal{I}^k)_{b+1} = t_l$ with $(\mathcal{I}^k)_k = t_h$.

The condition is as follows:

$$\frac{1+c_h}{1+c_l} > \hat{\alpha}(k, N)$$

where

$$\hat{\alpha}(k,N) = \begin{cases} (1-\mu)^{b+1} & \text{if } N < k \\ \left(\frac{1-(1-\mu)^b}{1-(1-\mu)^{b+1}}\right)(1-\mu)^{N+k-2b-3} & \text{if } N = k \\ \frac{(1-\mu)^{N+k-2b-3}\{1-(1-\mu)^b\}}{1-(1-\mu)^{b+1}-b\mu(1-\mu)^{k-b-1}+b\mu(1-\mu)^{k-1}} & \text{if } N > k \end{cases}.$$

Proof. Throughout the proof, I will assume that b is fixed. Firstly, consider comparing the following two inventories. The first inventory is $\mathcal{I}^k$ as defined above, with $k \in \{b+2, b+3, \cdots, M\}$; the second, to be denoted $\hat{\mathcal{I}}^k$, is acquired by removing the element $(\mathcal{I}^k)_k = t_h$ from $\mathcal{I}^k$. Assuming that there are $N' \leq N$ PSs to enter, one can compare two inventories' expected payoffs as follows:

$$EV(\mathcal{I}^k, N') = EV(\hat{\mathcal{I}}^k, N') + \underbrace{\{\rho \cdot 1 - (1-\rho)c_h\}}_{\text{benefit from the added } t_h}$$

where $\rho \in [0,1]$ is the probability that $(\mathcal{I}^k)_k = t_h$ will yield a match. While having been written as the "benefit" from adding a t_h-type (at the kth position), it is not clear whether $\rho - (1-\rho)c_h$ is nonnegative.

To be more precise, ρ is a function of k and N', such that

$$\rho(k,N') = \begin{cases} 0 & \text{if } N' < k \\ \big\{(1-\mu)^{k-b-1} + (1-\mu)^{k-b} + \cdots + (1-\mu)^{k-1}\big\}\mu & \text{otherwise} \end{cases} \tag{19}$$

where

$$\big\{(1-\mu)^{k-b-1} + (1-\mu)^{k-b} + \cdots + (1-\mu)^{k-1}\big\}\mu = (1-\mu)^{k-b-1} - (1-\mu)^{k-1}$$

by geometric series sum. The said term accounts for all possibilities in which an H-type PS appears after at least $(k-b-1)$ L-type PSs arrive, so that the kth position t_h-type PR can be matched with the said H-type PS. However, as written above, if there are less than k PRs to arrive in the first place (i.e., the condition $N' < k$), then the said t_h-type at the kth position cannot be matched at all.

Subsequently, consider comparing the following two inventories. The first is $\mathcal{I}^k$ as in the previous case; the second, denoted $\dot{\mathcal{I}}^k$, is obtained by switching the $(\mathcal{I}^k)_k = t_h$ to a t_l-type PR. By Lemma 2.1, one can notice that $EV(\mathcal{I}^k, N')$ greater than or equal to $EV(\dot{\mathcal{I}}^k, N')$. Let us denote their difference as $\alpha(k,N')$, a function of k and N' by $\alpha(k,N')$; one can write:

$$EV(\mathcal{I}^k, N') = EV(\dot{\mathcal{I}}^k, N') + \alpha(k,N').$$

and

$$\alpha(k,N') = \begin{cases} c_l - c_h & \text{if } N' < k \\ \rho(k,N')\big[1 - \underbrace{\{(1-(1-\mu)^{N'-1}) - (1-\mu)^{N'-1}c_l\}}_{\text{benefit from the added } t_l \text{ in front}}\big] & \text{otherwise,} \end{cases} \tag{20}$$

where ρ is as defined as above. The "benefit" from adding t_l in front is calculated by taking the expectation of t_l-type in front never being matched (with probability $(1-\mu)^{N'-1}$) along with cost c_l and being matched together with the benefit 1. One may re-organize the term above as follows:

$$\rho(k,N')\big[1-\big\{(1-(1-\mu)^{N'-1})-(1-\mu)^{N'-1}c_l\big\}\big]=\rho(k,N')(1+c_l)(1-\mu)^{N'-1}.$$

With this preparation, let us compare the following two inventories. The first inventory remains as $\mathcal{I}^k$; the second, to be denoted $\mathcal{I}^{(k,b+1)}$, is the above-mentioned strategy garnered by exchanging $(\mathcal{I}^k)_k=t_h$ with $(\mathcal{I}^k)_{b+1}=t_l$. Now, consider the case in which an H-type PS arrives after $\nu\le b$ L-type PSs arrive. The probability of such an event occurring is $\mu(1-\mu)^\nu$. Further, within this case, the additional benefit for $\mathcal{I}^{(k,b+1)}$ as opposed to $\mathcal{I}^k$ can be written as:

$$1-\rho(k-\nu,N-\nu-1)+\big(1-\rho(k-\nu,N-\nu-1)\big)c_h=\{1-\rho(k-\nu,N-\nu-1)\}(1+c_h)$$

where the term $\rho(k-\nu,N-\nu-1)-\big(1-\rho(k-\nu,N-\nu-1)\big)c_h$ is the benefit from the added t_h-type analogous to that from the comparison of $\mathcal{I}^k$ and $\hat{\mathcal{I}}^k$.

Notice that there can be multiple cases of the type described above, depending on the value of $\nu\in\{0,1,\cdots,b\}$. On the other hand, there can also be the case where only $b+1$ L-types arrive from the beginning. In this case, the additional benefit for $\mathcal{I}^k$ as opposed to $\mathcal{I}^{(k,b+1)}$ can be written as $\alpha(k-b-1,N-b-1)$. Therefore, gather all cases, the difference between the expected payoffs of $\mathcal{I}^k$ and $\mathcal{I}^{(k,b+1)}$ can be written as:

$$\begin{aligned}\Delta EV:=EV(\mathcal{I}^{(k,b+1)},N)-EV(\mathcal{I}^k,N)=&\sum_{\nu=0}^{b}\mu(1-\mu)^\nu\big\{1-\rho(k-\nu,N-\nu-1)\big\}(1+c_h)\\&-(1-\mu)^{b+1}\alpha(k-b-1,N-b-1)\end{aligned}\tag{21}$$

Let us suppose that $N<k$. Then for all $\nu\in\{0,1,\cdots b\}$, $N-\nu-1<k-\nu$, meaning that $\rho(k-\nu,N-\nu-1)=0$ by (19). Further, by (20), $\alpha(k-b-1,N-b-1)=c_l-c_h$ since $N<k\ \Leftrightarrow\ N-b-1<k-b-1$. Therefore, by using (21) and geometric series' sum, one may write that if $N<k$, then

$$\begin{aligned}\Delta EV&=\big\{1-(1-\mu)^{b+1}\big\}(1+c_h)-(1-\mu)^{b+1}(c_l-c_h)\\&=(1+c_h)-(1-\mu)^{b+1}(1+c_l).\end{aligned}$$

Notice that for ΔEV to be positive in the case of $N<k$, it must be that

$$\frac{1+c_h}{1+c_l}\ge(1-\mu)^{b+1}.\tag{22}$$

Notice that if $\mu=0$, then $\Delta EV=1+c_h-(1+c_l)=c_h-c_l<0$.

Alternatively, consider the case $N=k$. Again, for all $\nu\in\{0,1,\cdots,b\}$, $N-\nu-1<k-\nu$. Therefore,

$\rho(k-\nu, N-\nu-1) = 0$. However, since $N-b-1 \not< k-b-1$,

$$\begin{aligned}\alpha(k-b-1, N-b-1) &= \underbrace{\rho(k-b-1, N-b-1)}_{=(1-\mu)^{k-2b-2}-(1-\mu)^{k-b-2}}(1+c_l)(1-\mu)^{N-b-2} \\ &= (1-\mu)^{N+k-3b-4}\{1-(1-\mu)^b\}(1+c_l).\end{aligned} \tag{23}$$

Therefore, ΔEV can be written as

$$\Delta EV = \{1-(1-\mu)^{b+1}\}(1+c_h) - (1-\mu)^{N+k-2b-3}\{1-(1-\mu)^b\}(1+c_l)$$

and in order to have ΔEV positive in the case of $N = k$, it must be that:

$$\frac{1+c_h}{1+c_l} > \Big(\frac{1-(1-\mu)^b}{1-(1-\mu)^{b+1}}\Big)(1-\mu)^{N+k-2b-3} \tag{24}$$

given that $\mu \neq 0$ (to divide both sides by a non-zero number). If $\mu = 0$, notice that $\Delta EV = 0$.

Finally, let us suppose that $N > k$. Note first that $\rho(k-\nu, N-\nu-1) = (1-\mu)^{k-\nu-b-1} - (1-\mu)^{k-\nu-1}$ since $N-\nu-1 \geq k-\nu$. Then, for any $\nu \in \{0, 1, \cdots, b\}$,

$$\begin{aligned}(1-\mu)^\nu\{1-\rho(k-\nu, N-\nu-1)\} &= (1-\mu)^\nu - \rho(k, N-1) \\ &= (1-\mu)^\nu - (1-\mu)^{k-b-1} + (1-\mu)^{k-1}\end{aligned}$$

Using this result in (21), one may further write that

$$\begin{aligned}\Delta EV =& \{1-(1-\mu)^{b+1} - b\mu(1-\mu)^{k-b-1} + b\mu(1-\mu)^{k-1}\}(1+c_h) \\ &- (1-\mu)^{N+k-2b-3}\{1-(1-\mu)^b\}(1+c_l)\end{aligned}$$

in which the result about $\alpha(k-b-1, N-b-1)$ from (23) is used. Then, as long as the following condition is met in the case of $N > k$ and $\mu \neq 0$ (to divide both sides by a non-zero number), ΔEV will be positive:

$$\frac{1+c_h}{1+c_l} > \frac{(1-\mu)^{N+k-2b-3}\{1-(1-\mu)^b\}}{1-(1-\mu)^{b+1} - b\mu(1-\mu)^{k-b-1} + b\mu(1-\mu)^{k-1}}. \tag{25}$$

Again, if $\mu = 0$, then $\Delta EV = 0$ in this case as well.

Gathering the inequalities in (22), (24), and (25) completes the proof. $\square$

While the term $\hat{\alpha}(k, N)$ is not very clean, notice that it approaches 0 as $\mu \to 1$. Therefore, since $\frac{1+c_h}{1+c_l} \in (0,1)$ (by $0 < c_h < c_l$), the chance of meeting the condition $\frac{1+c_h}{1+c_l} > \hat{\alpha}(k, N)$ in Proposition 2.3 rises as $\mu \to 1$. In this sense, the key idea behind Proposition 2.3 can be summarized as follows. If μ approaches 1 so that the probability of encountering an H-type PS increases, then it would make much sense to place the t_h-type PR at the $(b+1)$th position so that the said PR is reachable by an H-type entering.

Combining Theorem 2.2 and Proposition 2.3, the following corollary (Corollary 2.1) specifies which PR types should be placed in the positions 1 to $b+1$ to yield an optimal strategy.

Corollary 2.1. *(to Theorem 2.2 and Proposition 2.3)* Continue using the setting of Proposition 2.3. Suppose that $\mu \in (0,1)$ and that $\frac{1+c_h}{1+c_l} > \hat{\alpha}(k,N)$, where $\hat{\alpha}(k,N)$ is defined in Proposition 2.3.

Then, there exists an optimal inventory $\mathcal{I}^*$ (in terms of expected payoff) such that $(\mathcal{I}^*)_{b+1} = t_h$ and $(\mathcal{I}^*)_i = t_l$ for all $i \in \{1, 2, \cdots, b\}$.

Corollary 2.1 is the key theoretical result of this paper that may greatly reduce the number of potential strategies to consider depending on the value of b. Because it is an existence theorem, this is not to say that all optimal strategies have to satisfy the condition in Corollary 2.1. However, notice that the inclusive nature of the said corollary is to encompass the more trivial optimization cases such as $\mu \in \{0,1\}$ or $N < b$. In the instance of non-trivial optimization problems, the condition in Corollary 2.1 will help detect sub-optimal strategies.

2.4 Limitations and Potential Extensions

The key limitation of the economic framework described above is that results are highly dependent on the value of b, embodying the H-type PRs' willingness to search further. For instance, if H-types, similar to L-types, are absolutely unwilling to re-search, then at best one can characterize the PR type to be placed at the first position. Therefore, if b is not big enough, then the CSO might instead want to invest in ways to increase the level of b or simply resort to approximating optimal solutions by randomized searches across available strategies. From a more critical point of view, a reader might also question how the inventory designer can pinpoint the values of parameters like N, b, and μ. However, the said criticism is a relatively benign issue as the CSO in reality would have experienced multiple instances of PS arrivals, thereby gathering an insight into or expectation of the said parameters.

A potential extension to further theoretically remove other candidates for optimal inventory might be to identify the maximum number of t_l-types to be placed in between two t_h-types. Following is a yet to be proved claim that may prove to be false:

Claim 2.1. *(Maximum number of t_l-types between two t_h-types)* Continue using the setting of 2.3. Then, there is an optimal strategy where $|i-j| \leq b$ for any two t_h-types at positions i and j with $i \neq j$.

The intuition behind this unproven claim is as follows. If the far-away-yet-close point to have the first H-type PS arriving to acquire a match with a t_h-type is at $b+1$ (i.e., b positions away from the beginning), this rule should also apply for other H-type PSs who arrive later than the initial H-type when the said t_h-type at position $b+1$ is gone by matching. Should the claim be actually true, then it would help remove a number of potential optimal strategy candidates in conjunction with the result of Corollary 2.1. For instance, consider the simple case of $M_l = 12$, $M_h = 8$, $b = 4$, and $\mu \in (0,1)$. Without any guidance, there are 40,320 ($= \frac{20!}{12!8!}$) cases to choose from; with Corollary 2.1 (provided that its conditions are met), one may further reduce the number of cases to 6,435; with Claim 2.1, the cases to consider further reduces to 6120 (approximately 15.2% of the original number of cases). A follow-up theoretical study can be aimed at proving (or finding counterexample to) Claim 2.1.

3 Data and Empirical Framework

There are mainly two parts to the empirical framework of this paper: (1) assessing which features (variables) contribute to (or hinder) matching and (2) verifying whether there is evidence of variability in the said features as well as predicted match likelihoods across different pages. The first part is conducted to show support for the existence of multiple PR types within in the dataset. The second part is carried out to observe traces of the CSO (i.e., Compassion International) using an optimal strategy according to what the economic framework predicts in Section 2. This section describes the methods to conduct the two parts aforementioned, preceded by a description of the datasets used in conjunction with them.

Before proceeding with the discussion of datasets utilized, I state the assumptions maintained throughout the empirical analyses:

1. In the population, there are more than one types of PRs possible, yielding different levels of utility for the PSs;
2. In the population, there are more than one types of PSs possible, some of whom are more willing to accept PRs who yield lesser utility.

Notice that the above Assumptions 1 and 2 are what enables non-trivial optimization in the sense of the model in Section 2. However, the said assumptions do not translate to assuming that there are multiple types of PRs in the data, nor to assuming that multiple types of PSs interacted with the data. Also, it does not mean that the organization actually believes multiple types of PSs are possible (i.e., $\mu \in (0, 1)$ in the economic framework). While describing methods for the empirical framework, I will also remark on the identifiability of such claims using the said methods.

3.1 Data

Baseline and additional datasets. — For accomplishing the tasks mentioned above, this study has collected individual-level data of 9,518 children from the official website of Compassion International.[3] Among many organizations, I chose Compassion International (henceforth Compassion) as its website provides more individual-level information about the PR. For the types of information presented by other institutions, I again refer to Table 6 in the Appendix. Another practical reason was that Compassion's website had a relatively full list of children available for sponsorship. In contrast, other websites (e.g., that of Save the Children) only showed a few sample individuals through randomization.

I administered data collection daily from April 19, 2019, to May 3, 2019, using web scraping through Python's Selenium module. Each round of collection was made at around 9:00 PM Central Standard Time (CST) for approximately 3 hours, except for the sessions in May due to network-related problems where the processes started at 11:00 PM CST. The first data collection was to curate the characteristics of the PR in

[3] For a more recent list of children waiting for sponsorship, I refer to the following web page: https://www.compassion.com/sponsor_a_child.

the study's sample, which could influence the sponsorship decisions of the PS. Fourteen additional follow-ups were made to check the match/non-match status of each PR. The dataset is right-censored as it has a set window of observation, and the match/non-match statuses of the PR are not visible after the said window.

In addition to the above-mentioned dataset (henceforth the "baseline dataset") that is used for constructing models to assess factors that impact matching (methods described in Subsection 3.2), I collected additional datasets to confirm evidence of across-page variability (methods described in Subsection 3.3). To be referred to as simply "additional datasets," they were collected daily on July 15, 2020 and from July 19, 2020 to July 26, 2020. Gathering these datasets actually took a longer time than collecting the baseline dataset, in which each collection process (starting at approximately 9:00 PM CST) took about 16 hours.[4]

Table 7 of the Appendix provides the descriptive statistics of most of the available covariates. I separate the table by whether a PR associated with the data was matched within the 14-day window of observation or not. A rudimentary analysis with differences in the means and the corresponding p-values from t-tests (for continuous variables) and χ^2 tests of independence (for categorical variables), while not entirely conclusive, shows that there is heterogeneity between the two groups separated by match status.

Furthermore, Table 8 of the Appendix provides the descriptive statistics comparing the baseline and additional datasets. As seen from the t-test and χ^2 test of independence results, there is much heterogeneity between the two types of datasets. Judging from this information, one may consider it an unjustified extrapolation to apply the models built from the baseline dataset to the additional datasets. However, I note that a PS entering the online list of PRs is not likely to have the overall (descriptive) information about the entire list of PRs, as there are thousands of PRs to be matched within the inventory. Further, Compassion International does not provide summary statistics about the children's information. Based on these grounds, I impose the "no-learning assumption," which can be described as follows: given a PS and a PR, the PS yields unvarying utility from being matched with the corresponding PR regardless of the change in the composition of the inventory in which the PR is presented. A piece of evidence for the no-learning assumption is found when looking at the "badge variables," to be described below.

A closer look at the "badge" variables. — Among the covariates, those located in panel D of Tables 7 and 8 require more attention. They are "urgency" (indicating a PR has waited 180 days or more without a sponsor), "urgency×wait days" (0 if waiting less than 180 days, days of waiting without a sponsor if more than 180 days), "AIDS-affected," and "vulnerable to exploitation." For simplicity, I will refer to latter two as "AIDS-area" and "exploitation-area," respectively, and collectively call the four variables as "badge variables." These variables are the more-prominently displayed information, in which the official Compassion website highlights using "badges" or visual indicators right next to the photographs of the PR if applicable. Along with primary information such as the name, birth date, and picture of a PR, these variables are what a PS would first respond to even before clicking on a separate, individual page to acquire information about other PR attributes. I also note that the days of waiting for a sponsor are displayed *only* for the children with urgency badges.[5]

[4] It is suspected that the Compassion International website forced a longer loading time, in order to deter web-scraping of individual PR information. During the additional data collection stage, the website also regularly changed the URL of individual PRs, perhaps to disable tracking of information. Due to this unexpected behavior, the study was not able to collect useful data prior to July 15, 2020 and from July 16, 2020 to July 18, 2020.

[5] However, I also note that it is possible to discover an approximate order of days without sponsors for the PR without

Despite their importance, there are problems associated with the badge variables, especially AIDS-area and exploitation-area. While these variables seem to reflect the individual-specific vulnerabilities to AIDS and exploitation, they are actually geocoded (hence the re-naming of the variables to end with "area"). A more in-depth look at the dataset shows that the Compassion website has coded all PR living in Africa and *only* them as those in AIDS-affected areas. It is true that, on a continent-level inspection, those in Africa are more susceptible to HIV/AIDS. However, this is false on a country- or region-level investigation. For instance, the west-African nation of Burkina Faso has a lower HIV/AIDS prevalence than Haiti (CIA n.d.). Both are nations where portions of the overall sample PR reside in, yet the latter country (being in the Caribbean) is not marked as AIDS-area whereas the former is.

With respect to exploitation-area, the official website does not give any guideline of what constitutes vulnerability to exploitation and what type of exploitation (e.g., labor, sexual) the badge is referring to. Due to these issues, any PS may have misinterpreted the information shown by this badge as well. As previously mentioned, the said variable is geocoded as well; in fact, a number of countries are marked as vulnerable to exploitation and only them. While it is understandable that there would be additional costs incurred to collect individual- or finer-regional-level information about exposure to exploitation, these issues seem to be more related with careless data management. I note that the lists of countries marked as AIDS-area or exploitation-area are provided in Table 11 of the Appendix.

"Page" and its relation to inventory positions. — For both the baseline and additional datasets, I collected individual PR information that is organized on different but consecutive webpages (listed from 1 to 139) in which each page holds 72 PRs. Throughout the empirical part of the paper, I will consider the said pages as inventory positions described in Section 2. The baseline dataset pages are ordered in decreasing days of waiting, whereas those for the additional datasets are not ordered in any particular manner (i.e., the default page ordering by Compassion International).

The problem is that, as opposed to Section 2 where there was one PR per inventory position, now there are 72 PRs. This alteration may make the theoretical predictions based on Section 2 to be not as accurate. With respect to this potential incompatibility, I suggest the use of page-level average PR types while acknowledging the difference between theory and reality. By using this specification, it would still be possible to extend the theory results as the proposed goal of the CSO remains the same: to maximize sponsorship and minimize the costs incurred to care for unmatched PRs.

3.2 Identification of Factors Significant to Match Results

A. Identification and Limitation

This study employs two classes of estimation strategies for this subsection – (semi-)parametric and nonparametric – to identify which variables in the base dataset as mentioned above contribute significantly to match results. While there is no absolute boundary that separates the two classes of strategies, the former

urgency badges. The said order be can found by setting the display option to "longest waiting" and taking note of the pages they are shown on.

allows an econometrician to understand the sign and degree to which a variable affects match status. On the other hand, the latter produces less explicable results while having better prediction accuracy (and other prediction-related metrics) for match statuses, as to be confirmed in Section 4. However, the said difficulty in interpretation when using nonparametric methods does not mean that there are no metrics to gauge which variable is more profound; in specific, I will use either feature importance based on impurity or permutation (e.g., see Han et al. 2016).

As previously mentioned at the outset of this section, the purpose of confirming factors significant to match results is to see whether there are multiple PR types in the dataset. The methods used in this part assess how much a change in one variable affects or is associated with the degree of change in the dependent variable (e.g., probability of matching). However, since this study's datasets are confined to the information from a single organization (i.e., Compassion International), it would be difficult to extrapolate the estimation results to apply to the entire CSP market. Further, if one applies the logic of nested logit choice (as in Goldberg 1995, for instance), then all effects estimated using the Compassion International data will be relevant within the said organization's range of child sponsorship and not across those of different organizations.

Despite this limitation, the methods to be described below can be crucial as they suggest the possibility of multiple PR types, which in turn leads to the idea that the CSO may consider non-trivial optimization strategies. The commonality in the models to be discussed below is that they all examine how the variability in feature data is related to match statuses or probabilities through measures like statistical significance (for parametric methods) and feature importance (for nonparametric methods). Assuming there were multiple PS types interacting with the data, I provide an illustration with the parametric approach *while controlling for positions in the inventory* (i.e., the pages) of how the said approach may or may not buttress the possibility of multiple PR types. Suppose that, out of the independent variables used, none of them yielded statistical significance. Then, this result would suggest that despite the variability in data, they were ineffective in influencing the PSs whose decisions are as good as random. According to the economic framework of Section 2, as-if-random choices of PR is only available when there is only a single type of PRs; this is because if there were both t_l- and t_h-types in the data, the those classified as t_h-types would be more likely to yield matches by being wanted both by L- and H-type PSs. Analogous statements can be made with nonparametric methods.

The above example as well as the methods elaborated below assumes that there is PS type variability within the data. This assumption is provided as there is no explicit information about the PSs who interacted with the dataset, and therefore identifying their respective types is not possible. If the said assumption of multiple PS types is removed, statistical significance of certain variables can be due to only H-type PSs entering the inventory and picking only the t_h-type PRs.[6] One possibility ruled out by the exposition of statistical significance is that there are only L-type PSs interacting with the data, as the theory assumes that they are willing to match with any PR (and therefore would be choosing as if random).

Beyond the inability to identify information about the composition of PSs, the methods in this subsection

[6]Nonetheless, if any statistical significance is shown, it would suggest that there is variability in PR types regardless of variability in PS types.

have another limitation: while they are able to provide evidence for (or against) multiple PR types in the data, they are not able to particularize the *number* of PR types in the data. As mentioned in Section 2, the assumption that there are two types each of PRs and PSs is a simplification to capture the essence of CSP activity. In fact, it is very likely that there are multiple types of PRs perceived by the PSs, in which the most extreme case is that a PS recognizes each individual PR as a separate type. While this limitation is worthy of recognition, the point of confirming that there are multiple PR types is to check whether the CSO has any motivation to design a non-trivial, optimal solution. Therefore, the various approaches introduced in this subsection will still have their merits.

Before proceeding with the description of parametric and non-parametric approaches, I remark that the below approaches should be distinguished with simply checking for the variability in data of independent variables or using the said variability to conduct operations like clustering. The key difference is that the approaches to be described below try to get at how the variability is perceived by the PSs and is reflected on the dependent variable having to do with match statuses. Checking simply for measures like variance for continuous variables or categories for discrete variables are important for data inspection, but whether PSs will interpret such as different types of PRs or levels of utility towards them is unclear just by doing so.

B. Parametric Approaches

Three types of (semi-)parametric approaches will be utilized, which are (1) survival analysis (mainly Cox regression), (2) fuzzy regression discontinuity design (RDD), and finally (3) random coefficients logit model. The former two methods utilize data as is, with observations at the individual PR level. On the other hand, the third method uses data that is aggregated at the group level. The said groups are made by classifying according to the country of origin, specific Compassion regional center that a PR belongs to, or assigning group labels by clustering methods.[7]

Survival analyses. — Survival analyses with Cox proportional hazards model (i.e., Cox regression) will be the primary (semi-)parametric strategy that this study employs to understand the relationship between successful matches and the displayed PR characteristics. The critical object of interest in a survival analysis approach is the conditional hazard; it is the probability that an observation is removed (or reaches "death") from the list in the directly subsequent period given that the said observation is on the list (or "alive") while also accounting for the covariates (Ciuca and Matei 2010). Therefore, conditional hazard rates directly correspond to how long it takes for a match to be made (or not made during the survey window of observation) in the CSP environment.

Among a number of different types of survival analysis methods, I follow Cox (1972) to apply the semi-parametric Cox regression. The difference between purely parametric versions of survival analysis versus the Cox regression is that the former versions assume a predefined functional form for the baseline hazard, while the latter estimates the said baseline hazard using the data (Rodriguez 2010). Many of the functional forms used in parametric survival analyses, including Weibull and log-logistic functions, are monotonic in survival time. However, the assumption of monotonic hazard rate seemed unfitting for the purposes of this

[7]I note that while the said groups are an attempt to approximate the different PR types, they should not be understood as equivalent to the PR types.

study for the following reason. If monotonic hazard functions are used, as it would mean that the likelihood of matching either increases or decreases as waiting days (without a sponsor) becoming longer; however, in the data, it is observable that there is a very long tail to the right in terms of waiting days while longer days increases the chance of meeting a sponsor. Therefore, a non-monotonic functional form for baseline hazard may be more suitable, most likely a positive and concave one (so that chances of matching increases to a certain point in observation time but starts to decrease again). Since it is not possible to specify exactly where the maximum hazard rate occurs in the proposed concave function, I instead use the data to fit the baseline hazard's functional form.

As a complement to the Cox regression model, I also utilize a simple logistic regression in which the dependent variable is the binary (= 1 if matched during the window of observation). Comparing the logistic and Cox regression results will help one better understand any potential non-linearities of independent variables affecting the match status across time.

To further elaborate on Cox regression, I follow Rodriguez (2010) to write the continuous-time model of general survival regression as follows:

$$\lambda_i(t|X_i) = \lambda_0(t)\ \exp(X_i'\beta)$$

where $\lambda_0(t)$ is the baseline hazard function at time t, X_i' is the vector of covariates for individual i, and $\lambda_i(t|X_i)$ is the conditional hazard function at time t given the said covariates. Because the study does not observe the data in continuous time, one may re-write the above model as follows:

$$\lambda_{ij}(X_i) = \lambda_j\ \exp(X_i'\beta)$$

where j denotes the time interval instead of continuous time, λ_j the baseline hazard function at time interval j, and $\lambda_{ij}(X_i')$ the conditional hazard function at time interval j given the covariates X_i of individual i.

For the covariates (independent variables), this study will primarily consider the treatment variables as the three badge variables (i.e., urgency, AIDS area, and expropriation) and urgency$\times$wait days with the additional variables as age, biological sex, dummy variables for living with father and with mother, continent dummies, parent employment statuses, schooling status of the PR, and page dummies. Unless described otherwise, the page variable used in the parametric methods will refer to the specific page a PR is displayed on when shown in the order of the longest time waiting. Regarding the previously-mentioned additional variables, one may also interpret some (including guardian employment statuses and age of the PR) as treatments that influence sponsorship matching as well.

Before proceeding with the description of other parametric and nonparametric methods for this subsection, I note the following about the standard errors calculated, multiple hypotheses testing, as well as potential worry for endogeneity problems. Firstly, I assume that there may be structural unobservables at the page level that influence sponsorship matches. For instance, it may be that the PS' attention is restricted to a per-page level, and they may only consider the PR on a specific page to be in their set of alternatives. Per such considerations, the study will use clustered standard errors (SE) at the page level.

Moreover, note that the study regards the PS to be observing multiple PR characteristics at once, and such information is contemporaneously affecting the PS' decision-making. Following this viewpoint, the survival regression strategy described above tests several covariates together for their statistical significance in determining matches. Therefore, I consider multiple hypotheses testing and the corresponding correction for SE. A relatively simple way of implementing the said procedure is to use the Bonferroni correction, which is to test individual hypotheses at the significance level of α/K when the target significance is at α, and there are K hypotheses to test (Dunn 1961). As specifications may vary across regressions, I set $K = 50$ and $\alpha = 0.05$ so that one may regard a coefficient statistically significant if the corresponding p-value turns out to be lesser than 0.001.[8] While this can be a harsh standard, it may be a necessary one to understand how different types of information affect PSs' decision-making and, by extension, to verify the existence of multiple PR types.

Finally, the study acknowledges that there are variables not included in the data collection process that could be influential to PSs' decisions. Had this study been about, for instance, calculating elasticities of match demand with respect to the variables included in the statistical analyses, such excluded or unobservable variables that affect the estimation may contribute to inaccurate projections of such statistics. And in such cases, measures to account for the potential endogeneity problem (such as difference-in-differences or instrumental variables approaches) may be essential. However, because this study focuses on trying to detect whether there are any significant variables that affect match statuses (and this is done to hopefully give grounds to assuming multiple PR types), I will acknowledge the said potential issues but will not take additional measures to correct for them.

Regression discontinuity design. — The fuzzy RDD conducted in this study specifically has to do with the dual nature of urgency variable. While it informs the PS about how dire a PR's situation is, it is also an indirect way to show how long the said PR has been on the CSP database waiting for sponsors. This latter part poses a problem in interpreting the sign or statistical significance of the coefficient estimated from survival analysis, as it may highlight the extended exposure to being on the list (and have higher chances of yielding matches) rather than a PR's exigency for help. In light of this, the study employs the RDD approach to disentangling the components of the said dual nature. Specifically, it looks at the conditional average treatment effect (CATE) of the urgency variable on match probability at the threshold of it turning on; by doing so, it is possible to estimate the urgency badge's impact on sponsor-recipient matches with less concern for the impact of extended exposure.

As previously explained, the said threshold of urgency badge turning on is precisely 180 days after being on the list, and the days of waiting are unobservable to the PS when 180 days have not passed. This inability to pinpoint the days of waiting for the majority of observations prevents the use of sharp RDD.[9] However, the dataset has page information that serves as a proxy for days without sponsorship. As seen from Figure 12 in the Appendix, there is a sharp, monotonic increase of percent showing the urgency badge between the pages three and four; this motivates the use of fuzzy RDD. The probability of being "treated" by the

[8]The high number of K is to account for not only the variables mentioned above but also the page-level or country-level fixed effects to be used.

[9]I note that I also considered the method of assuming a parametric distribution for the days of waiting and fitting the known data points using maximum likelihood estimation. However, because the days of waiting are observed only for the first three pages (approximately only 2% out of the total 140 pages), I deemed that this would be highly inaccurate.

urgency badge does not jump from 0 to 1 (as in the sharp RDD setting), but there still exists a sizeable, discontinuous change in the said probability of treatment at the threshold (i.e., page 3).

I follow Hahn et al. (2001) and use the instrumental-variables version of fuzzy RDD, which can be written as follows:

$$Y_i = \alpha + \gamma \text{urgency}_i + \beta(\text{page}_i - 3) + \delta Z_i(\text{page}_i - 3) + \varepsilon_i$$

where, for individual i, Y_i is the match status (= 1 if matched during the window of observation), urgency_i is the urgency badge variable (= 1 if the said badge is shown), and page_i is the page i was displayed on. urgency_i is instrumented by $Z_i \equiv 1\{\text{page}_i \leq 3\}$. The parameters α, γ, β, and δ are estimated by local linear regression as follows:

$$(\hat{\alpha}, \hat{\tau}, \hat{\beta}, \hat{\delta}) = \underset{\alpha,\tau,\beta,\delta}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau \text{urgency}_i - \delta Z_i(\text{page}_i - 3)\right)^2 K_h(\text{page}_i - 3)$$

where $K_h(\cdot)$ is the kernel function with bandwidth h. This paper uses triangular kernel with optimal bandwidth h selected by the method of Calonico et al. (2014). For sensitivity analysis of the threshold being page 3, I will also conduct regressions with similar designs using page 2 and page 4 as thresholds.

Note that because the fuzzy RDD described above utilizes IV, it must satisfy relevance and exclusion restriction. The relevance condition follows from the existence of a threshold for being "treated" with a urgency badge. On the other hand, showing exclusion restriction can be dubious as it may be the case where the page variable itself affects match status. For example, the PS might be negligent of looking at the latter pages in comparison to the earlier ones due to limited time and attention. For example, the PS might be negligent of looking at the pages near the end in comparison to the earlier ones due to limited time and attention. However, because the problem at hand is to estimate the CATE of the urgency badge at the threshold, I argue that the page variable would not locally affect the decision-making of a PS. As an illustration, consider a PS browsing a set of PR for sponsorship decisions on page n. While it would be difficult to argue that the said PS would view each page with equal carefulness, it is not unrealistic to assume that he or she may inspect pages $n-1$ to $n+1$ with similar amounts of attention. That is, there is no local direct impact from page numbers to match statuses. This relaxation helps one to apply the exclusion restriction condition.

Random coefficients logit. — As opposed to the models described above, the random coefficients logit considers potential variability in coefficients by individual PSs or by types of PSs. Using the developments of the seminal paper by Berry, Levinsohn, and Pakes (1995) and the overview of the said paper by Conlon and Gortmaker (2020), the model can described to fit the CSP setting of this paper as follows. Let the indirect utility of an individual PS i choosing to match with a PR from group j (to be further explained below) as follows:

$$U_{ij} = \delta_j + \mu_{ij} + \epsilon_{ij} \tag{26}$$

in which $j \in \{1, 2, \cdots, J\}$ (i.e., there are J different groups to consider). Further, let $j = 0$ denote the outside option of not matching with any of the PR groups in $\{1, 2, \cdots, J\}$. Note that δ_j is the component of the indirect utility that does not vary across PSs; on the other hand, μ_{ij} represents the component that does. ϵ_{ij} refers to the error term specific to i and j. Because the indirect utility in (26) assumes a linear form, I will dub this as the "(demand-side) linear equation."

Further, denoting d_{ij} as the binary decision of i selecting ($d_{ij} = 1$) or not selecting ($d_{ij} = 0$) the PR from group j, d_{ij} can be written out as:

$$d_{ij} = \mathbb{1}(\forall \hat{j} \in \{0, 1, \cdots, J\} \setminus \{j\},\ \ U_{ij} > U_{i\hat{j}})$$

for any $j \in \{0, 1, 2, \cdots, J\}$ (where it is assumed that $U_{i0} = 0$ for all i) and $\mathbb{1}(\cdot)$ is the indicator function.[10]

Note that d_{ij} is a function of $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{iJ})'$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_J)'$. Then, denoting the share of group-j PRs being matched within the window of observation and over all PRs in the CSP as s_j, the said share can be written as

$$s_j = \int d_{ij}(\boldsymbol{\delta}, \boldsymbol{\mu}_i)\ d\boldsymbol{\mu}_i\ d\boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{iJ})'$. Conlon and Gortmaker (2020) further suggest that ϵ_{ij} be assumed i.i.d. with type I extreme value distribution for identification (p.7). If this assumption is placed, s_j can be written as a function of $\boldsymbol{\delta}$ and the parameter $\tilde{\theta}$ used for the distribution $f(\boldsymbol{\mu}_i|\tilde{\theta})$:

$$s_j(\boldsymbol{\delta}, \tilde{\theta}) = \int \frac{\exp \delta_j + \mu_{ij}}{\sum_{\hat{j}=1}^{J} \exp \delta_{\hat{j}} + \mu_{i\hat{j}}} f(\boldsymbol{\mu}_i|\tilde{\theta})\ d\boldsymbol{\mu}_i. \tag{27}$$

The model in (27) is referred to as random coefficients logit, where the "randomness" comes from the fact that individual specificity (i.e., $\boldsymbol{\mu}_i$) is randomized over the mixing distribution $f(\boldsymbol{\mu}_i|\hat{\theta})$ (Conlon and Gortmaker 2020, p. 7).

As opposed to the previous two parametric approaches discussed, various independent variables seem to be missing from the equation (27). However, they are actually used in determining the values of δ_j for various j, the individual-invariant component to U_{ij}. Let σ_j denote how much, in percentage, the group-j PRs occupy the inventory (I refer to this as "occupancy"). Also, let the vectors of other observable (average) characteristics of group-j PRs be denoted $\boldsymbol{x}_j$. Then, with ζ_j and ξ_j respectively denoting the structural unobservables and the error term, one may write δ_j as follows:

$$\delta_j = \boldsymbol{x}_j'\beta + \alpha\sigma_j + \zeta_j + \xi_j \tag{28}$$

where $\delta_j = \delta_j(\boldsymbol{S}, \tilde{\theta})$ by solving the system of equations using (27) (for all $j \in \{1, 2, \cdots, J\}$) in terms of δ_j, and $\boldsymbol{S} = (s_1, s_2, \cdots, s_J)'$.

One key difference between the random coefficients logit model of Berry, Levinsohn, and Pakes (1995) and that in equations (27) and (28) is that the model of this paper does not include a price term. This is due

[10] Notice that the assumption of $U_{i0} = 0$ for all i may seem as if it is against the assumption of differing outside option by PS types. However, since utility from sponsorship is ordinal and the expression (26) allows it to vary by individual PS, one may think of U_{ij} as the difference in utility between matching with a PR from group j and the outside option.

prices being homogeneous for all PRs within the same CSP, and I only have data from one CSO. Instead, I include the term σ_j to understand how occupancy of a certain group of PRs in the inventory may affect the utility of matching with a group-j PR. Further, in this paper's exercise, I do not specify the "supply-side" equations (e.g., that of a CSO determining σ_j due to having monopoly over information of PRs in the said organization's program). The main reason is that all data used in this paper comes from Compassion International, and supply-side equations can be useful only if one observes variability in different CSOs (or firms, more generally).

The parameters in (28) can be estimated using GMM. Specifically, I use the Python module PyBLP described in Conlon and Gortmaker (2020). The said module also deals with the creation of "optimal instruments" for the unobservables in $\boldsymbol{v}_j$. In reporting the results, I state the estimates and standard errors for β and α in (28), which can be understood as the across-group average contribution towards (linear) indirect utility of a PS. I refer to Conlon and Gortmaker (2020) for further details about the algorithms used in estimation.

With respect to the groups $j \in \{1, 2, \cdots, J\}$, I have previously stated that the said groups are not to be considered equivalent to the PR types described in Section 2. In contrast to the settings of papers like Berry, Levinsohn, and Pakes (1995) where the products or groups are easily distinguishable or predefined, that of CSP does not have clear boundaries between PRs other than those by countries or Compassion centers the PRs belong to. Further, for sound statistical estimation, it is ideal to have a large number such groups; therefore, options like binary grouping (e.g., those exposed to exploitation versus those not) would be difficult to use. In light of this, this study will use countries and Compassion centers as standards for creating groups due to their large variability. Also, as mentioned before, I use k-means clustering to produce additional group labels based on the data.

C. Nonparametric Approaches

In place of finding features important to the match status via statistical significance, one may alternatively try to identify features that help to correctly predicting the match status (i.e., contribution to prediction accuracy). In light of this, I suggest nonparametric approaches such as decision tree, random forest, and discriminant analysis, which can be used in classifying PR match statuses. Due to the lack of balance in the numbers of PRs yielding matches after certain days, I consider the simpler problem of classifying whether a PR yields a match or not during the window of observation (instead of trying to also predict on what day the PRs yield matches).

Especially with decision trees and random forests, Kuhn (2014) and Bou-Hamad et al. (2011) show that such methods are natural extensions to survival analyses including Cox regression as described earlier. Specifically, Kuhn (2014) suggests the use of information gain as a splitting criterion to achieve similarity to the survival analysis methodology. While it is not mentioned that other nonparametric methods are also directly related to survival analyses, the said methods can be compared with random forest and decision tree via metrics like prediction accuracy, recall, and precision.

By using minimal mean squared error (MSE) as the criterion for choosing the best model, I will race different methodologies listed in Table 12 together with hyperparameter tuning (HPT) criteria. The features used in these methods are summarized in Table 13. After acquiring the best model in terms of prediction accuracy, I will calculate feature importances of each variable used in the feature set. If the best model turns out to be either decision tree or random forest, then impurity-based feature importances will be presented; if not, permutation feature importances will be reported.

3.3 Identification of Potential Optimization by the CSO

A. Overview and Limitation

Given that the methods in the previous subsection are able to detect variables significant to match results, variability of such variables across the different pages will allude to not only multiple PR types but also how the CSO arranges the inventory. I note once again that, using the data of this study, it is impossible to firmly state about how the CSO perceives different types of PRs (or how many types there are). However, by seeing page-to-page variations in the aforementioned significant variables as well as the said variations' degrees, it would be possible to get the type of strategy being employed by the CSO.

More precisely, this study seeks to assess whether there is any reason to be suspicious about the CSO's strategy going against the results in Corollary 2.1. The null hypothesis that the study will test is not that "the CSO is optimizing" or its negation. Rather, it would be the following: "given assumptions about parameters like the maximum search amount, there is reason to believe that CSO is adhering to the results of Corollary 2.1"; I will denote this hypothesis as H_0^*. To slightly paraphrase, Corollary 2.1 states that if $\mu \in (0, 1)$ is close enough to 1 (i.e., such that $\frac{1+c_h}{1+c_l} > \hat{\alpha}(k, N)$), then the optimal strategy has t_l-types up to the bth position and has a t_h-type at the $(b+1)$th position. Therefore, the said corollary does not absolutely produce a closed-form solution but rather provides one of necessary conditions for optimal strategies. While a less strict null hypothesis, it would still be challenging to test due to at least two reasons: (1) the value of b is unknown to an outsider and (2) there are multiple positions within one page, as mentioned in Subsection 3.1 (maximum of 72 children displayable). However, the issue of not knowing the exact value of b can be overcome by considering different values of the said parameter. Additionally, if one interprets b as the maximum average number of pages that a PS will view, one can calculate the average characteristics within pages and compare statistics page-to-page, as suggested above.

The page-by-page comparisons will be conducted as follows. As previously argued, the only variations that are important are the ones in variables that seem to significantly impact the match results. Therefore, statistics (to be further described below) used in this subsection will be produced only from the said significant variables. Once the statistics are garnered, I will compare those from a single page with those from a neighboring page. By using appropriate statistical tests, it will be possible to gather p-values from the page-by-page comparisons (with the null hypothesis being that the two neighboring pages' statistics are the same on average). By assuming different cutoffs for p-values, one may produce hypothetical statements about the said two neighboring pages containing similar types or not. These results will then be compared with that of Corollary 2.1 to qualitatively evaluate H_0^*. I further remark that I will present results of up to the 15th page (approximately 11% of the entire additional data), assuming that the PS would search a small number of pages on average and provided that Corollary 2.1 is useful in designing the front part of the inventory rather than the back.

B. Statistics for Page-by-page Comparisons

There will be three types of statistics used in page-by-page comparisons. The first type is using the aforementioned significant variables as they are. In this case, the continuous variables can be tested using the t-test and the categorical variables can be tested using the χ^2 test of independence. However, because the types are likely to be co-determined by the variables, it may be necessary to compare them together in a multivariate sense. The difficulty in doing so is that there is a mixture of categorical and continuous variables, where it is impossible to apply the normality assumption to the former type of variables and tests like MANOVA or Hotelling's T^2 test (which rely on the said assumption) cannot be properly used (Johnson and Wichern 2007).

Therefore, the two additional types of statistics are suggested to overcome said difficulties. One is factor scores from exploratory factor analysis (EFA) using the features mentioned above. Specifically, I calculate the heterogeneous correlation matrix and use this matrix as an input to EFA. Here, the heterogeneous correlation matrix refers to one that records Pearson correlation between continuous variables, tetrachoric correlation between binary variables, and other polychoric correlation for all other relationships. Once the EFA is conducted, data can be used to generate factor scores, which is a way of dimenality reduction with the end-products being continuous variables. The said scores are then compared across pages using MANOVA to test for the equivalence of mean vector of factor scores.

An additional statistic is to use the best (in the sense of greatest prediction accuracy) nonparametric model described in Subsection 3.2 and use this to generate predicted match probabilities across the additional datasets. Once again, the aforementioned no-learning assumption is in effect while justifying for the use of this approach. Those with higher match probabilities are likely to be higher-utility yielding, but one would need to conduct appropriate tests to corroborate the idea that predicted match probabilities are indeed different with statistical significance. In this aspect, I will compare the average predicted match probabilities across different pages using ANOVA.

4 Estimation Results

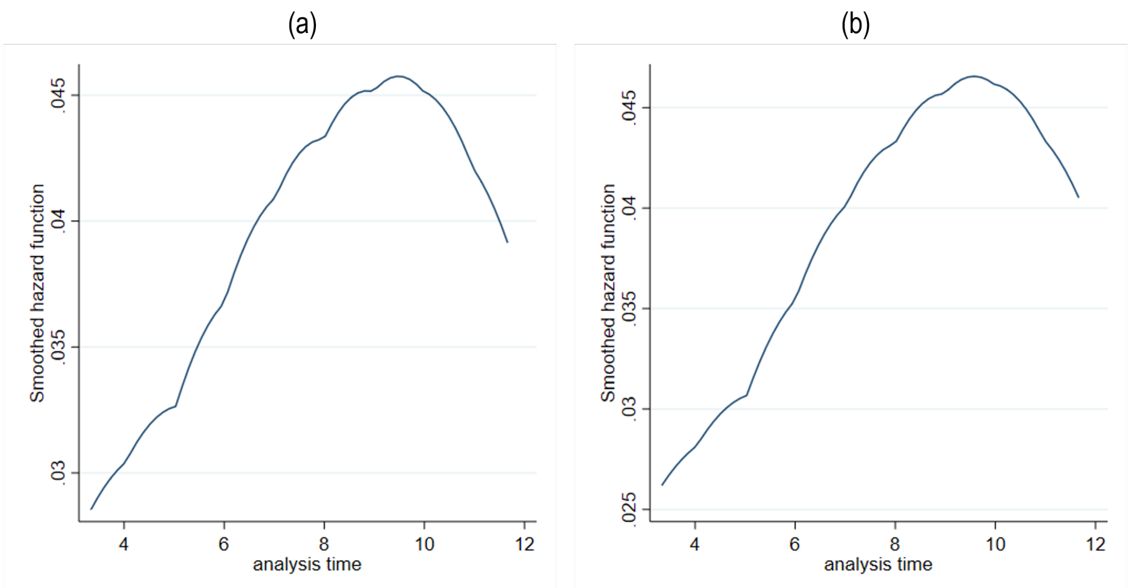
4.1 Results from Identifying Factors Significant to Match Results

A. Results from Parametric Methods

Survival analyses with comparisons to logistic regression. — Presented in Table 9 of the Appendix are the hazard rate estimates (with page-clustered SEs) from the Cox regressions. Presented together with this information are the results from logistic regressions. In both types of regressions, I conducted two variants of model specification in which one incorporates the variables mentioned in Subsection 3.2 as well as page-level fixed effects, while the other includes country-level fixed effects and removes country-specific variables (e.g., Christianity %) as well as continent dummies.

While the results show statistical significance for many of the variables in each model estimation regardless of using Bonferroni corrections and align with one another in general, they do have some discrepancies. Across different regression methods, an example is with the variable urgency, in which it is noted as negatively impacting matches (in terms of hazard rates) for the Cox regressions while being noted as positively impacting for the logistic regressions. Within the same regression method, there are also variables like urgency$\times$wait days (for the Cox regressions) whose coefficients are significant (at $\alpha = 0.001$) in one specification and not in the other. Fortunately, within the same class of model choice, there are no variables whose coefficients have different signs and have statistical significance (at $\alpha = 0.001$) at the same time.

Figure 5: Baseline Hazard Functions for the Cox Regressions

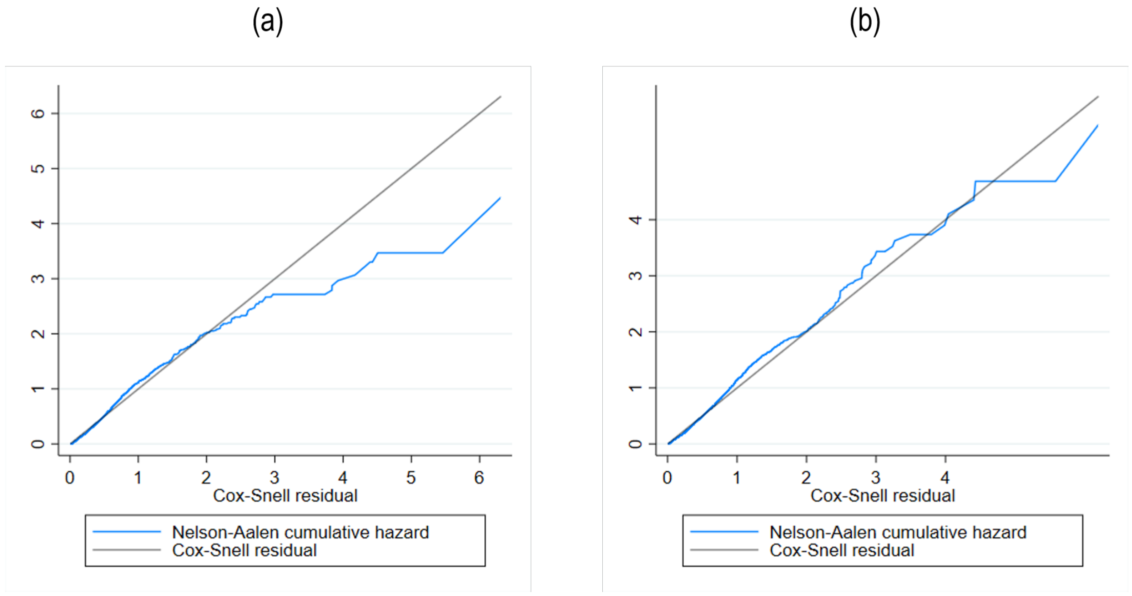


NOTE: Panels (a) and (b) each refer to smoothed hazard rates Cox regressions as corresponding to columns (1) and (2) of Table 9.

These observations suggest that there is not only nonlinear relationship between the dependent and independent variables, but also that among the independent variables as well. While it is difficult to say for certain from where the disagreement between Cox and logistic regression results originates from, this could be due to the nonlinear shape of the baseline hazard function. While logistic regression looks at data before and after the window of observation, survival regression takes into account for many of the within-window dynamics with respect to the dependent variable. Therefore, while a certain variable may seem to promote matching when considering the net effect during the window of observation, it may be hindering matching at certain ranges of time within the window. As a piece of evidence, I present the baseline hazard functions of the regressions corresponding to columns (1) and (2) Table 9 in Figure 5.

The disagreement between the Cox regression models seems to originate from the aforementioned nonlinear relationship of the covariates. Specifically, as the two regressions in columns (1) and (2) of Table 9 differ only by the addition of country-level fixed effects, it is likely that there are across-countries difference in tendencies of variables interacting with one another. While this is not to say that using continent dummy variables are invalid (as many are statistically significant in column (1) of Table 9 at $\alpha = 0.001$), such dummies are finer and therefore more suitable if the relationships between independent variables vary even within the same continent designation.

Figure 6: Comparison of Cox Regression Goodnesses of Fit



NOTE: Panels (a) and (b) each refer to goodness-of-fit graphs for Cox regressions as corresponding to columns (1) and (2) of Table 9.

Accounting both for the across-countries variations in covariate relationships and across-time non-monotonicity in hazard rate (or match probability, in extension), the Cox regression in column (2) of Table 9 seems to be the most reasonable choice of modelling among the four specifications explored. I also provide a comparison of goodnesses of fit for the survival regressions in the columns (1) and (2) of Table 9 in Figure 6. The said goodnesses of fit can be assessed by plotting the cumulative hazard function against Cox-Snell residuals, and seeing whether the former agrees with the latter approximately (Xue and Schifano 2017). As visible from Figure 6, the model with country-level fixed effects has a hazard-to-residuals graph closer to the 45-degree line (representing the residuals-to-residuals graph), meaning that the overall goodness of fit is higher for the said model.

Based on the Cox regression model with country-level fixed effects, then, it can be said that urgency, urgency$\times$wait days, biological sex, and geographical variables (e.g., country dummy variables; if unavailable, at least continent dummies with the exploitation-area variable) are significant in determining the match statuses of PRs. This result, while not entirely conclusive, gives weight to the idea that some PRs are more sought after than others (i.e., multiple PR types).

Fuzzy RDD. — As elaborated in Section 3, estimations using fuzzy RDD around page 3 (and around pages

2 and 4 for sensitivity) are conducted to address concerns regarding the dual nature of the urgency badge. Table 2 shows the results.

Table 2: Results for Fuzzy Regression Discontinuity Design (Using IV)

	Threshold `page`		
	`page` 3	`page` 2	`page` 4
Panel A. Structural Estimates			
Urgency	0.1067^{***}	-0.0194	-0.2142
	(0.0278)	(0.0469)	(0.3236)
Panel B. First-stage Estimates			
Page	-0.8179^{***}	-0.3412^{***}	-0.2190
	(0.0331)	(0.0570)	(0.1477)
Total N	9518	9518	9518
Left N	109	45	168
Right N	950	666	261
Optimal bandwidth	13.093	9.150	3.445

NOTE: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$. Heteroskedasticity-robust SE in parentheses. Out of the three threshold pages, using page 3 produces the main result; others are used for sensitivity analysis. Dependent variable is binary match status during the window of observation (i.e. = 1 if matched in the window). Left and right N refer to the numbers of observations with `page` less than or equal to the threshold and greater than the threshold used for fitting the local linear regression. Triangle kernel function was used.

I begin by examining the first-stage estimates (in panel B of Table 2), which will gauge how strong the relationship is between the variables page and urgency. The coefficients on the page variable are all shown to be negative, which means that as the page numbers increase, it is less likely to see the urgency badge. This intermediate result is correct as the pages are in the decreasing order of duration of waiting without sponsorship. I also note that both the magnitude and statistical significance of the coefficient on the page variable are the most considerable for the threshold page of 3. This observation is anticipated, as seen in Figure 12 of the Appendix.

Next, I examine the CATE of urgency on match probability at the designated thresholds. Table 2 presents the results of the structural estimates of the fuzzy RDD. Note that positive values for the coefficients mean that the urgency badge increases the chances of a sponsorship match. The study finds the only positive and the single statistically significant result when using the threshold of page 3. This finding confirms that, despite the worry for the urgency badge's dual nature as previously mentioned, the said variable positively affects match probability at the threshold where it turns on.

There may still exist lingering concerns when comparing the results of fuzzy RDD and Cox regression. As seen from column (2) of Table 9, the hazard rate estimate on urgency badge was lesser than 1 (0.282), implying negatively affecting match probability. However, when considering also the urgency×wait days variable (with the estimate of 1.007), it is seen that there is altogether a positive impact. For instance, consider comparing two individual PRs with all else equal but the urgency-related variables; let us further assume that the urgency badge has barely turned on for one of the PRs, such that wait days is equal to 180 days. Then, the PR with 180 days of waiting has a hazard rate (approximately) $3.5(\approx 0.282 \times (1.007^{180}))$ times that of the PR without the urgency badge. Therefore, there is no contradiction between the two methods regarding the urgency variable.

Table 3: Comparison of Accuracy and AUC for the Nonparametric Strategies

Method	Accuracy	AUC	Brief HPT Details
DT	0.581	**0.915**	Minimum samples for split: 9
RF	**0.607**	0.713	Minimum samples per leaf: 8
SVM	0.586	0.854	Kernel: radial basis, C: 2.194
QDA	0.575	0.614	Tolerance: 10^{-6}, reg. param.: 0.8
MLP	0.580	0.634	Hidden layer size: 17, α: 3.905

Note: In bold are the maximum values for each column (by panel). The version designation of either "whole" or "feature selected" (abbreviated to FS above) is based on the features used in estimation; the lists for the said features are given in Table 13. "AUC" refers to the area under the ROC curve.

Random coefficients logit results. — As visible from Table 10 of the Appendix, the linear parameter estimates through the random coefficients logit model are mostly very statistically significant. While this is a positive result, there are two conspicuous problems. Firstly, the said estimates' signs are inconsistent across different grouping methods. For instance, while urgency is associated with higher PS utility in columns (1), (2), (4) and (5), the opposite result is shown in column (3). This discrepancy might be due to how the grouping criteria are correlated with the variables used in this set of analyses.

Secondly, the within-grouping ratios of one estimate to another are inconsistent across the columns as well. That there is scale difference across columns is not too problematic as the said coefficients are estimates to the equation for the linear indirect utility, and the said utility is ordinal. For instance, consider the father employment variables. Despite those estimates in column (1) being in the order of at least 10^{19} and those in other columns being in the order of at most hundreds, they would be reasonable estimates as long as the ratios of coefficient on father irregular employment to that on father regular employment across different grouping methods are similar. However, the coefficient ratios are very uneven despite using the same variables across different . Again, the problem might be due to the correlation between grouping criteria and the variables.

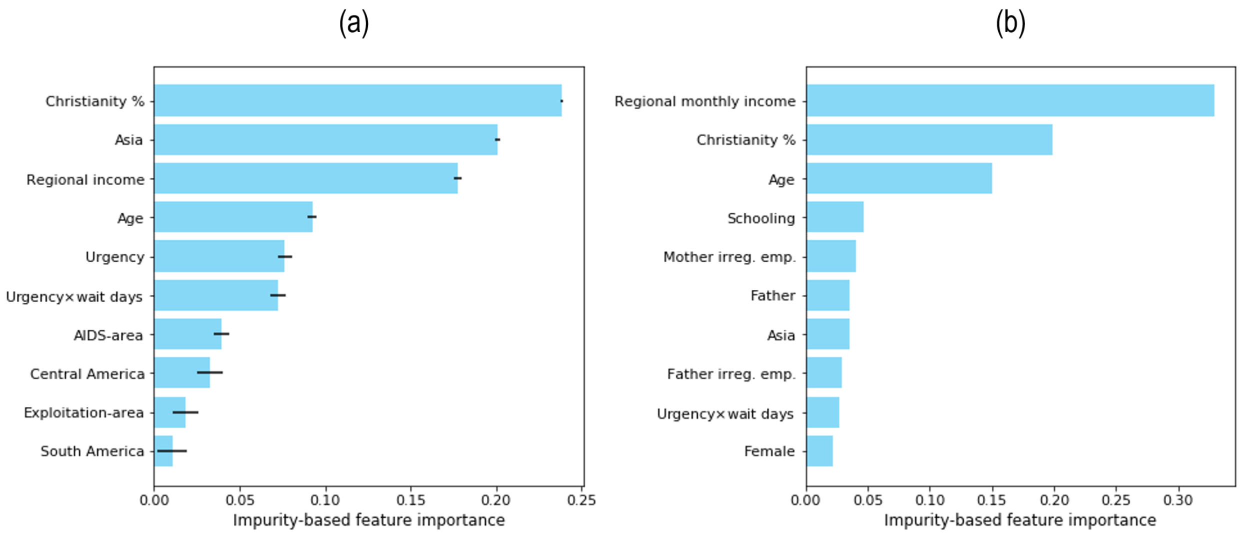
Considering the above problems, random coefficients logit results do not add much to the results garnered with Cox proportional hazards models, logistic regressions, and fuzzy RDD. Along with arbitrary grouping criteria, perhaps another source of problem is the linear form of the equation for indirect utility. This suspicion is similar to that in the results for Cox and logistic regressions, where a number of independent variables were hinted to have a nonlinear relationship with match status.

B. Results from Nonparametric Methods

The prediction accuracy and area under the ROC curve (AUC) results are shown in Table 3. They suggest that, out of the nonparametric strategies considered and with hyperparameter tuning, random forest yields the best result in terms of prediction accuracy. However, the said accuracy is actually not too distinct from those of other methods. In terms of AUC, however, the decision tree model outperforms others by far. Therefore, I will report the feature importance based on impurities for both the decision tree and random forest models.

Feature importances (showing top 10 variables each) are plotted in Figure 7. The results, however, are

Figure 7: Impurity-based Feature Importance Results for Random Forest and Decision Tree



Note: Panels (a) and (b) each refer to bar graphs showing top 10 highest impurity-based feature importances for random forest and decision tree, respectively.

slightly different from what were seen as significant variables in the parametric methods. While geography-related and urgency-related variables are shown as relatively important in the random forest model, variables like Christianity % and regional monthly income are shown as few of the most influential features in predicting the match status. A potential explanation is that the said two variables represent finer geographical segmentation than continents; Christianity % is country-based and regional monthly income has to do with finer regions within countries. Another explanation is that both variables are continuous ones and therefore their importances likely to be inflated (Parr et al. 2018). The latter explanation may also be the reason why the age variable, also roughly continuous, has a high feature importance value.

C. Conclusion and Limitations

To summarize, the parametric and nonparametric methods do show that there are significant (in terms of statistical significance for parametric methods and feature importance for nonparametric ones) variables for determining match statuses. Such important variables include the badge variables (i.e., urgency, urgency×wait days, exploitation-area, and AIDS-area) as well as geography-related variables like the continent and country dummy variables. This information can be implemented in arguing that there should be more than one type of PRs in terms of utility yielded from matching, which leads to the statement that the CSO should employ non-trivial inventory optimization strategies. Among the estimations used above, it seems that survival regression and fuzzy RDD are the most useful in understanding how different variables might affect match status.

However, the above results come short in the following areas, in addition to the limitations mentioned in Subsection 3.2. Firstly, while the goal of nonparametric methods was not to build a model for prediction's sake, accuracy of roughly 60% is not too great given that approximately 43.4% of the PRs in the baseline dataset acquire matches by the end of window of observation. This deficiency may suggest that more variables, such as visual cues within the PR photographs, may be necessary for not only better prediction but also in building parametric models like survival regression. Further, even if there are multiple types of PRs as this subsection suggests, one cannot confirm whether the cost of caring for unmatched lower-utility-yielding PR types is higher for the CSO. In order to verify the said point, one will need, for instance, gather additional data about the CSO's financial situations or seek the said CSO's cooperation.

4.2 Results from Identifying Potential Optimization by the CSO

Individual variables comparison. — As mentioned in Subsection 3.3, the tests involving only one variable is limited in their abilities to holistically assess whether information is displayed heterogeneously across different pages. However, it may still give one insights into how the information is structured throughout the pages as well as what variables might the CSO be more interested in varying to yield more sponsorship. The p-values from comparing each neighboring pages using t-tests (for continuous variables) and χ^2 tests of independence are presented in Table 4. I report the said statistics for six variables that yielded some statistical significance in affecting match status as observed from the previous subsection.[11]

If one uses $\alpha = 0.05$ as the threshold for detecting heterogeneity of information across neighboring pages, it seems that (according to Table 4) there is not enough evidence to confirm there is significant variation in information representation. The statement is true except for the variables such as urgency$\times$wait days and continents where evidence for variation is visible as early as page 8 for urgency$\times$wait days and page 3 for continents. As manifested in the results of Subsection 4.1, geographic variables and signs of urgency were seemingly important in determining the match statuses, more so than some of the other variables involved. Therefore, while still inconclusive, the variability in the said features can be understood as a part of the CSO's attempt for optimization.

Factor scores comparison. — After viewing the proportion of variances explained by each factor, I choose to extract five factors through EFA using the heterogeneous correlation matrix. This choice is due to the said number of factors explaining more than 60 percent of the total variance in the data, a threshold suggested in Hair et al. (2010, pp.109-110). In constructing the heterogeneous correlation matrix, the following variables were utilized: urgency $\times$wait days, age, and dummy variables for schooling, female, AIDS-area, exploitation-area, Asia, South America, and Central America. Note once again that Figure 8 depicts the percentage of variance explained by each factor up to six factors.

The rows indicating factor scores in Table 5 exhibit the results from the MANOVA tests across different pages with the factor scores produced using the aforementioned five factors. If one considers the variability in factor scores as that in PR types and set a threshold of $\alpha = 0.05$, then the earliest change in types occurs on average at around page 4 based on the mean p-values reported. This partial conclusion is also true considering the said table's standard errors and also their depiction in Figure 9. It may also be remarked that the said result is in alignment with the initial changes for the variables urgency$\times$wait days and continents

[11] I note that "continents" refer to the categorical variable that combines the variables AIDS-area (Africa), Asia, South America, and Central America into one. Previously, in Subsection 4.1, the dummy variables for each continent was used because the said categorical variable is not ordered.

Table 4: p-values for Page-by-page Comparisons of Individual Variables

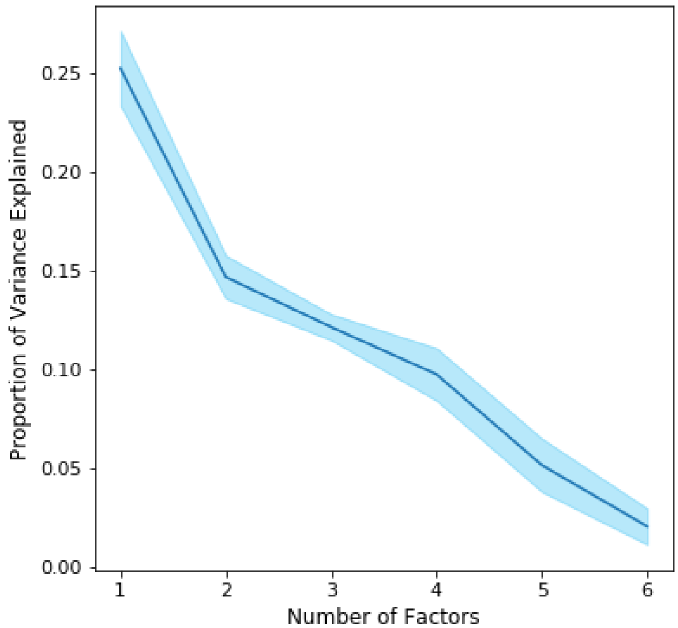
	Pages						
	2	3	4	5	6	7	8
Urgency×wait days	0.109	0.091	0.054	0.111	0.059	0.082	0.000
	(0.062)	(0.076)	(0.041)	(0.068)	(0.036)	(0.072)	(0.000)
Age	0.398	0.334	0.120	0.164	0.299	0.265	0.127
	(0.105)	(0.102)	(0.048)	(0.093)	(0.101)	(0.090)	(0.082)
Schooling	0.393	0.341	0.113	0.133	0.397	0.327	0.140
	(0.113)	(0.127)	(0.05)	(0.081)	(0.137)	(0.104)	(0.096)
Female	0.402	0.734	0.521	0.378	0.332	0.394	0.594
	(0.117)	(0.070)	(0.092)	(0.108)	(0.097)	(0.111)	(0.107)
Continents	0.168	0.088	0.033	0.017	0.102	0.096	0.039
	(0.074)	(0.080)	(0.020)	(0.008)	(0.056)	(0.083)	(0.023)
Exploitation area	0.201	0.243	0.140	0.339	0.188	0.250	0.132
	(0.083)	(0.123)	(0.049)	(0.123)	(0.082)	(0.115)	(0.080)
Average N per page	35.7	36.5	38.3	38.1	37.3	38.1	38.2
	Pages						
	9	10	11	12	13	14	15
Urgency×wait days	0.106	0.012	0.058	0.055	0.167	0.075	0.029
	(0.084)	(0.007)	(0.054)	(0.045)	(0.087)	(0.042)	(0.027)
Age	0.121	0.206	0.072	0.123	0.351	0.348	0.374
	(0.04)	(0.076)	(0.053)	(0.075)	(0.121)	(0.119)	(0.099)
Schooling	0.333	0.413	0.156	0.097	0.311	0.183	0.226
	(0.103)	(0.133)	(0.09)	(0.063)	(0.13)	(0.08)	(0.104)
Female	0.506	0.490	0.310	0.222	0.479	0.540	0.468
	(0.106)	(0.128)	(0.092)	(0.094)	(0.114)	(0.106)	(0.110)
Continents	0.164	0.080	0.011	0.023	0.192	0.197	0.081
	(0.074)	(0.036)	(0.007)	(0.021)	(0.078)	(0.103)	(0.050)
Exploitation area	0.301	0.360	0.045	0.062	0.403	0.215	0.233
	(0.115)	(0.129)	(0.024)	(0.051)	(0.109)	(0.124)	(0.124)
Average N per page	35.7	35.1	36.9	35.7	35.2	33.0	33.8

NOTE: SE across different inventories in parentheses. Average N per page for page 1 is 33.6. All p-values are garnered by comparing the indicated page (e.g., page 2) with the previous page (e.g., page 1). Those for continuous variables (i.e., Urgency×wait days and Age) were acquired by using two-sample t-test; those for categorical variables (i.e., all others) were acquired by using χ^2 test of independence. Data used in acquiring these statistical results were collected daily from the Compassion International website (on July 15, 2020 and from July 19, 2020 to July 26, 2020).

occurring at similarly early pages.

Despite that the results show evidence of the CSO roughly following what is anticipated by Corollary 2.1, it is worrisome that the standard errors are too large. One interpretation is that perhaps the CSO anticipates different PS search behavior across days, and tries to alter the position for initial change of PR types. For instance, if the PRs are shown to search more deeply into the inventory during the weekend than during the weekdays, then it would make sense for the CSO to have the initial change of types at the later positions during the weekends. Because the current data is limited in its coverage of different days, this is one hypothesis that would be difficult to confirm within the scope of this study. Another, more plausible reasoning may have to do with inadequacy in the use of EFA. As the five factors extracted from EFA only account for 60 to 70 percent of the total variance in the data, factor scores created from the said factors may be exaggerated by extrapolating how much they can explain. A follow-up study involving more data points and a more carefully-designed EFA may be useful in confirming such concerns.

Figure 8: Exploratory Factor Analysis; Proportion of Variance Explained by Each Factor



NOTE: Solid line represents the mean, and the shaded area represents the range between 2 standard errors above and below the said mean.

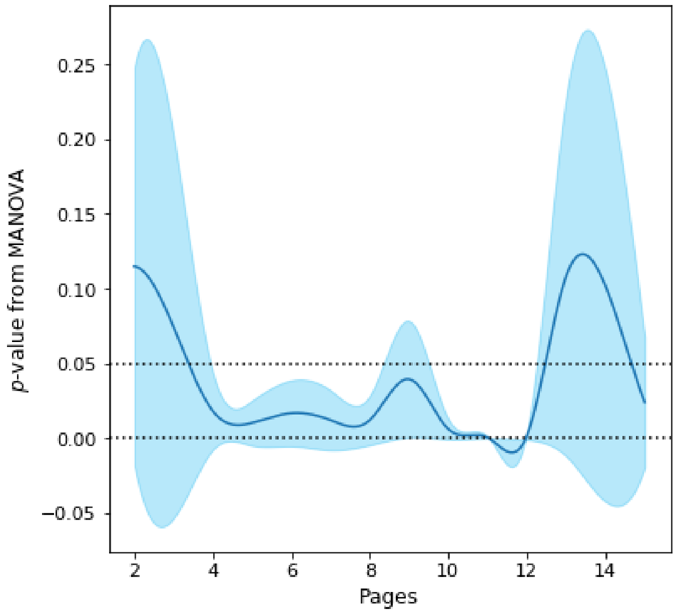
Table 5: p-values for Page-by-page Comparisons of Factor Scores and Predicted Match Probabilities

	Pages						
	2	3	4	5	6	7	8
Factor scores	0.115	0.074	0.018	0.010	0.017	0.012	0.012
	(0.067)	(0.064)	(0.013)	(0.008)	(0.011)	(0.010)	(0.008)
Pred. match prob.	0.211	0.400	0.192	0.163	0.101	0.299	0.077
	(0.032)	(0.043)	(0.034)	(0.034)	(0.021)	(0.043)	(0.018)
Average N per page	35.7	36.5	38.3	38.1	37.3	38.1	38.2
	Pages						
	9	10	11	12	13	14	15
Factor scores	0.039	0.006	0.000	0.001	0.108	0.102	0.024
	(0.020)	(0.003)	(0.000)	(0.001)	(0.060)	(0.072)	(0.022)
Pred. match prob.	0.176	0.172	0.060	0.045	0.261	0.231	0.210
	(0.026)	(0.036)	(0.013)	(0.014)	(0.034)	(0.037)	(0.033)
Average N per page	35.7	35.1	36.9	35.7	35.2	33.0	33.8

NOTE: SE across different inventories in parentheses. Average N per page for page 1 is 33.6. All p-values are garnered by comparing the indicated page (e.g., page 2) with the previous page (e.g., page 1). Those for factor scores were acquired by using MANOVA; those for pred. match prob. (predicted match probabilities) were acquired by using ANOVA. Data used in acquiring these statistical results were collected daily from the Compassion International website (on July 15, 2020 and from July 19, 2020 to July 26, 2020).

Predicted match probabilities comparison. — The match probabilities in this exercise are predicted using the hyperparameter-tuned random forest model in conjunction with the baseline dataset. This model is exactly the one with the highest prediction accuracy among different competing models as described in the previous subsection. Note that this "best" model's prediction was still quite low (at around 60 percent); while it was proficient at predicting those that will *not* yield a match after the window of observation, it was

Figure 9: p-values for Page-by-page Comparisons Using Factor Scores



NOTE: Solid line represents the mean, and the shaded area represents the range between 2 standard errors above and below the said mean. The graphs have been smoothed using cubic interpolation.

poor at guessing those that actually will.

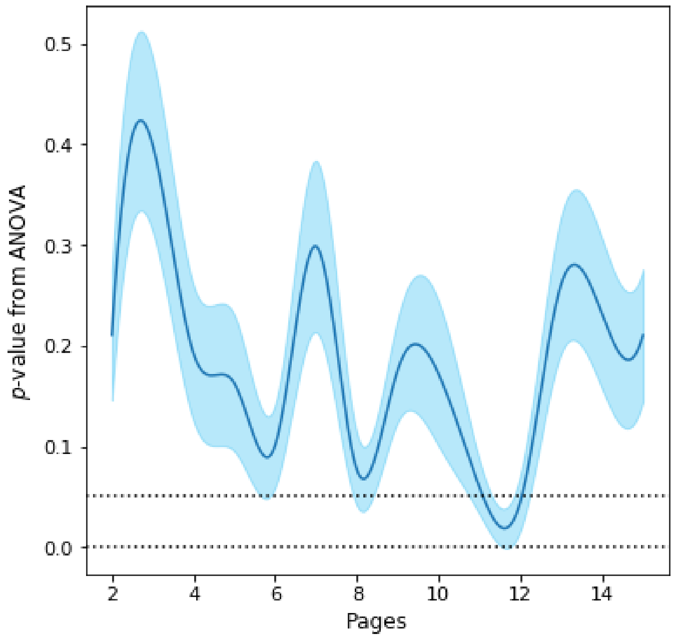
Perhaps owing to this imprecision, the ANOVA results using predicted match probabilities (presented in Table 5) exhibit a lack of evidence in much variation across pages. Similar to the case using factor scores, if one understands the evidence for variability as that for change in PR types, then on average, the initial change in PR types is projected to happen at page 12 (using $\alpha = 0.05$). As also depicted in Figure 10, the standard errors are not as large as those in estimations with factor scores. However, this quality might be due to, once again, the "best" model's shortcomings in prediction accuracy.[12]

Despite the ambiguous results above, a positive aspect in using predicted match probabilities is that, instead of only finding evidence on where the initial PR type changes occurs, one can also infer the types that the said changes occur from and to. As described in Section 3, one may understand PRs with higher match probability as more utility-yielding for the PSs when matched. It is visible for Figure 11 that the average per-page predicted match probabilities are actually higher not for the forefront pages, but slightly beyond that at pages 5 to 7. With the idea linking match probabilities and PS utility levels, the first initial transition from lower-utility to higher-utility set of PRs is more likely to occur at around page 6 rather than at page 12 (where the predicted match probability is seen to actually decrease).

Conclusion and limitations. — In summary, the tests above show that the initial transition of PR types is likely to occur at around page 6. Further, the said initial transition is likely to be one from the lower-utility PR types to higher-utility PR types (i.e., from t_l-types to t_h-types). If it is assumed that the CSO expects

[12]For instance, despite a PR eventually yielding a match, the match probability may be predicted too low, as the "best" model's greatest problem is in its exaggerated projection of false negatives.

Figure 10: p-values for Page-by-page Comparisons Using Predicted Match Probabilities



NOTE: Solid line represents the mean, and the shaded area represents the range between 2 standard errors above and below the said mean. The graphs have been smoothed using cubic interpolation.

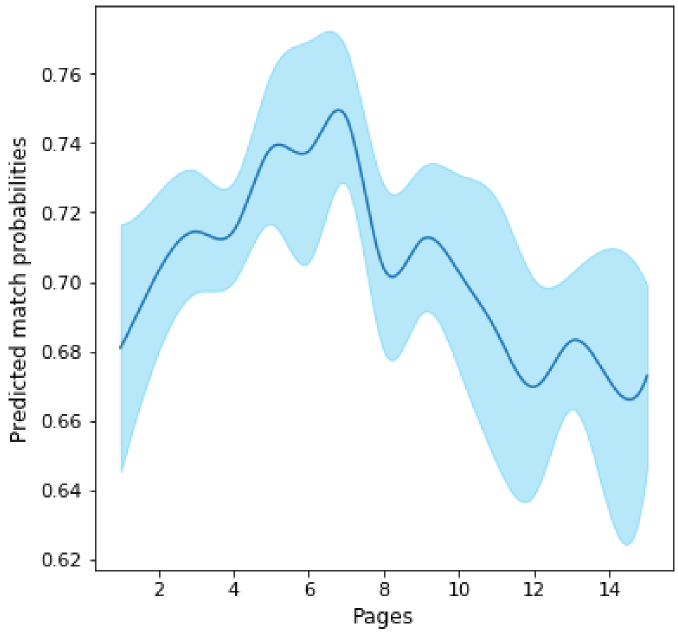
that the maximum search depth is around 5 pages (i.e., $b = 5$), then the results of this subsection are actually in accordance with that of Corollary 2.1. While it is not to state that the CSO is optimizing throughout its entire set of inventories, it can be concluded that its behavior is within the theory predictions of Section 2.

Nonetheless, the above results come short in the following areas. Firstly, it is not very certain that the true maximum search depth is around 5 pages. If the PSs actually exit the inventory much earlier than 5 pages, then one may refer to Corollary 2.1 to state that the CSO is not optimizing. Unfortunately, because the data collected for this paper does not encompass those for tracking online search or click activities, it would not be possible to firmly state about maximum search depths. Secondly, as previously mentioned, the models used in creating data to be used for hypotheses testing (i.e., EFA and random forest) may need improvements in their respective criteria to produce more unambiguous results. For the future studies in both EFA and the various machine learning classification models, it may be necessary to gather more data about other relevant variables that this study failed to collect and use them to better metrics like explained variance (for EFA) and prediction accuracy (for the machine learning methods).

5 Conclusion

In this paper, I explored the CSO's problem of maximizing its expected payoff by considering the strategies in ordering different PRs' information sets, referred to as inventory design (optimization). In doing so, I suggested a simple economic framework involving different types for PRs and PSs, which led to a partial

Figure 11: Predicted Match Probabilities from the Random Forest Model



NOTE: Solid line represents the mean, and the shaded area represents the range between 2 standard errors above and below the said mean. The graphs have been smoothed using cubic interpolation. The random forest model (utilized in producing the predicted match probabilities) has been cross-validated (using 5-folds CV).

characterization of an optimal strategy in inventory design. This result, in turn, helped remove many potential candidates for the said optimal strategy, thereby lessening the computational task of finding the optimal inventory.

Furthermore, in conjunction with data from Compassion International, I employed various empirical strategies to identify variables influential to match statuses, so that the hypothesis for multiple PR types could be confirmed. I also drew upon additional empirical approaches to verify any discrepancy between what the economic framework predicts versus the strategy that the CSO seems to be implementing. The results, while not conclusive, show that there are significant features such as geography- and urgency-related variables affecting match results. Additionally, the strategy that the CSO supposedly uses is within the boundaries of economic theory's prediction.

Future theoretical extensions should attempt to find rules about spacing between different types of PRs as I exhibited that the so-called bunching strategies can be sub-optimal in many versions of the inventory optimization problem. Regarding empirical extensions, it would be crucial to acquire data about additional variables that can tell a researcher more about the behaviors of PSs. Such information can be utilized to confirm some of the restrictive assumptions that the study took advantage of and build better empirical models to describe the CSP environment.

References

Amos, Orley. 1982. "Empirical Analysis of Motives Underlying Contributions to Charity." *Atlantic Economic Journal* 10 (4): 45-52, December.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *Economic Journal* 100 (401): 464-477, June.

Baccara, Mariagiovanna, Allan Collard-Wexler, Leonardo Felli, and Leeat Yariv. 2014. "Child-Adoption Matching: Preferences for Gender and Race." *American Economic Journal: Applied Economics* 6 (3): 133-158.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–1678.

Bennett, Roger. 2003. "Factors Underlying the Inclination to Donate to Particular Types of Charity." *International Journal of Nonprofit and Voluntary Sector Marketing* 8 (1): 12-29.

Berry, Steven, JamesLevinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841-890.

Bou-Hamad, Imad, Denis Larocque, and Hatem Ben-Ameur. 2011. "A Review of Survival Trees." *Statistics Surveys* 5: 44-71.

Brest, Paul. 2012. "A Decade of Outcome-Oriented Philanthropy." *Stanford Social Innovation Review*: 42–47, Spring.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica*, 82 (6): 2295-2326, November.

Carpenter, Jeffrey, and Caitlin Knowles Myers. 2010. "Why Volunteer? Evidence on the Role of Altruism, Image, and Incentives." *Journal of Public Economics* 94: 911-920.

Central Intelligence Agency. n.d. "Country Comparison: HIV/AIDS Adult Prevalence Rate." `https://www.cia.gov/library/publications/resources/the-world-factbook/fields/363rank.html`.

Ciuca, Vasilica, and Monica Matei. 2010. "Survival Analysis for the Unemployment Duration." *Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation* 1: 354-359.

Compassion International. 2018. "Annual Report 2017-2018." `https://www.compassion.com/multimedia/2018-annual-report-compassion-international.pdf`

Conlon, Christopher, and Jeff Gortmaker. 2020. "Best Practices for Differentiated Products Demand Estimation with PyBLP." Forthcoming in the *RAND Journal of Economics*, July 24.

Cox, David R. 1972. "Survival Analysis for the Unemployment Duration." *Journal of the Royal Statistical Society, Series B (Methodological)* 34 (2): 187-220.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (6): 1431-1451.

Dunn, Olive Jean. 1961. "Multiple Comparisons Among Means." *Journal of the American Statistical Association* 56 (293): 52-64, March.

Evans, William N., and W. Kip Viscusi. 1991. "Estimation of State-Dependent Utility Functions Using Survey Data." *Review of Economics and Statistics* 73 (1): 94-104, February.

Gentzkow, Matthew, and Emir Kamenica. 2016. "A Rothschild-Stiglitz Approach to Bayesian Persuasion." *American Economic Review: Papers & Proceedings* 106 (5): 597-601.

Goldberg, Pinelopi Koujianou. 1995. "Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry." *Econometrica* 63 (4): 597-601, July

Gravois, John. 2004. "Why Do Adoptive Parents Prefer Girls?" *Slate*, January 14. https://slate.com/news-and-politics/2004/01/why-do-adoptive-parents-prefer-girls.html

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69 (1): 201-209, January.

Hair, Jr., Joseph F., William C. Black, Barry J. Babin, and Rolph E. Anderson. 2010. "Factor Analysis." In *Multivariate Data Analysis*. Upper Saddle River, New Jersey: Pearson Prentice Hall, 91-152.

Han, Hong, Xiaoling Guo, and Hua Yu. 2016. "Variable Selection Using Mean Decrease Accuracy and Mean Decrease Gini Based on Random Forest." *7th IEEE International Conference on Software Engineering and Service Science*: 219-224, August.

Huber, Chris. 2019. "Facts about How Child Sponsorship Works," March 7. https://www.worldvision.org/sponsorship-news-stories/facts-about-how-child-sponsorship-works

Johnson, Richard A., and Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101 (6): 2590-2615.

Keane, Michael P., Petra E. Todd, and Kenneth I. Wolpin. 2011. "The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications." In *Handbook of Labor Economics* 4a, edited by Orley Ashenfelter and David Card. North Holland: Elsevier, 331-461.

Kremer, Michael, Sylvie Moulin, and Robert Namunyu. 2003. "Decentralization: A Cautionary Tale." *Poverty Action Lab Paper* 10, April 15.

Kuhn, Ben. 2014. "Decision Tree for Survival Analysis," November. https://www.benkuhn.net/survival-trees

Lipnowski, Elliot, and Laurent Mathevet. 2017. "Simplifying Bayesian Persuasion." Unpublished. http://www.laurentmathevet.com/wp-content/uploads/2017/03/BPCover.pdf

Ostrower, Francie. 1995. *Why the Wealthy Give: The Culture of Elite Philanthropy*. Princeton: Princeton University Press.

Parr, Terence, Kerem Turgutlu, Christopher Csiszar, and Jeremy Howard. 2018. "Beware Default Random Forest Importances." *explained.ai*, March 26. https://explained.ai/rf-importance/.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825-2830.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2019. "Selecting the Number of Clusters with Silhouette Analysis on KMeans Clustering." Unpublished. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Rabbitts, Frances. 2014. "Give and Take? Child Sponsors and the Ethics of Giving." In *Child Sponsorship: Exploring Pathways to a Brighter Future*, edited by Brad Watson and Matthew Clarke. New York: Palgrave Macmillan, 280-296.

Rodriguez, Germán. 2010. "Survival Models." In *Lecture Notes on Generalized Linear Models*, September. https://data.princeton.edu/wws509/notes.

Tardella, Fabio. 2008. “On the Existence of Polyhedral Convex Envelopes.” In *Frontiers in Global Optimization*, edited by Christodoulos A. Floudas and Panos N. Pardalos. Boston: Kluwer, 149-188.

Tzortzis, Grigorios F., and Aristidis C. Likas. 2008. “The Global Kernel k-Means Algorithm for Clustering in Feature Space.” *IEEE Transactions on Neural Networks* 20 (7): 1181-1194, July.

Wydick, Bruce, Paul Glewwe, and Laine Rutledge. 2013. “Does International Child Sponsorship Work? A Six-Country Study of Impacts on Adult Life Outcomes.” *Journal of Political Economy* 121 (2): 393-436.

Xue, Yishu, Elizabeth D. Schifano, and Laine Rutledge. 2017. “Diagnostics for the Cox model.” *Communications for Statistical Applications and Methods* 24: 583-604, November.

A Appendix

A.1 Appendix Tables

Table 6: Types of Information Revealed by Child Sponsorship Organizations

Organization	$/Month	Types of Information
World Vision	$39	• Basic country-level information • Advanced individual-specific information • Individual health conditions also revealed, but almost all potential recipients indicated as "in satisfactory health" • Basic family-specific information • Short video accompanied
Compassion Int'l	$38	• Basic country- and region-level information • Advanced individual-specific information • Basic family-specific information • Specific badges to indicate vulnerability to AIDS and that to child exploitation, but these are binary (badge or no badge)
Save the Children	$39	• Basic country-level information • Basic individual-specific information with days without a sponsor
ChildFund Int'l	$36	• Advanced individual-level information without days without a sponsor • Basic family-specific information
Plan Int'l	$35	• Basic individual-specific information • Information about the CSO's operations in the respective country, but no basic country-level information • Detailed information about access to water (e.g., what the main source of water is, how long it takes to reach that source) • Information about the PR's housing type and how they cook • Individual health conditions also revealed, but in all cases the following message is displayed: "According to the family, (name) has been healthy and not suffered from any serious illness"

NOTE: "Int'l" refers to "International." "$/Month" refers to the monthly payment amount, in USD, required per child sponsorship in the year 2019. "Basic country-level information" refers to items such as climate, terrain, major diseases, and main diet at a nationwide scale. "Basic region-level information" refers to the same items but at a smaller, regional scale. "Basic individual-specific information" refers to the name, age, country, spoken language, and photos of the children. "Advanced individual-specific information" refers to items such as days of waiting without a sponsor, hobbies, school enrollment status, and family duties of the children in addition to the basic individual-specific information. "Basic family-specific information" refers to items like the status of guardians living with the children, guardians' employment status, and siblings (if any).

Table 7: Descriptive Statistics of the Baseline Dataset, by Match Status

	Matched	Not matched	p-value
Panel A. Demographics			
Age	4.86	4.71	< 0.001
	(1.56)	(1.36)	
Female	47.34	44.65	0.010
Schooling	65.49	67.17	0.088
Panel B. Family Information			
Living with mother	92.67	91.24	0.013
Living with father	69.51	63.59	< 0.001
Mother employment status			0.214
Regular employment	6.1	5.85	
Irregular employment	42.76	41.2	
Father employment status			< 0.001
Regular employment	14.67	14.91	
Irregular employment	49.23	43.06	
Panel C. Geographic Information			
Christianity %	72.87	77.75	< 0.001
	(28.46)	(20.93)	
Regional average monthly income (USD)	86.82	74.21	< 0.001
	(81.37)	(75.46)	
Continent			< 0.001
Africa	43.97	53.49	
Asia	9.78	2.15	
South America	21.01	16.97	
Central America	15.78	22.47	
North America and Caribbean	9.46	4.92	
Panel D. Variables with Specific Badges			
Urgency	3.36	0.02	< 0.001
Urgency× wait days	7.02	0.03	< 0.001
	(38.31)	(2.56)	
AIDS-affected area	43.97	53.49	< 0.001
Vulnerable to exploitation	34.92	43.85	< 0.001
N	4132	5386	-

NOTE: Mean entries for categorical variables are in percentage terms, and those for continuous variables are averaged by group. Standard deviations are only reported for continuous variables. p-values are calculated using t-tests for continuous variables, and χ^2 tests of independence for categorical variables. Baseline for parents' employment status is "parent not living with the child or unemployed." "Christianity %" refers the percentage of Christians in the PR's country. This is considered due to Compassion International being a Christian non-profit organization.

Table 8: Descriptive Statistics, Comparing the Baseline and Additional Datasets

	Matched	Not matched	p-value
Panel A. Demographics			
Age	4.77	4.67	< 0.001
	(1.56)	(1.36)	
Female	45.82	32.97	< 0.001
Schooling	66.44	44.31	< 0.001
Panel B. Family Information			
Living with mother	91.86	94.29	< 0.001
Living with father	66.16	77.96	< 0.001
Mother employment status			< 0.001
Regular employment	5.96	5.53	
Irregular employment	41.88	52.74	
Father employment status			< 0.001
Regular employment	14.8	2.52	
Irregular employment	45.73	43.59	
Panel C. Geographic Information			
Christianity %	75.63	63.5	< 0.001
	(28.46)	(20.93)	
Regional average monthly income (USD)	79.68	70.15	< 0.001
	(81.37)	(75.46)	
Continent			< 0.001
Africa	49.36	47.96	
Asia	5.46	22.45	
South America	18.72	11.72	
Central America	19.56	12.77	
North America and Caribbean	6.89	5.10	
Panel D. Variables with Specific Badges			
Urgency	1.47	99.98	< 0.001
Urgency$\times$ wait days	3.07	318.44	< 0.001
	(38.31)	(2.56)	
AIDS-affected area	49.36	47.96	0.014
Vulnerable to exploitation	39.98	43.39	< 0.001
N	9518	42631	-

NOTE: Mean entries for categorical variables are in percentage terms, and those for continuous variables are averaged by group. Standard deviations are only reported for continuous variables. p-values are calculated using t-tests for continuous variables, and χ^2 tests of independence for categorical variables. Baseline for parents' employment status is "parent not living with the child or unemployed." "Christianity %" refers the percentage of Christians in the PR's country. This is considered due to Compassion International being a Christian non-profit organization.

Table 9: Survival and Logistic Regression Results

	Cox		Logistic	
	(1)	(2)	(3)	(4)
Urgency	0.641	$0.282^{\dagger}$	$108.9^{\dagger}$	$87.0^{\dagger}$
	(0.212)	(0.097)	(10.0)	(9.5)
Urgency×wait days	1.002	$1.007^{\dagger}$	$-0.580^{\dagger}$	$-0.463^{\dagger}$
	(0.002)	(0.002)	(0.054)	(0.051)
Age	0.946	1.034	0.143^{***}	$0.161^{\dagger}$
	(0.063)	(0.030)	(0.047)	(0.041)
Female	$1.220^{\dagger}$	$1.298^{\dagger}$	$0.297^{\dagger}$	$0.3745^{\dagger}$
	(0.036)	(0.044)	(0.046)	(0.050)
Schooling	1.044	0.899^{*}	-0.179^{*}	$-0.275^{\dagger}$
	(0.037)	(0.049)	(0.091)	(0.079)
Living with mother	0.885^{*}	1.048	−0.128	0.092
	(0.057)	(0.071)	(0.105)	(0.116)
Living with father	1.391^{***}	1.070	0.272^{*}	-0.075
	(0.129)	(0.085)	(0.138)	(0.137)
Mother employment				
Irregular	1.280^{***}	1.094^{*}	$0.322^{\dagger}$	0.096
	(0.065)	(0.056)	(0.082)	(0.083)
Regular	1.068	1.004	0.085	0.014
	(0.086)	(0.087)	(0.126)	(0.131)
Father employment				
Irregular	0.814^{**}	0.952	−0.097	0.112
	(0.073)	(0.077)	(0.134)	(0.136)
Regular	$0.693^{\dagger}$	0.915	-0.314^{**}	0.021
	(0.068)	(0.075)	(0.143)	(0.140)
AIDS-area (Africa)	$0.566^{\dagger}$		-0.658^{***}	
	(0.096)		(0.221)	
Exploitation area	0.654^{***}		$-0.880^{\dagger}$	
	(0.105)		(0.204)	
Asia	1.683^{**}		$1.079^{\dagger}$	
	(0.366)		(0.278)	
South America	$0.653^{\dagger}$		$-0.747^{\dagger}$	
	(0.065)		(0.151)	
Central America	$0.434^{\dagger}$		$-1.213^{\dagger}$	
	(0.044)		(0.147)	
Regional monthly income	1.001	1.000	0.001^{**}	0.001
	(0.001)	(0.000)	(0.001)	(0.001)
Christianity percent	1.002		0.000	
	(0.003)		(0.004)	
N	9518	9518	9518	9518
Pseudo-R^2			0.152	0.205
Page dummies	Yes	Yes	Yes	Yes
Country dummies	No	Yes	No	Yes

NOTE: † :< 0.001, *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$. SE clustered at the page-level in parentheses. Columns (1) and (2) report hazard ratio estimates from Cox regression, with the dependent variable being the conditional hazard rate; (3) and (4) report coefficients from logistic regression, with the dependent variable being the binary match status after the window of observation (= 1 if matched). Values less than 1 for columns (1) and (2) indicate reducing the hazard rate (i.e. correlated with longer period of not being matched). Baseline for parent employment is "unemployed or absence of parent"; that for page dummies is page 1. Constant estimates omitted for logistic regressions.

Table 10: Random Coefficients Logit Results, Linear Equation Estimates

	Country (1)	Center (2)	(3)	k-means (4)	(5)
Constant	3×10^{21}	82.20	-117.9	51.16	310.2
	(4×10^{19})	(0.021)	(0.001)	(0.004)	(0.009)
Occupancy	-3×10^{20}	0.065	1.698	-15.20	0.182
	(7×10^{19})	(0.000)	(0.003)	(0.003)	(0.000)
Urgency	1×10^{20}	84.90	-62.9	346.7	147.4
	(3×10^{18})	(0.003)	(0.000)	(0.005)	(0.008)
Urgency×wait days	-4×10^{18}	-3.180	-5.350	-6.080	-5.478
	(7×10^{17})	(0.155)	(0.004)	(0.100)	(0.097)
Schooling	-2×10^{20}	15.30	-341.4	-70.41	-6.972
	(4×10^{19})	(0.062)	(0.004)	(0.032)	(0.085))
Mother	-2×10^{21}	-19.51	-89.74	-73.49	16.654
	(6×10^{19})	(0.086)	(0.001)	(0.006)	(0.091)
Father	-4×10^{20}	-40.89	41.46	-217.6	-49.36
	(8×10^{19})	(0.186)	(0.005)	(0.040)	(0.089)
Age	-1×10^{20}	-1.069	225.6	63.00	-0.621
	(2×10^{19})	(0.267)	(0.017)	(0.067)	(0.296)
Female	3×10^{20}	-1.231	61.79	-58.73	-22.55
	(5×10^{19})	(0.128)	(0.000)	(0.045)	(0.083)
Mother employment					
Irregular	1×10^{20}	-6.868	102.3	19.16	6.403
	(7×10^{19})	(0.087)	(0.010)	(0.036)	(0.164)
Regular	-9×10^{20}	-15.39	19.76	-50.34	39.35
	(2×10^{20})	(0.078)	(0.000)	(0.026)	(0.080)
Father employment					
Irregular	7×10^{19}	34.23	82.27	38.40	23.37
	(7×10^{19})	(0.226)	(0.010)	(0.051)	(0.103)
Regular	2×10^{20}	48.94	40.21	301.4	47.77
	(2×10^{20})	(0.030)	(0.001)	(0.010)	(0.037)
Regional monthly income	-8×10^{17}	-1.425	-8.219	-7.356	-6.701
	(1×10^{17})	(0.338)	(0.056)	(0.358)	(0.363)
Exploitation area	-7×10^{19}	-27.05	-360.1	-31.99	-46.64
	(2×10^{19})	(0.102)	(0.016)	(0.054)	(0.127)
Aids area	-4×10^{20}	-45.26	-136.5	1.029	-119.7
	(3×10^{19})	(0.056)	(0.004)	(0.028)	(0.059)
Asia	-3×10^{20}	-15.00	-138.3	275.7	-34.92
	(3×10^{19})	(0.041)	(0.000)	(0.042)	(0.122)
South America	-2×10^{20}	16.95	292.3	363.8	33.06
	(2×10^{19})	(0.055)	(0.002)	(0.016)	(0.108)
Central America	-2×10^{20}	13.57	-187.8	186.8	16.98
	(2×10^{19})	(0.159)	(0.001)	(0.040)	(0.110)
Page	-2×10^{18}	-2.568	-8.504	-4.796	-5.755
	(3×10^{17})	(0.140)	(0.054)	(0.390)	(0.161)
N	24	1518	12	50	489
Convergence	Yes	Yes	Yes	Yes	Yes

NOTE: All estimates statistically significant at the $\alpha = 0.001$ level, except for those in column (1) of father employment variables, mother irregular employment (significant at $\alpha = 0.1$), and exploitation area (significant at $\alpha = 0.01$). SE in parentheses. Indications above the column numbers refer to the method by which group-level statistics were created. For columns (3), (4), and (5), the grouping was completed by using k-means clustering with the numbers of clusters being 12, 50, and 500. Observations with the "market share" being 0 are removed. Total market share is initialized as $\frac{2}{9}$.

Table 11: Countries Marked as AIDS-area and Exploitation-area

Badge variable	Countries classified as vulnerable
AIDS-area	Burkina Faso, Ethiopia, Ghana, Kenya, Rwanda, Tanzania, Togo, Uganda
Exploitation-area	Bangladesh, Burkina Faso, Ethiopia, Haiti, Rwanda, Tanzania, Togo, Uganda

NOTE: Countries listed neither as AIDS-area nor as exploitation-area are as follows: Bolivia, Brazil, Colombia, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Indonesia, Mexico, Nicaragua, Peru, Philippines, Sri Lanka, and Thailand.

Table 12: Nonparametric Methods Used in the Study

Classifier	Abbrev.	Lin.	HPT Strategy
Decision Tree	DT	N	Splitter, minimum samples for split
Random Forest	RF	N	Number of trees, max. depth, max. # of features, etc.
Support Vector Machine	SVM	Y	Penalty parameter (C), gamma, shrinking
Quadratic Discriminant Analysis	QDA	N	Regularization parameter, tolerance
Multilayer Perceptron	MLP	N	Hidden layer size, $l2$, activation fn.

NOTE: For linear discriminant analysis, this study uses the least-squares solver. "Lin." refers to whether a method is a linear one or not. HPT and CV refer to hyperparameter tuning and k-folds cross-validation, respectively. "fn." refers to function.

Table 13: Features Used in the Nonparametric Methods

List of Features
Age, female, urgency, urgency×wait days, AIDS-area, exploitation-area, Christianity %, Schooling, Living with mother, Living with father, Mother employment status dummies, Father employment status dummies, Regional monthly income, Asia, South America, Central America

A.2 Appendix Figures

Figure 12: Percent of Showing Urgency by Pages