The Structure of the Non-Profit Organization

Canice Prendergast*

University of Chicago

October 7, 2025

Abstract

We address how non-profits allocate services to clients in need. We offer a pecking order for mechanisms, ranging from assignment through rigid rules to allowing clients to choose at subsidized prices. When its funding is low, the non-profit allocates by bureaucratic rule or delegates to a social service provider. Neither outcome reflects client preferences. This is because efficiency is sacrificed for needed funding, a feature that does not arises in the market design literature. It incorporates client preferences through Deferred Acceptance only when clients can pay little but it has enough alternative funds. Client choice with subsidized prices arises when average client income is sufficiently high and clients exhibit decreasing risk aversion. Limited choice may arise at lower client income, including one analogous to markets with artificial currency.

*Thanks to Jacob Leshno, Eduardo Acevedo, and seminar participants at the NBER, University of Mannheim and Northeastern University for helpful comments.
canice.prendergast@chicagobooth.edu.

"You have not told me all concerning yourself;
and how shall I choose better than you?"
J.R.R. Tolkien,
The Fellowship of the Rings.

The private social sector is enormous. Non-profits employ more people than public education or manufacturing, accounting for 10% of private sector U.S. employment and expenditures of $3 trillion.[1] The majority of non-profits distribute services to clients in areas such as social assistance, housing, education and health care.[2] This involves an amalgam of market and non-market features: sometimes clients choose services while in other instances services are assigned (rationed). This choice occurs in the shadow a need for funding. Non-profits consistently report lack of funding as their greatest concern. Furthermore, they receive 49% of funding from fees from clients, with an additional 22% from donors. There is almost no economic research on how non-profits distribute services to clients. Our interest is in how best to do so.

The design of mechanisms to transfer goods and services to clients has been a central theme of modern microeconomics. For the for-profit sector, this has resulted in fields like optimal auctions, non-linear pricing, quality distortions, and bundling. The literature on matching without prices assigns students to schools, patients to kidneys, and so on. Our starting point is that neither reflects the reality of the non-profit. The literature on profit-seeking firms typically designs mechanisms that best sell to high willingness-to-pay clients. For non-profits that exist to serve those in financial need, this is anathema to their mission. On the other hand, the market design literature on matching does not address the firm's need to fund itself. Social organizations sit squarely in the middle, needing funding but not using willingness-to-pay to get it. We model that reality and offer a pecking order for the non-profit - ranging from rigid bureaucratic rules to client choice - based on the source and level of its funding.

The model has two components: a mechanism to provide services, and a budget constraint is limited by revenue from clients and donors. The firm chooses both which clients to serve and which service to offer each client. We assume that clients have private information: for instance, a client who seeks housing knows where she wants to live. As a result, client choice has obvious appeal in overcoming "how shall I choose better than you?" above. However, choice involves prices and non-profits do not typically price based on willingness-to-pay: instead, services are subsidized and often rationed. We assume that a client's need and her ability-to-pay can diverge.

[1] 12.8m people work in non-profits, compared to 12.7m in manufacturing, roughly 1.5m in public high schools, 8.8m overall in public schools and universities, 10.9m in construction and 6.8m in finance and insurance. The non-profit sector is also growing far more rapidly than employment overall. Non-profit employment has increased by 33% over the last fifteen years compared to 9% in the for-profit sector (Bureau of Labor Statistics, 2024, USA Facts, 2023, BLS, 2023).

[2] 83% of non-profits are in the social assistance, health care and educational sectors, which have obvious clients (BLS, 2023).

The problem with pricing is that poorer clients may suffer compared to wealthier ones.[3] To address this tradeoff, the firm must take a position on access to service. We assume that pricing is governed by a *Fairness Constraint*, where no client can be denied for lack of income. For many service providers, this constraint defines their mission of serving those in need.

Since Gale and Shapley, the gold standard for understanding non-market assignment has been matching without prices. That literature emphasizes incorporating client preferences through mechanisms like Deferred Acceptance or Serial Dictatorship.[4] We instead study the canonical non-market setting where a firm provides services to clients in need, where funding (with fair prices) limits coverage. We show a wider set of optimal assignment mechanisms, including the kind of impersonal allocation rules common to the sector. This arises as ex post efficiency - typically a *sine qua non* of the matching literature - is optimally sacrificed for funding. More than this, client-focused Pareto efficient outcomes are a luxury good, only used only when firm funding is sufficiently high.

We begin by addressing a non-profit choosing between four common mechanisms. *Unrestricted Choice* allows a covered client to choose any service provided at fair prices. It is the analog of Serial Dictatorship in the market design literature. It can be used only if it satisfies the Fairness Constraint that chosen outcomes reflect client need rather than ability to pay. When Choice fails that test, Assigned Consumption arises. *Deferred Acceptance* incorporates client preferences through strategy proof rankings in a familiar way. *Bureaucratic Assignment* assigns treatments based on the firm's information alone - for example, a client's family structure, housing status or employment history - where client rankings play no role. Finally, non-profits use social service providers or agents to implement service. *Agent Assignment* allows autonomy to that social service provider to choose services, again without input from clients.

First consider Assigned Consumption. Deferred Acceptance has higher allocative efficiency - matching services to need - than the other Assigned Consumption mechanisms. This is because outcomes are Pareto efficient. However, incorporating client preferences in this way is costly for two distinct reasons. First, Deferred Acceptance has the foundational component of strategy proof rankings by clients. These are costly here, because the feature that limits client choice - that fairness may be violated by benefiting higher income clients - also constrains clients from ranking them. We show that this problem can be overcome only by reducing prices paid by clients, thereby reducing funding.

[3]This was the central problem in two recent market design initiatives. Feeding America was aware that choice with pricing in fiat money could avoid many of the inefficiencies of assigning food to food banks. Yet it did not do so as some food banks had more monetary resources than others. Similarly, Covax, which was charged with distributing a billion vaccines during the Covid-19 pandemic, was aware that pricing could mitigate countries taking vaccines that they could not distribute. Yet if they were to price vaccines, higher income countries would disproportionately benefit.

[4]See Roth, 2008, Pathak and Sonmez, 2008, Abdulkadiroglu et al., 2005, etc.

The second reason relates to the importance of agent autonomy, a central concern of the extensive empirical literature reviewed below. Disagreement among stakeholders is a characteristic of many mission-based organizations. We assume that agents disagree both with the firm and each other over client needs. Because of this, Agent Assignment reduces allocative efficiency. However, agents are motivated by altruism towards clients, and value their job by "making a difference" to client welfare. This implies Agent Assignment also reduces wages. This arises not because there is an agency problem per se, but instead because the identify of the agent has no impact on outcomes under Deferred Acceptance. This reduces his marginal impact, thereby increasing wages. We also offer a second reason why Pareto efficiency is costly, as it increases wages even though client welfare is commonly known to be higher.

Because of these two issues, the firm faces a tradeoff between allocative efficiency and client coverage. Deferred Acceptance has the lowest level of client coverage. More generally, client coverage and allocative efficiency are substitutes among Assigned Consumption mechanisms. Given these problems, the appeal of client choice is obvious. Unrestricted Choice can be used only when the interests of the firm and client are aligned at optimal prices. This alignment is endogenous. When clients exhibit Decreasing Absolute Risk Aversion, (subsidized) Unrestricted Choice satisfies Fairness if and only if average client income is sufficiently high. If this is the case, it raises the most revenue from clients.

Non-profits have many constituencies and objectives: employees, clients and a desire for consistency in outcomes. We offer a simple pecking order for Assigned Consumption. When the firm is poorly funded, it prioritizes employees by using Agent Assignment. With somewhat higher funding, it transitions to an emphasis on process through Bureaucratic Assignment. Among Assigned Consumption mechanisms, Deferred Acceptance is only used if the firm has high funding and is poorly informed about client needs. Client input to allocations in this sense is a luxury good. As a counterfactual, if funding does not depend on the firm's allocative choices, as arises in the existing literature, the firm only used Deferred Acceptance. However, we review an extensive empirical literature on social service providers showing that the key tradeoff is between Agent Assignment and Bureaucratic Assignment.

The pecking order changes when average client income is sufficiently high. Here Unrestricted Choice is possible and only competes with Agent Assignment. Once again, incorporating client preferences - here through choice - is a luxury good. The link between choice and client income generates endogenous increasing returns to client ability-to-pay. Non-profits that serve poor clients not only receive little revenue from them, but distort allocative efficiency by for example rendering service for clients dependent on the identity of one's bureaucrat. By contrast, wealthier clients not only get to choose their preferred service, but the additional revenue generated by choice allows more wealthier clients to be served. In this way, we identify how within the non-profit sector there are endogenous increasing returns to ability to pay.

These outcomes assume indivisible services and strategy proof mechanisms. Two

commonly used market designs relax these: when goods are divisible and when strategy proofness is not imposed. First, when services are divisible, *Restricted Choice* gives clients a choice between a larger quantity of a less desired good and a smaller quantity of a more desired one. With appropriate quantity restrictions and prices, this mechanism overcomes the Fairness problem for any level of client income. It also raises more revenue than Assigned Consumption. Quantity distortions decline with average client income and converge to Unrestricted Choice. Restricted Choice can be implemented by a common market design, namely allowing clients access to artificial currency. The central problem with Restricted Choice is poor targeting of high priority clients. We use two recent market design settings, Feeding America and Covax, to illustrate this targeting issue.

Second, many social organizations implement a mixture between assignment and choice, where clients can choose from undersubscribed services but oversubscribed ones are assigned. We show that such *Partial Choice* can be implemented at lower levels of client income than Unrestricted Choice through a (non-strategy proof) First Choice mechanism. Our objective here is not to show that welfare can exceed that of Assigned Consumption: this has been shown before. Instead, it raises more revenue than Assigned Consumption and can be implemented at lower client income than Unrestricted Choice.

Despite its size, there is almost no work on the non-profit's assignment problem.[5] However, in its focus on assignment with inequality concerns, our outcomes relate to the literature on inequality averse mechanism design (Dworczak et al., 2021, Dworczak, 2024, Akbarpour et al., 2024). That literature similarly assumes a divergence between willingness-to-pay and welfare. Its objective is largely to identify conditions under which non-market mechanisms - those that involve rationing - arise. We also address when rationing arises. However, our primary interest is to offer guidance on the optimal *form* of rationing, and how it depends on key parameters. Our assumptions also differ in important ways from that literature to more closely reflect the reality of the non-profit. This is discussed below.

We begin in Section 1 by describing a model of the non-profit. Section 2 considers Assigned Consumption, while Section 3 addresses Choice. Section 4 offers a pecking order for the non-profit and considers partial choice outcomes. Fairness is key for the pecking order. We show that the results continue to hold when the firm can exclude low income clients from service. Section 5 considers extensions. Section 6 concludes.

1 The Model

A firm provides treatments (services) $t_i \in \{0, 1\}$ with intensity $I_i \in \{0, 1\}$ to clients i, where $i = 1, 2, 3, ...\infty$. The return to treatment depends on a client i's privately

[5]Rose Ackerman, 1982, addresses how non-profits allocate excessive resources to fund-raising, as a way of business stealing from other non-profit organizations.

known type $\tau_i \in \{0, 1\}$. The surplus from treatment for client i is

$$T(\tau_i, t_i) = t_i\mu + \phi\mathbb{1}(t_i = \tau_i). \tag{1}$$

The parameter $\mu > 0$ scales vertical differentiation and $\phi > 0$ reflects horizontal differentiation or the return to matching τ_i to t_i. Our objective is to characterize outcomes in terms of these two parameters.

Our basic model assumes that the firm wishes to match t_i to τ_i. We assume that the supply of types $\tau \in \{0, 1\}$ and the distribution of treatments $t = \{0, 1\}$ are balanced and uniform, where there are $\frac{1}{2}$ of each type and treatment.[6] (Below we show how the qualitative results extend to both unbalanced demand and supply and a more general technology on T.) We also assume that the number of clients is large to avoid sampling error. The objective of the firm is to distribute its services to as many deserving candidates as its funding will allow. Its surplus is

$$S = \sum_{i=1}^{\infty}(v_i + T(\tau_i, t_i))I_i \tag{2}$$

where v_i represents a client-specific value that the firm gets from treating client i. A central feature of the market design literature is firm priority over clients, here represented by v_i. Clients are ordered to reflect diminishing returns, where $v_i = v - \delta i$, where $\delta \geq 0$. To allocate services to clients, the firm chooses an assignment mechanism A described below. The firm charges prices $P = \{p_0, p_1\}$, where $p_j \geq 0$ is the unit charge for treatment t_j.

Clients: The utility of client i with type τ_i and marginal utility of income m_i, receiving treatment t_j for a unit price p_j, is given by

$$U_i^c = (T(\tau_i, t_j) - m_i p_j)I_i. \tag{3}$$

We assume that m_i is privately known and (among those served) is defined over a support $[\overline{m}, \underline{m}]$. Let the poorest clients have marginal utility $\underline{m}$ and the richest $\overline{m} < \underline{m}$. Income and type τ are independent. (We relax this below.)

There are many settings in the social sector that fulfill this combination of vertical and horizontal differentiation.[7] As a leading example, imagine that a housing

[6] The interpretation of the supply of treatments is that as the firm expands, it hires more employees (agents) or expands its assets. Some employees (assets) are better than others, with type $t = 1$ the better one. Each client is matched with an employee (asset), and most clients would prefer the better one. However, higher quality assets are better matched to the $\tau = 2$ client. This is discussed below.

[7] In ascending desirability for the typical client: (i) Covid-19 vaccines: Sinopharm, Astra Zeneca, Pfizer, (ii) food banks: produce, dairy, frozen chicken, (iii) housing: larger units, or units in more desirable locations, and (iv) mental health care: drugs, drugs plus therapist check-in monthly, intensive therapy.

association is assigning clients to two cities, *Chicago* ($t = 0$) and *NewYork* ($t = 1$). Then $T(k, j)$ for clients is

$$\begin{array}{c} \tau = \\ Chicago \\ NewYork \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} \phi & 0 \\ \mu & \mu + \phi \end{bmatrix} \end{array}$$

The problem is only of interest if client preferences do not lead to trivial outcomes. The technology $T(\tau_i, t_i)$ and supply of types implies that τ_i should receive treatment $t_i = \tau_i$. If clients always preferred that allocation, the firm would always allow choice. We choose preferences such that client ranking in the absence of prices is partially informative. We assume that ϕ is random, equal to $\overline{\phi}$ with probability h (h for "horizontal" preferences) and $\underline{\phi} < \overline{\phi}$ with probability $1 - h$. ϕ is private information for the client. To render the problem interesting, we make Assumption 1.

Assumption 1: $\overline{\phi} > \mu > \underline{\phi}$.

Absent pricing, those with $\phi = \underline{\phi}$ prefer $t = 1$, while those with $\overline{\phi}$ prefer τ_i. In the example above, a majority $1 - \frac{h}{2}$ prefer *New York* to *Chicago*. Let $\hat{t}_i = (t_i^1 \succ t_i^2)$ be the preference ranking of client i, where t_i^j is her jth preferred treatment.[8] Let $\hat{\phi} = h\overline{\phi} + (1 - h)\underline{\phi}$. Efficient assignment entails $\tau_i = t_i$, with expected surplus $T^* = \frac{\mu}{2} + \hat{\phi}$.

Agent: The firm hires an agent to implement t_i. Any agent a observes a (contractible) signal τ_i^a on each served client's type where $\tau_i^a = \tau_i$ with probability q, where $\frac{1}{2} \leq q < 1$. The signal τ_i^a represents the kind of observable data - family size, employment history, age, residential location etc. - used to allocate social services. Agent a has utility U^a defined over wages w and client welfare:

$$U^a = \sum_{i=1}^{\infty} (w + \alpha \hat{T}_i^a) I_i. \tag{4}$$

where α reflects altruism towards the client and $\hat{T}_i^a$ is agent a's about client i's expected utility from services. The agent has a reservation utility $\overline{w}$ per unit of I.[9] A random agent a is offered the position at a wage of w per unit of I. If he accepts the position, his utility per client receiving service is $U^a = w + \alpha \hat{T}^a$. If he does not,

[8] As we assume that all clients must be served, we do not need to include the null set in the preference ranking.

[9] First, $\overline{U}$ need not be the compensation that the agent could receive elsewhere, but instead could include any "warm glow" discount for working for the social organization. Second, the model has constant returns to scale with respect to $I = \sum I_i$. This is assumed to avoid wired-in benefits from the firm becoming large and to address the allocation of fixed costs.

let $U^a = \overline{w} + \alpha \overline{T}^a$ be his utility per client, where $\overline{T}^a$ is agent a's belief of the client's utility if another agent is employed.

We assume that the agent accepts any offer per client where

$$w \geq \overline{w} + \alpha(\overline{T}^a - \hat{T}^a), \tag{5}$$

supported by a belief by an agent that *some* other agent will accept the position if he does not.[10] $\overline{T}^a - \hat{T}^a$ derives from differences in beliefs about τ_i. After observing τ_i^a, the agent holds that posterior with probability λ, but with probability $1-\lambda$, has a prior that a client is type $\tilde{\tau}$ with probability ≈ 1, where $\tilde{\tau}$ is uniformly distributed. We assume λ is independently distributed across agents. Conditional on two agents holding a belief $\tilde{\tau}$, there is a probability $\rho \geq \frac{1}{2}$ of the agents agreeing. Let $\tilde{\tau}_i^a$ be the agent's posterior belief over τ_i after observing τ_i^a.

Non-profits rely on donors. External donations are D.[11] Our interest is in the non-profit organization, which is legally defined by the Non-Distribution Constraint (Hansmann, 1980). Accordingly, all revenue is spent on programming. Let $p = \frac{p_0+p_1}{2}$ be average revenue per client. Client coverage depends on two cases. When $w > p$, donor-funding limits client coverage as

$$I = \frac{D}{w-p}. \tag{6}$$

The novelty of the paper derives from (6) as w and p are endogenous to the mechanism. The first best level of treatment for the firm is $\frac{v+T^A}{\delta}$. To avoid trivial outcomes, we assume $I < \frac{v+T^A}{\delta}$. When $p \geq w$, the firm is self-funded without donors. Unless otherwise stated, we consider the case where the firm is not self-funded.[12]

Inequality Aversion: The firm has preferences to reduce inequality across clients. To reflect this, we assume that treatment is independent of m_i:

Fairness Constraint: $Et_i(\tau_i|m_i) = Et_i(\tau_i), EI_i(\tau_i|m_i) = EI_i(\tau_i)$, for all m_i.

Note that this assumption implies that not only must all clients be served equally $EI_i(\tau_i|m_i) = EI_i(\tau_i)$, but they also must receive the same expected type of service $Et_i(\tau_i|m_i) = Et_i(\tau_i)$ independent of income. This assumption, allied to the objective of the firm, is a key difference from the literature on inequality averse mechanism design (Dworczak et al., 2021, Dworczak, 2024, Akbarpour et al., 2024). This is discussed below.

[10] This rules out implausible equilibria where all agents believes that if he does not accept the position, no other agent will do so.

[11] D refers to any source of funding aside from clients, such as endowments, government or foundation support.

[12] This characterization rules out the use of Rotating Savings mechanisms, where clients pool income and randomize who receives treatment. See Besley et al., 1993.

1.1 Assignment Mechanisms

We now address the set of assignment mechanisms considered. There are three actors: clients, the firm, and the agent. Allowing for input from each of these, and for now considering only strategy proof mechanisms with indivisible services, results in the following matrix of mechanisms:

Client Input	**Firm Input**		**Agent Input**		**None**
Yes	*Deferred*	*Acceptance*	*Agent*	*DA*	*Choice*
No	*Bureaucratic*	*Assignment*	*Agent*	*Assignment*	.

Assigned Consumption: when the firm or agent has input to outcomes, the four items on the left of the matrix.

- *Deferred Acceptance*: Assignment is based on τ_i^a and the client's ordinal ranking $\hat{t}_i$ in a standard DA mechanism.
- *Bureaucratic Assignment*: Assignment is based on τ_i^a
- *Agent Assignment*: Assignment is based on $\tilde{\tau}_i^a$.

One final case is a combination of two mechanisms:

- *Agent Deferred Acceptance*: Assignment is based on $\tilde{\tau}_i^a$ and the client's ordinal ranking $\hat{t}_i$.[13]

Choice: when the firm plays no role other than through price setting.

- *Unrestricted Choice*: The client chooses either treatment at price p_i.

Strategies and timing: The timing is as follows:

1. The firm chooses A, P^A and w^A. All are observed by clients and potential agents.
2. The firm offers the position to one agent subject to (5).
3. All clients $i = 1, 2, ...I^A$ report $\hat{t}_i$. Equilibrium requires that $\hat{t}_i \perp m_i$.
4. The firm and agent observes τ_i^a for all $i = 1, 2, ..., I^A$.
5. A is implemented, p_i^A paid by clients, and w^A is paid, where

[13]Random Assignment has input from neither, but is always dominated by Bureaucratic Assignment.

- If A = Deferred Acceptance, t_i is allocated through an algorithm described below.
- If A = Bureaucratic Assignment, $t_i = \tau_i^a$,
- If A = Agent Assignment, $t_i = \tilde{\tau}_i^a$,
- If A = Unrestricted Choice, $t_i = \hat{t}^1$ if services are not oversubscribed.

Given this timing, the objective of the firm is to choose A, w^A and P^A to maximize surplus subject to (3), (4), (5), (6), the Fairness Constraint, and Assumption 1. The equilibrium concept used is Bayesian Nash. The program is formally defined in the Appendix.

The purpose of the model is to offer a tractable structure to address tradeoffs between allocative efficiency and funding. Let T^A be average per capita treatment surplus from assignment mechanism A and p^A be average revenue per client. Then $I^A = \frac{D}{w^A - p^A}$. As a result, surplus is

$$S^A = \frac{D}{w^A - p^A}\left(v + T^A - \frac{\delta D}{2(w^A - p^A)}\right). \tag{7}$$

Surplus depends on three endogenous variables: (i) wages w^A, (ii) client payments p^A, and (iii) assignment surplus T^A.

Note one feature before describing outcomes. This is not a setting where the client cannot identify the appropriate treatment. In such instances, the optimal outcome often entails the firm assigning consumption to the client.[14] Our interest is not in such cases, but rather why Assigned Consumption would arise with informed clients.

2 Assigned Consumption

Assigned Consumption arises when the firm or agent at least partially determines services. From the timing, the firm knows I^A when allocating t_i. It offers $I_i = 1$ to all $i \leq I^A = \frac{D}{w^A - p^A}$ and 0 to all other clients. The allocations from A are as follows, and follow the choice of P^A and w^A:

Bureaucratic Assignment: $\tau_i^a = t_i$.

Agent Assignment: $\tilde{\tau}_i^a = t_i$.

Unrestricted Choice involves the firm choosing prices for each service. Depending on parameters, choice may entail efficient allocations where $t_i = \tau_i$ or may involve

[14] For instance, Bansak et al., 2025, design a mechanism for allocating asylum seekers to regions based on employment prospects in those regions, a factor that is unlikely to be known by the clients.

standard monopoly distortions. To focus on the novel components of the analysis, we consider the case where the market allocation is efficient, $t_i = \tau_i$. Below we extend all results to the monopoly pricing case. As a result, we restrict attention here to:

Unconstrained Choice: $\tau_i = t_i$ for all τ_i and m_i.

Deferred Acceptance:

1. All clients $i < I^{da}$ report rankings $\hat{t}_i = (t_i^1 \succ t_i^2)$. The firm and agent observe $\hat{t}_i$.
2. The firm allocates $t = 1$ to all clients $i < I^{da}$ for which $\tau_i^a = 1$ and $\hat{t}_i = (1 \succ 0)$. If there are remaining $t = 1$ slots, it randomly assigns them to clients for which $\tau_i^a = 0$ and $\hat{t}_i = (1 \succ 0)$ until there are no more $t = 1$ slots.
3. The firm allocates $t = 0$ to all other clients.

Two features determine outcomes: (i) when Choice can be used, and (ii) the level of client coverage under each mechanism. Lemma 1 simplifies.

Lemma 1:

- *Unrestricted Choice satisfies the Fairness Constraint if and only*

$$\frac{\overline{m}}{\underline{m}} \geq \frac{\mu - \underline{\phi}}{\mu + \underline{\phi}}. \tag{8}$$

  In that case, maximum prices are $P^{uc} = \{p_0^{uc}, \frac{\underline{\phi}+\mu}{\underline{m}}\}$, *where* $p_0^{uc} = \min\{\frac{\underline{\phi}-\mu}{\overline{m}} + \frac{\mu+\underline{\phi}}{\underline{m}}, \frac{\underline{\phi}}{\underline{m}}\}$. For $\frac{\overline{m}}{\underline{m}} \geq 1 - \frac{\underline{\phi}}{\mu}$, $p_0^{uc} = \frac{\underline{\phi}}{\underline{m}}$.

- *Agent Assignment and Bureaucratic Assignment can be implemented for any income level with maximum prices* $P^{ac} = \{0, \frac{\mu}{\underline{m}}\}$ *with revenue per client* $p^{ac} = \frac{\mu}{2\underline{m}}$.

- *Deferred Acceptance can be implemented for any income level. For* $\frac{\overline{m}}{\underline{m}} < \frac{\mu-\underline{\phi}}{\mu+\underline{\phi}}$, *maximum prices* $P^{da} = \{0, \frac{\mu-\underline{\phi}}{\underline{m}}\}$ *with revenue per client* $p^{da} = \frac{\mu-\underline{\phi}}{2\underline{m}} < p^{ac}$.

There are three significant components of Lemma 1. First, (8) implies that Unrestricted Choice can only arise if clients are sufficiently similar in their willingness to pay. Intuitively, this is more likely when there is more horizontal differentiation and less likely with more vertical differentiation.

Second, Unconstrained Choice raises more revenue than Assigned Consumption. In any equilibrium where type $\tau = i$ receives treatment $t = j$ with positive probability, Individual Rationality implies

$$p_j \leq min_i \frac{T(i,j)}{\underline{m}}. \tag{9}$$

As $q < 1$, any Assigned Consumption outcome implies that each type receives both services in equilibrium, so Individual Rationality requires that $p_0 = 0$, and $p_1 \leq \frac{\mu}{\underline{m}}$. Maximum revenue from any Assigned Consumption mechanism is $p^{ac} = \frac{\mu}{2\underline{m}}$. All clients receive the efficient outcome under Unrestricted Choice so maximum revenue is $p^{uc} = \frac{\frac{\mu}{2}+\underline{\phi}}{\underline{m}}$.

Finally, Deferred Acceptance raises less revenue than any other mechanism. In addition to Individual Rationality, Deferred Acceptance has an incentive compatibility constraint that $\hat{t}_i$ be independent of m_i. Consider the housing example with $\mu = 5$ and $\underline{\phi} = 2$. Then $T(k,j)$ for clients with $\phi = \underline{\phi}$ is

$$\begin{array}{cc} \tau = & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} Chicago \\ NewYork \end{array} & \begin{bmatrix} \mathbf{2} & \mathbf{0} \\ \mathbf{5} & \mathbf{7} \end{bmatrix} \end{array}$$

Bureaucratic and Agent Assignment raises revenue of $\frac{2.5}{\underline{m}}$ as Chicago is priced at 0 and New York at $\frac{5}{\underline{m}}$. From (8), the firm can use Unconstrained Choice if $\frac{\overline{m}}{\underline{m}} \geq \frac{3}{7}$.[15] If $\frac{\overline{m}}{\underline{m}} < \frac{3}{7}$, the only client information that can be revealed through Deferred Acceptance is their ranking of $T(k,j)$. Fairness requires that all clients with type $\underline{\phi}$ report *New York* $\succ$ *Chicago*, which requires that the price difference between them cannot be larger than $\frac{5}{\underline{m}} - \frac{2}{\underline{m}} = \frac{3}{\underline{m}}$. As a result, maximum DA prices are *Chicago* $= 0$ and *New York* $= \frac{3}{\underline{m}}$, with expected revenue of $\frac{\mu - \underline{\phi}}{2\underline{m}} = \frac{1.5}{\underline{m}}$. In this way, incorporating client preferences in allocations is costly, by reducing the number of clients that can be served.

Said differently, fairness requires that outcomes cannot vary between richer and poorer clients. If both services were free, this automatically arises. If not, poorer clients may be excluded from services in a way not true for wealthier clients. This limits prices.

Proposition 1 *Surplus from Deferred Acceptance is*

$$S^{da} = \frac{D}{\overline{w} - p^{da}} \left(v + T^{da} - \frac{\delta D}{2(\overline{w} - p^{da})} \right). \tag{10}$$

where

$$T^{da} = \frac{\mu}{2} + q\hat{\phi} + \frac{h(1-q)\overline{\phi}}{2} + \frac{r}{2}((1-q)\hat{\phi} - q(1-h)\underline{\phi}), \tag{11}$$

[15] It can charge $p_0 = \frac{2}{\underline{m}}, p_1 = \frac{7}{\underline{m}}$ when $\frac{\overline{m}}{\underline{m}} \geq \frac{3}{5}$. This yields revenue of $\frac{\frac{\mu}{2}+\underline{\phi}}{\underline{m}} = \frac{4.5}{\underline{m}}$.

and $r = \frac{h(1-q)}{1-hq}$.

Outcomes are independent of agent beliefs, so $\hat{U}^{ca} = \overline{U}^{ca}$ and hence $w^{da} = \overline{w}$ from (5). T^{da} reflects the benefits of Deferred Acceptance. Pareto efficiency requires that no two clients would swap allocations (and prices) while still satisfying Fairness. Deferred Acceptance is Pareto efficient, and the final two terms in (11) reflect its utilitarian benefits.[16]

Following Balinski and Sonmez, 1999, this outcome is equivalent to the one-sided matching outcome where each service is a player with a priority list given by τ_i^a, where r reflects the required lottery.[17] This interpretation allows us to relate to another common mechanism, Random Serial Dictatorship. This arises here if the firm is uninformed of the client's type, $q = \frac{1}{2}$.

Proposition 2 *Surplus from Bureaucratic Assignment is*

$$S^b = \frac{D}{\overline{w} - p^{ac}} \left(v + T^b - \frac{\delta D}{2(\overline{w} - p^{ac})} \right). \tag{12}$$

where $T^b = \frac{\mu}{2} + q\hat{\phi}$.

This outcome follows automatically because the allocations do not depend on information provided by the client. Bureaucratic Assignment has worse allocative efficiency than Deferred Acceptance as it is Pareto inefficient but higher client coverage.

Proposition 3 *Surplus from Agent Assignment is*

$$S^a = \frac{D}{w^* - p^{ac}} \left(v + T^a - \frac{\delta D}{2(w^* - p^{ac})} \right). \tag{13}$$

where

$$T^a = \mu + (\lambda q + \frac{1-\lambda}{2})\hat{\phi}, \tag{14}$$

and $w^* = \overline{w} - \alpha(1-\lambda)(1 - \frac{\lambda}{2} - (1-\lambda)\rho)\hat{\phi}$.

Agent Assignment has lower allocative efficiency than Bureaucratic Assignment, where $T^b - T^a = (1-\lambda)\hat{\phi}(q - \frac{1}{2})$. Yet it has higher client coverage. This is because $\hat{T}^a > \overline{T}^a$, as she now "makes a difference". As a result, $\overline{w} - w^* \equiv \Delta w = \alpha(1-\lambda)(1 - \frac{\lambda}{2} - (1-\lambda)\rho)\hat{\phi} > 0$. Intuitively, the wage discount is higher when the extent of disagreement $(1-\lambda)$ is greater, when the agent cares more about the client (α), and when the agent believes that an alternative agent is less likely to share his views (ρ lower). Agent Assignment has the worst allocative efficiency among Assigned Consumption mechanisms but the highest client coverage.

[16] This is through the $\frac{h}{2}$ clients that do not report $t = 1$ as their first choice, and through r, the lottery over assignments to $t = 1$ among those who rank $t = 1$ highest but whose signal to the firm suggests assignment to $t = 0$.

[17] Those who receive a signal $\tau_i^a = 1$ are given a choice over both services. A fraction $\frac{h(1-q)}{2}$ will choose $t = 0$ and the remainder choose $t = 1$. The firm then randomly chooses from those with $\tau_i^a = 0$ the remaining $\frac{h(1-q)}{2}$ slots for $t = 1$ and gives all other clients $t = 0$.

Agent Deferred Acceptance and the Cost of Pareto Efficiency: Agent Deferred Acceptance is where the DA algorithm is implemented using the agent's posterior rather than τ_i^a. In addition to lower client revenue, there is a novel cost associated with Pareto efficiency.

There is an intuitive appeal to Agent Deferred Acceptance. To see this, consider a counterfactual if wages remained at w^*. Prices are p_j^{da}, so surplus would be $\tilde{S}^{da} = \frac{D}{w^* - p^{da}} \left(v + \tilde{T}^{da} - \frac{\delta D}{2(w^* - p^{da})} \right)$, where $\tilde{T}^{da}$ is the surplus from a Deferred Acceptance algorithm with signal $\tilde{\tau}_i^a$ rather than τ_i^a.

However, this outcome does not arise. This is because wages increase, even though the agent is altruistic and average client welfare strictly higher. Agent Deferred Acceptance allows the $I^a \frac{h}{2}$ clients who in our example prefer *Chicago* to *New York* to be assured of receiving it. This increases client welfare, and the agent's welfare as he is motivated by client altruism. However, wages depend on marginal client welfare which falls because $\frac{h}{2}$ clients in effect allocate themselves. This shows the cost of a typically desired feature of market design outcomes, Pareto efficiency.

This allows us to distinguish our outcome from that of the agency literature. Agent Assignment has a flavor of an agency problem, as the agent receives benefits from his perception of the client's welfare.[18] To see this, note that the agent is motivated by altruism, not private benefits as is usually the case. The agent knows with probability 1 that client welfare is higher under Agent Deferred Acceptance than under Agent Assignment. Yet our altruistic agent demands a higher wage with Agent Deferred Acceptance. The key issue here is that the identity of the agent does not matter for outcomes under Deferred Acceptance and Bureaucratic Assignment. Said another way, impersonal algorithms may increase efficiency, but do little to recruit altruistic agents. A discount only arises when the agent influences the outcome, and fears her alternative may be worse.

[18] A common outcome in the agency literature on social organizations is the need to offer more autonomy to agents than in a standard profit-seeking firm. This can take the form of excessive delegation of control (Aghion and Tirole, 1997), hiring agents whose preferences do not coincide with that of the firm (Prendergast, 2007, Che and Kartik, 2009), or ignoring likely legitimate complaints against bureaucrats (Prendergast, 2003).

3 Unrestricted Choice

Many non-profits offer subsidized choice to their clients.[19] As Unrestricted Choice delegates to the client, it is the analog to Serial Dictatorship in the market design literature. After choosing $P^{uc} = \{p_0^{uc}, p_1^{uc}\}$, the mechanism to implement this is as follows:

1. Each client $i \leq I^{uc}$ reports $\hat{t}_i$.
2. The firm allocates all clients their first choice if i is undersubscribed.
3. For all $t = i'$ that are oversubscribed, the firm rations randomly.

So far, we have treated m_i as parametric. The parameter m_i derives from a client's indirect utility function $V(y_i)$, where $y_i \geq 0$ is her income, where her utility is $V(y_i - p_j) + T(\tau_i, t_j)$. (Below we interpret this.) We assume that $V'(0) = \infty, V^{'} > 0, V^{''} < 0$ and $V^{''} \to 0$ as $y \to \infty$.[20] To offer conditions for when (8) arises, let $y_i = \hat{y} + \eta_i \geq 0$, $E\eta = 0$ and $y \sim [\hat{y} + \underline{\eta}, \hat{y} + \overline{\eta}]$. Average income is $\hat{y}$. Finally, assume Non-Increasing Absolute Risk Aversion for clients, where $-\frac{V^{''}(y)}{V^{'}(y)}$ is non-increasing in y.[21]

Proposition 4 *With Decreasing Absolute Risk Aversion, (8) holds for all average income levels above $\hat{y}^{uc}$, where*

$$\frac{V'(\hat{y}^{uc} + \overline{\eta})}{V'(\hat{y}^{uc} + \underline{\eta})} \equiv \frac{\overline{m}^{uc}}{\underline{m}^{uc}} = \frac{\mu - \underline{\phi}}{\mu + \underline{\phi}}. \tag{15}$$

With Constant Absolute Risk Aversion, (8) either holds for all average income levels or for no average income levels. Surplus from Unrestricted Choice is

$$S^{uc} = \frac{D}{\overline{w} - p^{uc}} \left(v + T^* - \frac{\delta D}{2(\overline{w} - p^{uc})} \right), \tag{16}$$

where $T^* = \frac{\mu}{2} + \hat{\phi}$.

[19] There are many examples. First, housing associations often offer vouchers to qualified clients who choose from available housing units. Second, food pantries now typically allow clients to choose food (rather than being offered a pre-packed basket). Third, food pantries typically receive their food from the regional food bank, and choose food from a menu. Some foods are free, but the pantries must purchase other food at subsidized prices. Fourth, job seekers who receive training can tailor that retraining to their preferences (MacEachan et al., 2013). Finally, many mental health providers now offer patients choice over therapy options (Cook et al., 2019).

[20] We model m_i as variation in client income. This could alternatively be interpreted as savings, liquidity, an absence of debt, or access to alternative support. Clients do not care about v_i. This is to offer preferences for the firm that differ from those of the client. Allowing these as part of preferences only adds person-specific variation to prices.

[21] See Chiappori and Paiella, 2011, Guiso and Paiella, 2008, and Huber et al., 2023, for empirical evidence supporting DARA.

Wage are given by $\overline{w}$ as $\hat{U}^{ca} = \overline{U}^{ca}$. Prices are as in Lemma 1. The firm can induce efficient choice at $\hat{y}^{uc}$ with $p_0^{uc} = 0$. As average income rises above $\hat{y}^{uc}$, the firm can increase p_0^{uc} until it reaches its maximum of $\frac{\underline{\phi}}{\underline{m}}$ above average income of $\tilde{y}^{uc} > \hat{y}^{uc}$, where $\frac{V'(\tilde{y}^{uc}+\overline{\eta})}{V'(\tilde{y}^{uc}+\underline{\eta})} = 1 - \frac{\underline{\phi}}{\underline{\mu}}$. Proposition 4 illustrates how Choice not only yields efficient assignment but also raises more revenue than either Bureaucratic Assignment or Deferred Acceptance.

3.1 Three extensions

To focus on the key issues, we have restricted attention to Unrestricted Choice pricing where the efficient outcome $\tau_i = t_i$ arises. The firm may instead use monopoly pricing to garner more revenue at the expense of allocative efficiency. This arises as there is a fraction h clients who value the efficient match at $\overline{\phi}$ rather than $\underline{\phi}$. A sufficient condition for efficient pricing is $h\overline{\phi} \leq \underline{\phi}$. In the Appendix, we consider the possibility of monopoly pricing. This involves offering $t = 1$ for sure at a higher price than p_1^{uc} above, or a random allocation of $t = 0$ or $t = 1$ at a lower price. This extension does not change the logic and qualitative outcomes above.

Second, (8) relies on a single value of μ for all clients. Consider now the case where $\mu_i \in (\underline{\mu}, \overline{\mu})$. The analog of Assumption 1 is then $\overline{\phi} > \overline{\mu} > \underline{\mu} > \underline{\phi}$. The efficiency of matching clients based on ϕ now requires that $\overline{\mu} - \underline{\phi} \leq \underline{\mu} + \underline{\phi}$. The Fairness Condition now requires that $\frac{\overline{m}}{\underline{m}} \geq \frac{\overline{\mu}-\underline{\phi}}{\underline{\mu}+\underline{\phi}}$, which is harder to satisfy than (8). But $\frac{\overline{\mu}-\underline{\phi}}{\underline{\mu}+\underline{\phi}} \leq 1$ by the definition of efficiency, so it remains the case that for high enough income, Unrestricted Choice is feasible. However, this extends (8) to show that Unrestricted Choice requires not only limited variation in capacity to pay but also limited variation in the valuation of $t = 1$.

Finally, we have assumed that the range in client income does not depend on its average. More generally, let $\overline{\eta}$ and $\underline{\eta}$ be indexed by $\hat{y}$. The result that Choice can be used when average client income is high enough holds if $\frac{V'(\hat{y}+\underline{\eta}(\hat{y}))}{V'(\hat{y}+\overline{\eta}(\hat{y}))}$ does not increase too rapidly with $\hat{y}$.[22] When randomness scales in a linear way with $\hat{y}$, $\eta(\hat{y}) = \eta\hat{y}$, Unrestricted Choice arises for sufficiently high income if $\frac{V''(\hat{y}+\overline{\eta}(\hat{y}))}{V'(\hat{y}+\overline{\eta}(\hat{y}))}\frac{1+\overline{\eta}}{1+\underline{\eta}} < \frac{V''(\hat{y}+\underline{\eta}(\hat{y}))}{V'(\hat{y}+\underline{\eta}(\hat{y}))}$. This is the analog to Decreasing Relative Risk Aversion in our setting.

[22] When $\underline{\eta}'(\hat{y}) \geq 0$ the analysis holds if $\overline{\eta}' < \frac{\left(\frac{V''(\hat{y}+\underline{\eta}(\hat{y}))}{V'(\hat{y}+\underline{\eta}(\hat{y}))}\right)}{\left(\frac{V''(\hat{y}+\overline{\eta}(\hat{y}))}{V'(\hat{y}+\overline{\eta}(\hat{y}))}\right)} - 1$. This characterization holds for any parallel shift in supports, where $\underline{\eta}' = \overline{\eta}'$.

4 The Non-Profit Outcome

Our analysis is based on a premise uncontroversial to those in the social sector: that money matters, and that it primarily comes from clients. We now provide a pecking order of assignment mechanisms based on the level and source that funding. Outcomes vary in three ways: allocative efficiency, wages and prices, depend on four key parameters: the level of donations D, the ability of clients to pay $\underline{m}$, the quality of the firm's information, q, and the value of the highest priority clients, measured by δ.

We first show how the choice of Assigned Consumption mechanism depends on external funding D. Consider the choice between Deferred Acceptance and Bureaucratic Assignment. Let $\Delta p = p^{ac} - p^{da} = \frac{\phi}{2\underline{m}}$ and $\Delta T = T^{da} - T^{b}$. The firm chooses Bureaucratic Assignment for all $D \leq D_b^*$, where D_b^* is uniquely defined by

$$\Delta p(v + T^b) - \Delta T(\overline{w} - p^{ac}) = \delta D_b^* \left(\frac{(\overline{w} - p^{da})^2 - (\overline{w} - p^{ac})^2}{2(\overline{w} - p^{da})(\overline{w} - p^{ac})} \right). \tag{17}$$

Now let $\Delta \tilde{T} = T^b - T^a$. The firm chooses Agent Assignment over Bureaucratic Assignment for all $D \leq D_a^*$, where D_a^* is uniquely defined by

$$\Delta w(v + T^a) - \Delta \tilde{T}(w^* - p^{ac}) = \delta D_a^* \left(\frac{(\overline{w} - p^{ac})^2 - (w^* - p^{ac})^2}{2(\overline{w} - p^{ac})(w^* - p^{ac})} \right). \tag{18}$$

Proposition 5 follows.

Proposition 5 *For $\hat{y} < \hat{y}^{uc}$, the firm uses Agent Assignment for $D \in (0, D_a^*)$, Bureaucratic Assignment for $D \in [D_a^*, D_b^*)$, and Deferred Acceptance for $D \geq D_b^*$.*

Not all mechanisms need be used. If $\Delta p(v + T^b) - \Delta T(\overline{w} - p^{ac}) < 0$, Deferred Acceptance is always preferred to Bureaucratic Assignment, while if $\Delta w(v + T^a) - \Delta \tilde{T}(w^* - p^{ac}) < 0$, Bureaucratic Assignment is always preferred to Agent Assignment. This arises for instance if $\alpha = 0$ wage savings from Agent Assignment are zero.

This offers a pecking order for Assigned Consumption. Pareto efficiency is costly for funding. The cost of such reductions in coverage is greatest when funding is low, and so ex post efficiency, the *sine qua non* of the literature, is a luxury good. For $D < D_b^*$, the firm sacrifices Pareto efficiency as the value of additional funds is higher. At the lowest level of funding, Agent Assignment is used. The firm then transitions to Bureaucratic Assignment, and only with higher funding again uses Deferred Acceptance.[23] This tradeoff for the non-profit does not arises in the literature on matching without prices. If I is independent of w^A and p^A, the firm always uses client-focused Deferred Acceptance if $\hat{y} < \hat{y}^{uc}$ and Unconstrained Choice otherwise.

[23] The pecking order requires $\delta > 0$. When $\delta = 0$, one form of Assigned Consumption is preferred for all levels of funding. Bureaucratic Assignment is preferred to Deferred Acceptance if $\frac{\Delta p}{\overline{w} - p^{ac}} \geq \frac{\Delta T}{v + T^b}$ and Agent Assignment is preferred to Bureaucratic Assignment if $\frac{\Delta w}{w^* - p^{ac}} \geq \frac{\Delta \tilde{T}}{v + T^a}$.

The model differs from the standard setting of matching without prices in many ways: an agent is involved, the firm has information, and the firm charges prices. The use of mechanisms that do not offer input from the client depends only on pricing. Let $\alpha = 0$ and $q = \frac{1}{2}$. Then if $\hat{y} < \hat{y}^{uc}$, the firm chooses between Deferred Acceptance - here Random Serial Dictatorship - or Bureaucratic Assignment - here Random Assignment. It randomly assigns for all $D < D_b^{**}$ where $\frac{\mu}{2\underline{m}}(v+T^{da}) - \frac{h\bar{\phi}}{2(2-h)}(\overline{w}-p^{ac}) = \delta D_b^{**}\left(\frac{(\overline{w}-p^{da})^2-(\overline{w}-p^{ac})^2}{2(\overline{w}-p^{da})(\overline{w}-p^{ac})}\right)$ and $T^{da} = \mu + \frac{h\bar{\phi}}{2} + \frac{\hat{\phi}+(1-h)^2\underline{\phi}}{2(2-h)}$.

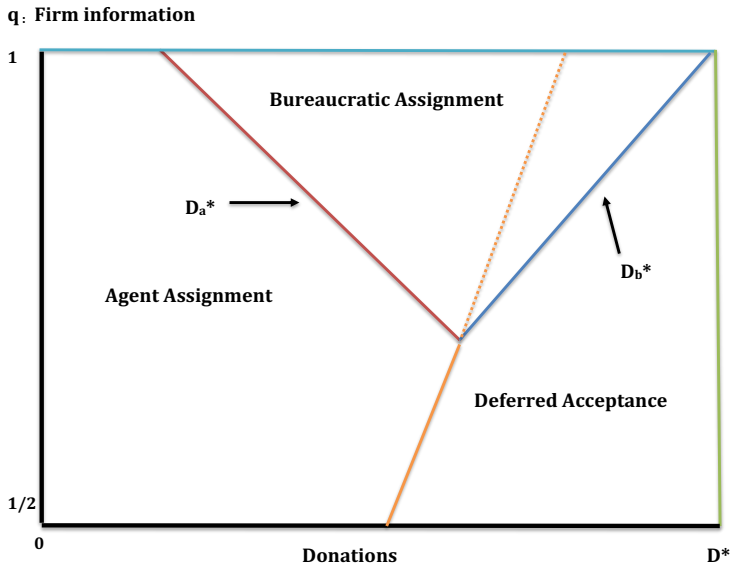


Figure 1: **Optimal Assigned Consumption.**

Figure 1 extends (17) and (18) by relating outcomes to another relevant variable, the quality of the firm's information. (D^* defines the maximum amount of donations that the firm would ever use: $D^* = \frac{v+T^{da}}{\delta}$.) Two additional outcomes arises. First, the firm never chooses Deferred Acceptance for q high enough. There is a fixed cost per client from client preferences revelation (Δp), but as $q \to 1$, $\Delta T \to 0$. As a result, Deferred Acceptance arises not only when the firm is well funded, but also poorly informed. Second, the firm always prefers Agent Assignment over Bureaucratic Assignment for q low. Agent Assignment raises more funds, but at the cost of using worse information to assign treatments. As $q \to \frac{1}{2}$, $\Delta\tilde{T} \to 0$.

Providing agency to clients through choice is a central theme of the literature on social service provision.[24] Figure 2 adds Choice and illustrates the role of the source of funding. First note that without uncertainty ($\frac{\overline{m}}{\underline{m}} = 1$), the firm can always achieve

[24]See for example, Chaumba and Locklear, 2021, and Cohen et al., 2016. From a practitioners perspective, see Food Bank News, 2024, for empirical evidence on how food bank clients value choice.

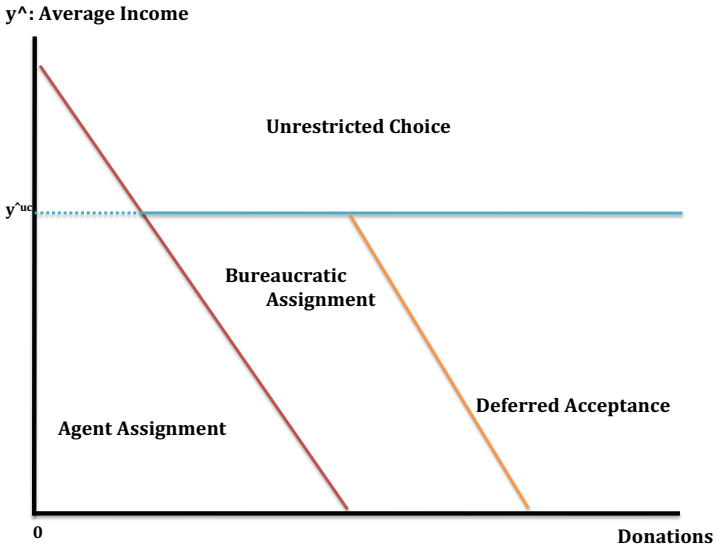


Figure 2: **Choice and the Source of Funding.**

the efficient outcome though Choice.[25] However, few non-profits have the capacity to determine a client's capacity to pay.[26] When $\frac{\overline{m}}{m} \neq 1$ and $\hat{y} < \hat{y}^{uc}$, the firm uses Assigned Consumption, in which case only total funding matters. The downward sloping lines below $\hat{y}^{uc}$ are combinations of donations and client revenue that hold total coverage constant. Deferred Acceptance is only used when donations are high, but client income is low.

The source of funding matters when $\hat{y} \geq \hat{y}^{uc}$. Then the firm chooses Unrestricted Choice over either Bureaucratic Assignment or Deferred Acceptance. The firm may choose Agent Assignment even when Unconstrained Choice is possible. This is likely if wage reductions are significant and little funding can be raised from clients. A relevant case here is the use of volunteers. 28% of Americans volunteered in 2022-23 (US Census Bureau, 2024). If $w^* = 0$ but $\overline{w} > 0$, agents are willing to volunteer but only when they are given autonomy. It is simple to show cases where the firm uses Agent Assignment in this setting, even when Unconstrained Choice is possible.

The firm has information on the cardinal preferences of clients. Ordinal preferences is that it does not impose such constraints. To address this, assume the firm only

[25] With common marginal utility m, the firm generates efficient outcomes with prices $p_0 \in [0, \frac{\phi}{m}]$ and $p_1 \in (\frac{\mu}{m}, \frac{\mu+\phi}{m}]$.

[26] Most are simply too small: 88% of non-profits have annual expenses of less than $500,000. Even those that do means testing for clients typically only use it to exclude high earning clients (Gai et al., 2021). A small number of large non-profits have access to say a client's tax return. Even for those, this may not suffice, as many non-profits exist to mitigate unexpected income shocks like unemployment or family dissolution. The uncertainty above reflects these constraints.

knows that the marginal utility of clients is positive. The worst case scenario is then that $\underline{m} = \infty$ so $p_i = 0$. The firm chooses between Agent Assignment and Deferred Acceptance, with Agent Assignment for $D < D_a^{**}$, where

$$T^{da} - T^a = \frac{\Delta w}{w^* \overline{w}} \left(v + T^a - \frac{\delta(D_a^{**} + \frac{\Delta w}{2})}{2w^*} \right). \tag{19}$$

This is given by the hashed line in Figure 1.[27]

Serving the Poorest Clients: Non-profits cover a wide range of needy clients, from the poorest members of society to those who have assets and income. Decreasing Absolute Risk Aversion allows a stark difference between non-profits that serve poorer clients and those serving wealthier ones. Those serving wealthier clients not only get more revenue through reduced m_i, but it allows the firm to transition to the higher efficiency of Unrestricted Choice, which in turn raises more revenue. By contrast, the revenue from serving the poor is so low that the firm further compromises allocative efficiency below that in Deferred Acceptance to raise needed funds. In this sense, a model that has no technological increasing returns to scale endogenously creates them.

Labor's Share: For modeling reasons, the only costs incurred by the firm are labor costs. In reality, labor is one of many costs. It is simple and intuitive to extend this to include other costs. This offers an additional implication that the gravity towards Agent Assignment increases when labor's share of costs is higher. Below, we describe empirical evidence where the benefit of Agent Assignment is high for social work, where labor's share is high. By contrast, it is unlikely that a housing association, whose greatest costs are real estate costs rather than labor, would be as likely to choose Agent Assignment.

A characteristic of mission-based organizations is many competing interests: those of clients, a desire that outcomes should be consistent, and those of the organization's employees. The analysis above offers a priority ranking based both on the level and source of funding and the quality of the firm's information. Less funding first prioritizes employees and then bureaucratic consistency, while either high level of donations or client payment lead to client preferences becoming pivotal. Note that outcomes in the non-profit sector are not a convex combination of market outcomes and those of the literature on matching without prices. Instead, mechanisms that offer clients less input than either of these are emphasized. The literature that we now review emphasizes these mechanisms.

[27] This extension may appear to offer a greater likelihood of Deferred Acceptance. However, Deferred Acceptance is only chosen when the firm has enough donations. When $p^A = 0$, fewer clients are covered for any level of donations and the required level of donations needed to transition to Deferred Acceptance is higher than above. Instead, it makes Agent Assignment more likely.

Empirical Literature: The modeling relies on three assumptions: that firms are concerned about funding, clients are an important source of revenue, and worker autonomy is important. We now provide evidence on each.

First, empirical evidence consistently points to financing concerns as predominant for non-profits: 42% report "lack of adequate finances/resources" as their biggest challenge, with 57% reporting "generating income and achieving financial sustainability" as their greatest concern (Statista, 2024, Small Charities Data, 2023).[28] Yet this is only relevant if the allocation mechanism affects revenue. As non-profits receive 49% of their revenues from fees to clients, it clearly does, with another 22% coming from donors (National Council of Nonprofits, 2024). Because of this, we focus on p^A and D as the primary sources of funds for the social organization.

Finally, the value of autonomy to social service providers is a central theme of the literature.[29] This is typically empirically identified when firms bureaucratize, where "lack of trust in social workers competence and ability to exercise discretion well has led municipalities to increasingly rely on manuals in an attempt to regulate their discretionary freedom" (Petersen et al., 2020). Pascoe et al., 2023, survey almost forty academic publications on the impact of the transition to Bureaucratic Assignment on social workers. The response is overwhelmingly negative, with many studies citing higher intention to quit after such transitions. As 29% of non-profits report "difficulty recruiting and retaining quality staff" as their biggest challenge, such hostility to bureaucracy likely has costs, particularly as workers are paid below market wages in the sector.[30] Employees feel they have become "basically glorified data-inputters" (Dustin, 2006), spend their days "checking whether you have completed your paperwork correctly" (Carey, 2009), and suffer from "mission drift from providing personalized care to cost-cutting and efficiency" (Groves et al, 2016).[31] For this large empirical literature, the key choice is between Agent Assignment and Bureaucratic Assignment.[32]

Theoretical Literature: Our work is similar in two ways to the literature on inequality averse mechanism design (Dworczak et al., 2021, Dworczak, 2024, Akbarpour et al., 2024). First, a principal assigns services, but wants to avoid inequality in outcomes. Second, a primary concern is conditions under which non-market assignments

[28] Concern about funding is even higher for small non-profits.

[29] The classic reference is Lipsky, 1980.

[30] See Statista, 2024, Leete, 2001, Ruhm and Borkowski, 2003, Salamon and Sokolowski, 2005, and Hirsch et al., 2017.

[31] For more studies corroborating this outcome, see Adler and Borys, 1996, Cantarelli et al, 2016, Langer et al, 2019, and Wright and Davis, 2003.

[32] We assume that agents know the assignment mechanism and their impact on client welfare when accepting a position at the firm. Realistically, the margin on which this matters is more likely retention of workers. With respect to this, there is a significant literature in the public administration literature correlating a service provider's intention to quit and the degree of autonomy on the job (for examples, see Liu at al., 2011, and Zhang et al., 2024).

(rationing) arise. Our work differs in both its interests and key assumptions. Most importantly, we are primarily interested in *how* services are rationed, motivated by both the market design literature without prices but also the extensive empirical literature focused on the tradeoff between autonomy and bureaucracy.

Our assumptions also have been chosen to more closely reflect a non-profit providing social services to clients. A key issue for an inequality-averse principal is access to an instrument to transfer utility to low income clients. Here this can only be achieved by making services sufficiently affordable. Dworczak et al., 2021, Dworczak, 2024, and Akbarpour et al., 2024, by contrast allow revenue generated by selling services to high income clients to be given as a lump sum transfer to low income clients.[33] This gives the principal more latitude than in our work.

Redistributing through lump sum payments to those who do not buy may be appropriate in many settings but it does not reflect the reality of the non-profit. Non-contingent money transfers from non-profits to clients are extremely rare, let alone given to clients who do not consume the firm's primary service. There are two primary reasons why this is the case. First, such transfers are illegal for the vast majority of non-profits. IRS rules allow non-profits to offer cash to clients only when those cash payments align with its primary charitable function.[34] Examples where cash is allowed are payments of scholarships by educational non-profits, hospital bills by community health agencies, or emergency financial assistance to pay specific bills like rent or utilities. Aside from these, they cannot be used.

Second, the objective of the typical non-profit in the social service sector is not to increase the utility of the disadvantaged. Instead, it is to increase the utility of the disadvantaged *through a specific mechanism.* For instance, Feeding America increases the welfare of the poor through food provision, Planned Parenthood does so through fertility services, and Medicins Sans Frontiers does so through medical expertise in lower income countries. They do not see their mission as giving money to the poor. Instead, any marginal revenue is spent on the mission. This is because non-profits perceive themselves to have a comparative advantage in providing their primary service or are inherently paternalistic. Our definition of firm surplus reflects this: it is not the utility of the client U_i^c, which would include her income, but the component that arises from the services that they provide T_i.

For these two reasons, a model of the non-profit that allows such transfers would not reflect reality. As a result, the Fairness Condition - which requires that services

[33]The authors more generally consider some alternative use for revenue that is not relevant for the non-profit comparison, and show that market mechanisms are preferred when the weight on revenue generation is sufficiently high. The market allocation - without rationing - arises when the virtual utility of the principal (plus a novel information rent term) is for wealthier clients to purchase with higher probability. Then the firm chooses the market clearing price, with revenue redistributed to those who do not buy.

[34]See for example wscadv.org/wp-content/uploads/2017/08/Grants-to-Individuals.pdf. The reason is to stop fraud, where a donor makes a tax-deductible donation and then receives a tax free benefit in return. The IRS has responded to this by disallowing unrelated cash transfers.

be affordable for all - is the only way that the firm can guarantee the absence of inequality.

More General Technology: The returns to v_i and T are separable in (7). This has been chosen to reflect the idea that for many social organizations, their first priority is to cover the most needy clients, and only then to worry about which services they receive. This implies that the indifference curve between p and T

$$\frac{dp}{dT}|_D = -\left(\frac{(w-p)^2}{(w-p)(v+T)-\delta D}\right) \tag{20}$$

flattens as D increases. Yet the result does not depend on this separability. A natural alternative allows the return to matching to increase in the priority of the client: $S = \sum_{i=1}^{\infty} v_i T(\tau_i, t_i) I_i = \frac{D}{w^A - p^A} T^A (v - \frac{\delta D}{2(w^A - p^A)})$. The analogous indifference curve becomes

$$\frac{dp}{dT}|_D = -\left(\frac{w-p}{T} + \frac{\delta(w-p)D}{2(w-p-\delta D)T}\right). \tag{21}$$

This is also increasing in D, so the indifference curve between revenue and allocative efficiency flattens. As a result, the priority ranking is retained with this technology.[35]

More General Preferences: The preference structure in (1) implies that horizontal preferences are sufficiently important. To see this, assume more generally that $T(k,j)$ for the horizontal ($\overline{\phi}$) types is $\overline{\mu}_{kj}$, and $\underline{\mu}_{kj}$ for the vertical ($\underline{\phi}$) types, where all are non-negative. The analog to Assumption 1 is (i) $\overline{\mu}_{kk} \geq \overline{\mu}_{kj}$, and (ii) $\underline{\mu}_{k1} \geq \underline{\mu}_{k0}$. Efficiency of assigning $\tau_i = t_i$ requires $\underline{\mu}_{01} - \underline{\mu}_{00} \leq \underline{\mu}_{11} - \underline{\mu}_{10}$. Assume that this is the case.

One substantive result can change with this more general technology: Deferred Acceptance may not reduce revenue.[36] Individual rationality with Assigned Consumption implies that $p_j = min_i \underline{\mu}_{ij}$ so $p^{ac} = \sum_j min_i \underline{\mu}_{ij}$. Let $p_0 \equiv min_i \underline{\mu}_{i0}$. Incentive compatibility with Deferred Acceptance implies that to rank $1 \succ 0$, the maximum

[35] One can construct examples where the priority ranking above no longer holds, by allowing more flexible diminishing returns via or by assuming the value of T is high for the highest priority clients but not those who are marginally lower on the priority rank. Yet these feel pathological, rather than the narrative here that coverage of more clients is especially important for poorly funded non-profits. Even in instances where the priority above no longer holds, the paper offers the novelty of mechanisms that ignore client preferences for fairness reasons.

[36] It remains the case that Unrestricted Choice arises for income high enough. Unrestricted Choice now requires two conditions: that type $\tau = 1, \phi = \underline{\phi}$ chooses $t = 1$ and $\tau = 0, \phi = \underline{\phi}$ chooses $t = 0$. The first condition requires that $\underline{\mu}_{11} - \frac{\overline{m}}{\underline{m}} p_1 \geq \underline{\mu}_{10} - \frac{\overline{m}}{\underline{m}} p_0$. The second requires that $\underline{\mu}_{00} - \frac{\overline{m}}{\underline{m}} p_0 \geq \underline{\mu}_{01} - \frac{\overline{m}}{\underline{m}} p_1$. Then let $p_0 = 0$ and $p_1 = \underline{\mu}_{11} - \underline{\mu}_{10}$. Then the first condition always holds. The second condition holds if $\frac{\overline{m}}{\underline{m}} \geq \frac{\underline{\mu}_{01} - \underline{\mu}_{00}}{\underline{\mu}_{11} - \underline{\mu}_{10}}$. As $\underline{\mu}_{01} - \underline{\mu}_{00} \leq \underline{\mu}_{11} - \underline{\mu}_{10}$, $\frac{\underline{\mu}_{01} - \underline{\mu}_{00}}{\underline{\mu}_{11} - \underline{\mu}_{10}} < 1$, and there there always exists an income level sufficiently high to allow Unrestricted Choice.

price of $t = 1$ is $p_1 = min_i(\underline{\mu}_{i1} - \underline{\mu}_{i0}) - p_0$. Deferred Acceptance does not reduce revenue if

$$min_i\underline{\mu}_{i1} - min_i\underline{\mu}_{i0} = min_i(\underline{\mu}_{i1} - \underline{\mu}_{i0}). \tag{22}$$

This arises if (i) $\underline{\mu}_{ij}$ is non-decreasing in j for all i, j, and (ii) $\underline{\mu}_{i1} - \underline{\mu}_{i0}$ is non-decreasing in j. This fails with (1). Deferred Acceptance does not reduce client revenue if and only if type (τ) and services (t) are weak complements for the $\underline{\phi}$ client.

This characterization allows us to identify settings in which client input is unlikely to reduce revenue. Consider a setting where a child receives tutoring but where t measures intensity: $t = 0$ is one hour per week and $t = 1$ is two hours a week. Some children value tutoring more than others. Then if the return to tutoring is linear in hours, the payoff would look something like the matrix below:

$$\begin{array}{cc} \tau = & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0(1-hour) \\ 1(2-hours) \end{array} & \begin{bmatrix} \mathbf{1} & \mathbf{3} \\ \mathbf{2} & \mathbf{6} \end{bmatrix} \end{array}$$

Because t_i and τ_i are complementary here, $min_i\underline{\mu}_{i1} - min_i\underline{\mu}_{i0} = min_i(\underline{\mu}_{i1} - \underline{\mu}_{i0}) = 1$ and there is no revenue loss from Deferred Acceptance over the other Assigned Consumption mechanisms. Note however that if all agents have such preferences - including the $\overline{\phi}$ agents - there is no value to Deferred Acceptance as all agents prefer $t = 1$ and Bureaucratic Assignment and Deferred Acceptance are identical.

So far, we have considered a firm that values revenue as it allows more clients to be covered. Consider the case where $p^A \geq w^A$, so the firm does not need donors. Then the level of client coverage is $I^* = \frac{v+T^A}{\delta}$. If (8) holds, the firm uses Unrestricted Choice. If (8) does not hold but $p^{da} \geq \overline{U}$, the firm uses Deferred Acceptance. If(8) does not hold, $p^{da} < \overline{U}$ and $p^{ac} \geq w^a$, then the firm uses either Agent Assignment with full coverage of $I^* = \frac{v+T^A}{\delta}$ or uses the preferred outcome of the other mechanisms above with lower coverage.

Donors: Donations are exogenously given by D. Donors are typically poorly informed of the internal workings of firms. More concretely, the literature on the responsiveness of whether donations are responsive to firm efficiency show at best a weak relationship (Buchheit and Parsons, 2006, Parsons, 2007, and Yoruk, 2013). However, donors at least likely know the kind of clients who will benefit, so D may be indexed by $\hat{y}$. All results continue to hold if $D(\hat{y}) + p^A$ is decreasing in $\hat{y}$. This simply means that donations only partially make up for lower revenue from clients. There is considerable evidence that donations in charitable settings partially offset funds raised elsewhere (Andreoni, 1993, Andreoni and Payne, 2009, Bolton and Katok, 1998, and Robert, 1984).

4.1 Relaxing the Fairness Constraint

The Fairness Constraint has been used to limit pricing. Here we relax it in two directions.

4.1.1 Excluding Low Income Clients

The Fairness Constraint does not allow the firm to sacrifice a high priority client to gain more lower priority clients. Some social organizations have no choice but to exclude some low income clients.[37] To address this, consider a relaxation of the Fairness Constraint where the firm chooses which clients to serve, but it but treats all those served fairly.[38] There are two purposes here. First, to show how our results continue to hold. Second, to show how varying δ allows us to identify two opposite outcomes: where the firm serves all clients as above, or where it maximizes revenue.

Assume a distribution over m_i, where a fraction $G(m)$ of clients have marginal utility below m, where $G(\underline{m}) = 1$, $G(\overline{m}) = 0$ and $G' = g > 0$. Let the firm now choose the fraction of clients that receive no service $\Gamma \geq 0$. We retain the assumption that conditional on $I_i = 1$, t_i is independent of m_i.

Previously the firm normalized prices by $\underline{m}$. Let it now do so by $M \leq \underline{m}$. First consider Unrestricted Choice. With this adaptation, maximum prices in Proposition 4 are now $p_0 = \frac{\underline{\phi}}{M}$ and $p_1 = \frac{\underline{\phi}+\mu}{M}$.[39] The fraction of clients not served is $\Gamma = (1-h)(1-G(M)) + h(\frac{(1-G(\tilde{m}_0))+(1-G(\tilde{m}_1))}{2})$, where $\tilde{m}_0 = \frac{M\overline{\phi}}{\underline{\phi}}$ and $\tilde{m}_1 = \frac{M(\overline{\phi}+\mu)}{\underline{\phi}+\mu}$.[40] This implies that the support of marginal utilities is $[\overline{m}, \tilde{m}_0]$. Revenue is now $\tilde{p}^{uc} = \frac{\frac{\mu}{2}+\underline{\phi}}{M}$. The expected number of clients before one is served is $\frac{1}{1-\Gamma}$, so the discount factor changes from δ to $\tilde{\delta} = \frac{\delta}{1-\Gamma}$. Surplus is

$$\tilde{S}^{uc} = \frac{D}{\overline{w} - \tilde{p}^{uc}} \left(v + T^* - \frac{\tilde{\delta} D}{2(\overline{w} - \tilde{p}^{uc})} \right). \tag{23}$$

Compared to (16), prices are higher but the discount factor increases. The same logic arises with the other assignment mechanisms. For example, Deferred Acceptance prices are now $\tilde{P}^{da} = \{0, \frac{\mu-\underline{\phi}}{M}\}$ where here $1-\Gamma = \frac{1}{2} + \frac{(1-h)(1-G(M))+h(1-G(\tilde{m}_{da}))}{2}$, where

[37] As an example, private housing associations have a minimum rent.

[38] A familiar example from our own profession is where admission rates to elite universities are higher for those who pay full tuition, but access to classes is equal across all students.

[39] We also need to extend the equilibrium strategy of the client to allow $\hat{t} = \emptyset$ where the client choose no service.

[40] All covered clients with $m_i \leq M$ and $\phi = \underline{\phi}$ choose $\hat{t} = 1$ and all covered clients with $m_i \leq \tilde{m}_j$ and type $\phi = \overline{\phi}$ choose $\hat{t} = 1$, where $\tilde{m}_j = \frac{M(\overline{\phi}+j\mu)}{\underline{\phi}+j\mu}$. The necessary condition for Unrestricted Choice is now $\frac{\overline{m}}{\tilde{m}_0} \geq \frac{\mu-\underline{\phi}}{\mu+\underline{\phi}}$.

$\tilde{m}_{da} = \frac{M\overline{\phi}}{\underline{\phi}-\mu}$.[41]

The firm chooses M. For Unconstrained Choice, the optimal choice M^* is defined by

$$2(\frac{\mu}{2}+\underline{\phi})(v+T^*) = \delta D \frac{(\frac{\mu}{2}+\underline{\phi})(1-\Gamma(M^*)) - M^{*2}(\overline{w}-\tilde{p}^{uc}(M^*))\Gamma'(M^*)}{(\overline{w}-\tilde{p}^{uc}(M^*))(1-\Gamma(M^*))^2}. \quad (24)$$

where $\Gamma'(M) = -(1-h)g(M) - h(\frac{g(\tilde{m}_0)\frac{\overline{\phi}}{\underline{\phi}}+g(\tilde{m}_1)\frac{\overline{\phi}+\mu}{\underline{\phi}+\mu}}{2}) < 0$. The left hand side is the value of additional revenue on clients covered. The right hand side is the cost of allocating treatments to lower priority clients. This allows us to identify the two limiting cases. First, as $\delta \to \infty$, the solution implies that no clients are skipped in the priority list. As the value of funding converges to zero here, the firm would use only Unconstrained Choice or Deferred Acceptance. This is the analog to the market design literature outcome which prioritizes incorporating client preferences in allocations. The other limiting case is where $\delta \to 0$. Then the solution entails $\Gamma \to 1$, in which case the firm sells only to the highest income clients, those with marginal utility $\overline{m}$. This maximizes revenue but at the cost of skipping almost all clients.

4.1.2 When Wealthy Clients Get Better Service

The firm's revenue is constrained by the requirement that the poorest clients must be able to afford the most desired service $t = 1$. Consider a modified Fairness Constraint where I_i is independent of m_i but t_i can depend on m_i. The central question in the theoretical literature reviewed above in identifying when market mechanisms - with no rationing - arises. Consider the same question here.

The firm can now use Unrestricted Choice for any income level. Market clearing prices when (8) fails are given by $p_0^* = 0$ and p_1^* defined by $1 = hG(m_{00}) + (1-h)G(m_{11}) + (1-h)G(m_{01}) + hG(m_{10})$, where $m_{00} = \frac{\mu-\overline{\phi}}{p_1^*}$, $m_{10} = \frac{\mu+\overline{\phi}}{p_1^*}$, $m_{01} = \frac{\mu-\underline{\phi}}{p_1^*}$ and $m_{11} = \frac{\mu+\underline{\phi}}{p_1^*}$. $p^* = 0$ as all clients must be served and some clients of type $\tau = 0$ receive $t = 0$.This raises more revenue than Assigned Consumption. Depending on q, it can either improve or harm allocative efficiency, as choice here partially screens on τ.

The problem is that it is unfair. The m_{0j} terms above determine the fraction of type $\tau = 0$ clients who receive $t = 1$ and the m_{1j} terms determine the fraction of type $\tau = 0$ clients who receive $t = 1$. The Fairness Constraint required that $G(m_{0j}) = 0$ and $G(m_{1j}) = 1$. Here the ratio of "undeserving" to "deserving" clients receiving $t = 1$ is $\frac{hG(m_{00})+(1-h)G(m_{01})}{hG(m_{10})+(1-h)G(m_{11})}$. Previously we treated the cost of this constraint as infinite at $\frac{hG(m_{00})+(1-h)G(m_{01})}{hG(m_{10})+(1-h)G(m_{11})} = 0$. A relaxation would be to assume that the cost to the firm is increasing in this measure.

[41] There is one change to outcomes. The definitions of the m_{ij} imply that the distributions are no longer balanced. However we show below that our results extend to unbalanced demand and supply.

The same capacity to trade revenue for unfairness holds for Deferred Acceptance.[42] The client plays no role in both Agent Assignment and Bureaucratic Assignment, so this makes no difference to outcomes there. Then relaxing the Fairness Constraint from a lexicographic one to a setting where fairness costs are increasing in $\frac{hG(m_{00})+(1-h)G(m_{01})}{hG(m_{10})+(1-h)G(m_{11})}$ offers more latitude to both Choice and Deferred Acceptance. However, the qualitative tradeoffs remain once the desire to avoid inequality remains. In this sense, it is similar to the outcome in Akbarpour et al., 2024, but with less ability to redistribute.

4.2 Other Forms of Choice

When $I_i \in \{0,1\}$ and $\hat{y} < \hat{y}^{uc}$, the firm uses Assigned Consumption. All the mechanisms above are strategy proof, and arise when services are indivisible. We now consider two additional choice outcomes when one of these fails:

- *Restricted Choice*: The client chooses either treatment at prices p_i and receives I_i of her highest ranked preference, where $I_i < 1$ for at least one i.
- *Partial Choice:* The outcome is partially chosen and partially assigned. It is implemented by a First Choice mechanism.

4.2.1 Restricted Choice

As an initial thought experiment, consider a case where the firm can degrade services, where instead of offering $I_j = 1$, it can choose a worse version $I_j = \kappa_j \in [0,1]$ at no cost. Client utility becomes $U_i^c = (T(\tau_i, t_j) - m_i p_j)\kappa_j$. Consider the κ_j that are consistent with Choice for any given income level, where $I_0 = 1$, $I_1 = \kappa_1 \leq 1$. We show in the Appendix that with per unit prices P^*, Choice satisfies the Fairness Constraint for any $\kappa_1 \leq \kappa_1^*$ where

$$\kappa_1^* = \frac{\underline{\phi}}{\mu - \frac{\overline{m}\underline{\phi}}{\underline{m}-\overline{m}}}. \tag{25}$$

This allows choice for any income level. However, the distortion $1 - \kappa_1^*$ is higher at lower levels: it is decreasing in $\frac{\overline{m}}{\underline{m}}$ and is 0 at $\hat{y} = \hat{y}^{uc}$. Expected treatment surplus is $\tilde{T}^{rc} = \frac{\hat{\phi}+\kappa_1^*(\hat{y})(\hat{\phi}+\mu)}{2}$. This shows how the firm can implement $\tau_i = t_i$ but at the cost of distorting services.

[42] There is a range of prices under which both revenue and allocative efficiency rise, but at the cost of unfairness. To see this, consider the firm increasing the price from the price above $p_1^{da} = \frac{\mu-\underline{\phi}}{\underline{m}}$ to a price $\tilde{p}$ where $\frac{\mu-\underline{\phi}}{\underline{m}} < \tilde{p} \leq \frac{\mu}{\underline{m}}$. Then revenue rises, but so also does allocative efficiency as the only clients who no longer report $\hat{t} = \{1 \succ 0\}$ are clients of type $\tau = 0$. As a result, pricing above p_1^{da} improves both revenue and allocative efficiency. The problem is that the outcome is unfair.

Now consider a setting where instead of degrading services, they are divisible $I_i \in [0,1]$. The client then chooses between *quantities* $I_0 = 1$, and $I_1 = \kappa_1^*$. (25) still applies. The only difference is that there is now residual supply of service $t = 1$ beyond client $I^{rc} = \frac{D}{\overline{w} - p^{rc}}$. After client $i = I^{rc}$ there is no more remaining supply of $t = 0$, but there is $I^{rc}\frac{1-\kappa_1}{2}$ of $t = 1$. The firm allocates $t = 1$ to all clients from I^{rc} to $I^{rc}(1 + \frac{1-\kappa_1}{2})$.

Proposition 6 *Let $\hat{y} < y^{uc}$ and $I^{rc} = \frac{D}{\overline{w} - p^{rc}}$. Assume that $P = P^*$ for all clients $i \leq I^{rc}$. Surplus from Restricted Choice is*

$$S^{rc} = I^{rc}\left(v + \mu + \hat{\phi}(1 - \frac{1-\kappa_1^*}{2}) + \frac{1-\kappa_1^*}{2}\frac{\hat{\phi}}{2}\right) - \delta\frac{I^{rc2}}{2}\left((1 - \frac{1-\kappa_1^*}{2}) + (1-\kappa_1^*)(1 + \frac{1-\kappa_1^*}{4})\right). \quad (26)$$

where $p^{rc} = (1 - \frac{1-\kappa_1^*}{2})p^{uc} + \frac{1-\kappa_1^*}{2}p^{ac}$.

The outcome has three notable components. First, clients $i \leq I^{rc}$ choose efficiently $t_i = \tau_i$ and pay unit prices P^*. Second, all covered clients $i > I^{rc}$ receive $t = 1$ so expected surplus is lower. Revenue is a weighted combination of that from Unrestricted Choice and random assignment. A new issue is *client targeting*. This is given by the second bracketed term in (26) and reflects the fact that the most desired service ($t = 1$) is offered to low priority clients. This is what we mean below by poor targeting of clients.[43]

Poor targeting leads to a more ambiguous ranking of this choice outcome relative to Assigned Consumption. When Unrestricted Choice satisfies fairness. it is preferred to Deferred Acceptance or Bureaucratic Assignment. This is not the case for Restricted Choice. It is both intuitive and straightforward to show that for δ low enough, the firm may prefer either of these Assigned Consumption mechanisms to Restricted Choice as it ensures complete coverage of the highest priority clients.

Artificial Currency: Restricted Choice screens on marginal rates of substitution. Given this, it should not be surprising that there is a relationship to competitive equilibrium with equal incomes (Foley, 1967). We show this in the Appendix, where the firm offers supply of $t_0 = \frac{I^{rc}}{2}$ and $t_1 = \frac{\kappa_1^* I^{rc}}{2}$ and gives all clients $i = 1, 2, ...I^{rc}$ one unit of artificial currency to bid on services. The outcome above is an equilibrium where each client chooses between $I_0 = 1$ and $I_1 = \kappa_1^*$. Competitive equilibrium with

[43]Proposition 6 assumes that the firm charges P^*. The issue with fairness is to induce clients with more income to choose $t = 0$ over the other services. There are two ways to do this: reducing quantities (κ_i) and reducing p_0. This implies that there is one other possible equilibrium outcome, where $p_0 = 0$. This is done to increase κ_i above the levels in (25). This does not change the tradeoffs, and is described in the Appendix.

artificial currency is the most common market design alternative to matching without prices (Budish, 2011, Budish and Cantillon, 2012, Prendergast, 2017, 2022, Goedde, 2024, and Wong, 2024).

In many of the applications where artificial currency has been used, it arises as outcomes would be unfair if fiat money were used instead. Such inequalities from using fiat money underlie much of the analysis here. This section offers a link between two existing mechanisms used. Yet the link is continuous in client income, where lower client income limits fiat money payments, resulting in more distorted service assignment. Furthermore, it shows how the outcome converges to Unrestricted Choice as $\hat{y} \to \hat{y}^{uc}$.[44]

We give two examples of the client targeting problem. First, Feeding America allocates food using an artificial currency, where food banks choose between higher quantities of a less desired good and lower quantities of a more desired one. The outcome with bidding with artificial currency is a Restricted Choice outcome, with prices (the analog to $\frac{1}{\kappa_i^*}$) given in Figure 2, taken from Prendergast, 2022. For instance, for every pound of produce ($I_i = 1$), a food bank can get $\frac{1}{7}$ pounds of processed fruit or $\frac{1}{50}$ pounds of cereal. Before introducing this system, the firm assigned the analog to $I_i = 1$ randomly. How does the Restricted Choice mechanism reallocate food across the national network? Feeding America had a desire that the amount of food would be reoriented towards smaller, poorer food banks.[45] The distributional impact of the Feeding America design reinforced its allocative benefits. Small food banks increased their food consumption by 72% compared to the ex ante allocation, at very little cost to nutritional value (Prendergast, 2022). In the language of the model, targeting improved though the move to Restricted Choice.

A different outcome arose for vaccine distribution during the Covid-19 pandemic. The international agency Covax allocated over a billion vaccines to 125 countries after early 2021. Here $S = \sum_{i=1}^{N}(v_i + T(\tau_i, t_i))I_i$ represents different characteristics of vaccines t_i to be allocated to the countries (i) needing those attributes most τ_i, where v_i represents Covax view of country priority, and I_i is coverage of vaccines in country i. Covax chose to allocate vaccines in a manner that largely ignored vaccine attributes or priority. Instead, it offered all vaccines to all countries in a lock-step fashion (Budish et al., 2022).

A problem that arose was that many vaccines were close to their expiry date. In the language of the model, $t = 0$ vaccines were close to expiry while $t = 1$ vaccines

[44]Restricted Choice arises above when services are divisible. We do not allow clients to purchase lotteries. If we did, an alternative would allow the κ_i^* to be probabilities of receiving the service, whether divisible or not. Screening can then arise through single crossing in probabilities. This approach would mimic Hylland and Zeckhauser, 1979. In reality most social organizations simply will not fail to offer service to its most needy clients *ex post*, even if it has appealing *ex ante* properties.

[45]The democratic nature of Feeding America restricted them beforehand to treat all areas equally (conditional on statistics measures of need), but Feeding America's v_i would have focused more on smaller food banks.

were far from their expiry date. Countries varied in their capacity to quickly distribute vaccines. As an example, Turkey's health care system ($\tau = 0$) was far more capable of responding than that of the typical West African nation ($\tau = 1$) . Efficiency requires distributing more "close to expiry" vaccines to Turkey than to West Africa. The allocation system did not use expiry date in the allocation mechanism. This led to remarkable waste, with estimates of almost 40% of all vaccines sent to sub-Saharan Africa not being used.[46]

Covax did not use expiry date because of a political issue deriving from countries not wanting an allocation mechanism that disproportionately gave them soon-to-expire vaccines. However, Restricted Choice is designed to solve such problems by offering countries a choice between a larger number of soon-to-expire vaccines and a smaller number of vaccines far from expiry. With an appropriately chosen rate of exchange (the κ_i^* of the model above), both sets of countries benefit. Covax did not pursue this. A likely concern was the perception that Covax was failing to focus on the most-needy countries, by giving higher income members *more* vaccines than the lower income countries of sub-Saharan Africa. Once again, this example focuses on how Restricted Choice mechanisms rest partially on how they affect the perception of targeting.[47]

4.3 Partial Choice and More Services

So far, we have assumed two services, $t_i \in \{0, 1\}$. This allows to describe all mechanisms above except Partial Choice. Consider a typical housing allocation problem, when a client is matched to a housing unit. The unit may not be the client's most desired type or in her preferred location. In the Assigned Consumption setting above, she has no choice but to take it. In reality, in many settings if she turns the offer down, she is not excluded from consuming a house. Instead, she is placed lower on waiting list going forward. As a result, the client needs to strategize about whether to accept. We model Partial Choice as arising when some services are chosen and some are assigned. This cannot happen when there are two treatments, because by definition if one is chosen, the remaining one must be the alternative.

To address this, now assume that the firm provides n treatments $j \in \{0, 1, ..., n-1\}$ to clients $i \in \{0, 1, ..., n-1\}$, both uniform and balanced, with returns given by (1). Further assume that t_i^a is correct with probability q, and equal to $t \neq t_i$ with probability $\frac{1-q}{n-1}$.[48] The analog to Assumption 1 is now given by:

[46] This led to well-publicized complaints from the African Union that "the majority of the donations have been ad hoc, provided with little notice and short shelf lives" (African Union, 2021).

[47] There is an important difference between both of these applications and the model above. The model assumes that the firm allocates services in the order of *its* priority ranking. Political considerations implied that both Feeding America and Covax had to offer equal access. Yet Feeding America's priorities were to reallocate food to poorer food banks, and Covax wished to vaccinate the poorest countries.

[48] Here the analogous condition for ρ is $\rho \geq \frac{1}{n}$.

Assumption 1a: $\overline{\phi} > n\mu$ and $\underline{\phi} < \mu$.

The analog to Lemma 1 arises:

Lemma 1a:

- *Unrestricted Choice satisfies the Fairness Constraint if and only* $\frac{\overline{m}}{\underline{m}} \geq \frac{(n-1)\mu - \underline{\phi}}{(n-1)\mu + \underline{\phi}}$. *When* $\frac{\overline{m}}{\underline{m}} \geq 1 - \frac{\underline{\phi}}{(n-1)\mu}$, *maximum prices are* $p_j^{uc} = \frac{j\mu + \underline{\phi}}{\underline{m}}$, with revenue per client $\frac{(n-1)\mu}{2} + \underline{\phi}$.
- *Deferred Acceptance can be implemented for any income level. For* $\frac{\overline{m}}{\underline{m}} < \frac{n\mu - \underline{\phi}}{n\mu + \underline{\phi}}$, *maximum revenue is* $p_j^{da} = \frac{j(\mu - \underline{\phi})}{\underline{m}}$ with revenue per client $\frac{(n-1)\mu}{2}$.
- *Bureaucratic Assignment and Agent Assignment can be implemented for any income level. Maximum revenue is* $p_j^{ac} = \frac{j\mu}{\underline{m}}$ with revenue $\frac{(n-1)(\mu - \underline{\phi})}{2}$.

This shows how the tradeoffs above generalize to more treatments.[49] It adds one potential additional optimal mechanism, Partial Choice.

Partial Choice: Let $n = 3$. Furthermore, let $\hat{y} < \hat{y}^{uc}$ and $I_i \in \{0, 1\}$, so neither Unconstrained nor Restricted Choice can be used. The novel mechanism allows the client to choose one t, where the alternative is an assigned mechanism over the other two services: Partial Choice. In our example above, assume that in addition to Chicago ($t = 0$) and New York ($t = 1$), we add Boston ($t = 2$).[50] Purely for simplicity, we assume that the assignment component of the mechanism is Bureaucratic Assignment. Partial Choice can then be implemented by a First Choice mechanism, which we now describe:

1. The firm chooses $P^{1st} = \{p_0^{1st}, p_1^{1st}, p_2^{1st}\}$.
2. Each client $i \leq I^{1st}$ reports $\hat{t}_i$.
3. The firm allocates $t = i$ to all clients k for which $\hat{t}^k = i \succ ...$ if i is undersubscribed.
4. For all $t = i'$ that are oversubscribed, the firm implements Bureaucratic Assignment among those where $\hat{t}^k = i' \succ ...$

[49] We also show in the Appendix that the Deferred Acceptance outcome has the same structure as above. Allocative efficiency retains the ranking above, so the tradeoff between the mechanisms retains the features of the two state example above.

[50] In any of the assignment mechanisms that allowed client input, the client always received Chicago ($t = 0$) in equilibrium if it was ranked first.

In the Appendix, we show that there is only one equilibrium of this form, where any client can choose $t = 1$ (New York) with the other two assigned. A Partial Choice outcome arises when only type $\tau = 1$ does so. Proposition 7 shows the outcome.

Proposition 7 *A Partial Choice equilibrium exists with* $T^{1st} > T^b$ *and* $p^{1st} > p^{ac}$ *if* $\hat{y} \geq \hat{y}^{1st}$*, where* $\hat{y}^{1st} < \hat{y}^{uc}$*.*

Partial Choice reduces assignment error from Assigned Consumption, as all $\tau = 1$ clients choose efficiently. The novelty here is not that non-strategy proof First Choice mechanisms can have higher allocative efficiency.[51] Instead, it can raise more funding than Bureaucratic Assignment. In the Appendix, we show that this arises if $\frac{\overline{m}}{\underline{m}} > 1 - \frac{(3q+1)\underline{\phi}}{(3q-1)2\mu}$ (when $q = 1$ this is identical to $\tilde{y}^{uc}$). Second, a Partial Choice equilibrium exists for lower client income than is needed for Unconstrained Choice.[52] For $q < q^* = \frac{2\mu+\underline{\phi}}{3(2\mu-\underline{\phi})}$, the Partial Choice outcome exists for any income level.

This argues that an appeal of non-strategy proof mechanisms is funding. A case familiar to readers may be students applying to universities via Early Decision. If the student is accepted by that school, she is obliged to accept the offer.[53] While this is not Partial *Choice* in the sense above, probabilities of acceptance are often four times higher than though the regular process. The use of ED has increased significantly over the last decade (Murphy, 2022).[54] The conceptual novelty here is that the firm uses Partial Choice to generate more funding than Assigned Consumption. It is well known that Early Decision students pay higher average tuition (Antecol and Smith, 2012, Murphy, 2023). Also consistent with the model is White, 2024, who provides data that colleges with lower endowments (the analog to D in the model) fill a larger fraction of their class through Early Decision.

5 Extensions and Observations

Here we discuss some modeling choices and examples.

Alignment of Interests: One-sided matching is typically defined as settings where one side (the firm here) is indifferent to assignments (services here). This yields Serial

[51] See Abdulkadiroğlu et al., 2011, Agarwal and Somaini, 2018, Calsamiglia et al., 2019, and Kapor et al., 2020, for differing views on the merits of this over Deferred Acceptance.

[52] The reason is simple: with Unconstrained Choice, the wealthier client has to be deterred from choosing $t = 2$ for sure when $\tau \neq 2$. This is not the case here - if the client does not choose $t = 1$, the alternative is random assignment, which is less valuable than the preferred one for sure.

[53] This is not a legally binding commitment, but is treated as binding by most students. They can also apply through Regular Decision, where the student is not committed to a school if accepted.

[54] Avery and Levin, 2011, show that it can be used to screen students on their enthusiasm for a school in a conceptually similar way to that described above.

Dictatorship, while two sided matching yields outcomes like Deferred Acceptance.[55] Our setting by that definition is two-sided, yet the outcome in many cases is the analog of Serial Dictatorship, namely Choice. Here the source of its use is not indifference but alignment of interests between the firm and client. When average income is high enough though (8), there is complete alignment through choice of prices, while when it lies below that the only possibility is partial alignment via Deferred Acceptance, but at the cost of lower prices and revenue. For low enough donations, there is no alignment.

Other Social Organizations: We have restricted attention to the non-profit organization by assuming that the funding-service provision link is closed to reflect the Non-Distribution Constraint. There are other settings where the tradeoff between funding and service arises, namely for-profit social enterprises and the public sector entities. First, there are estimated to be ten million social enterprises (World Economic Forum, 2024). They also face a tradeoff where funding limits their capacity for client coverage. However, as they also have a profit motive, the incidence of funding changes between service provision and profit would determine outcomes. Similarly, many public sector entities are required to raise funds through service charges in a setting where there are equity concerns. Although the insights again have relevance, we have ignored them here as the public sector allows other sources of revenue such as tax revenue to offset revenue shortfalls. As a result, we see the relevance of our insights extending beyond the non-profit sector.

Unbalanced Demand and Supply: The assumption of balanced supply and demand eliminates the need for rationing other than with Deferred Acceptance. It also means that with Unrestricted Choice, all clients receive their preferred service. The qualitative results extend to unbalanced demand and supply. In the Appendix, we show this with an excess of $t = 1$ slots relative to $t = 0$. The analysis is identical for the other case. Assume that the distribution of types $\{\tau_0, \tau_1\}$ is not uniform $(\frac{1}{2}, \frac{1}{2})$ but instead $\{\frac{1-\zeta}{2}, \frac{1+\zeta}{2}\}$ for $\zeta > 0$. The qualitative outcomes remain unchanged in the sense that (i) Unrestricted Choice is feasible if (8) holds, (ii) the revenue from Deferred Acceptance is p^{da} and the revenue from Bureaucratic and Agent Assignment is p^b, and (ii) there is still a wage reduction from Agent Assignment. This leads to the same qualitative outcomes. The only substantive differences are (i) Unrestricted Choice involves rationing as there is excess demand for $t = 1$ at the price of $\frac{\mu+\phi}{\underline{m}}$ (ii), $p_0 = 0$ under Unrestricted Choice as some $\tau = 1$ clients are mismatched, (iii) surplus is lower as the first best entails $\frac{\zeta}{2}$ type $\tau = 1$ clients being assigned to $t = 0$, and (iv) Unrestricted Choice becomes analogous to the familiar Random Serial Dictator

[55] The balance of demand and supply implies that there is no need for the common outcome of Random Serial Dictatorship in allocating service. Serial Dictatorship may also be preferred due to its ex ante combination of fairness and efficiency.

mechanism as not all clients demanding $\tau = 0$ can receive it.

Unbalanced supply allows us to address one final assumption above, namely that as the firm expands the distribution between supply of $t = 1$ and $t = 2$ remains balanced. In some settings, the firm chooses to spend its budget on a cheaper $t = 0$ or a more expensive $t = 1$. We have assumed that the outcome of this remains unchanged as the firm grows. However, the firm has one other instrument when it is short of resources, namely, to bias its supply towards $t = 0$ in order to cover more clients. When we add this extension, the firm can use both the assignment mechanisms - Agent Assignment and Bureaucratic Assignment - and excess supply of the cheaper service to cover more clients, but the conceptual tradeoffs remain.

Where m_i and τ_i are not Independent: Here we consider the case where say poorer clients are more likely to require more intensive services, $t = 1$. To model this, let a client with type τ_i have marginal utility distributed over a support $[\overline{m}_i, \underline{m}_i]$. Then prices for Bureaucratic and Agent Assignment are $p_0 = 0, p_1 = \frac{\mu}{max\{\underline{m}_0, \underline{m}_1\}}$, prices under Deferred Acceptance are $p_0 = 0, p_1 = \frac{\mu - \underline{\phi}}{max\{\underline{m}_0, \underline{m}_1\}}$, and the analog to (8) is $\frac{\overline{m}_0}{\underline{m}_1} \geq \frac{\mu - \underline{\phi}}{\mu + \underline{\phi}}$. This changes qualitative outcomes only if $\frac{\overline{m}_0}{\underline{m}_1} > 1$ for all income levels, in which case Unrestricted Choice is always possible. This is because the type $\tau = 0$ client is always too poor to mimic the $\tau = 1$ allocation as the utility cost of the price to be paid $\frac{\mu + \underline{\phi}}{\underline{m}_1}$ is too high. Otherwise the results remain unchanged.

Now consider cases where v_i and y_i are correlated. For instance, the firm may have objectives to prioritize the poorest clients. We have not addressed this because the level of funding changes the average income of client covered. This link is mechanical and simply means that the analysis would need to condition average income on I, but does not change the analysis.[56]

Interpreting the Client's Preferences: We have interpreted the client's preferences as $V(y) + T$, where y purchases a composite outside good. The central feature that causes the Fairness Constraint to fail is that higher income consumers are willing to pay more than those with lower incomes. While this may seem intuitive, it is not the outcome of textbook microeconomic choice. Consider textbook consumer choice. Client i consumes two goods z_1 and z_2 with income y_i, where the goods are

[56] We have assumed that the firm can only use a single mechanism for all clients. This is for the sake of realism. However, it maybe be desirable to use a distribution of mechanisms. Let z^j be the fraction of clients for which mechanism $j \in \{a, da, 1st, uc, rc\}$ is used, where $\sum_j z^j = 1$. Then the surplus obtained from the non-profit is $S^* = \frac{D}{w^A - \sum_j z^j p^j} \left(v + \sum_j z^j T^j - \frac{\delta D}{2(w^A - \sum_j z^j p^j)} \right)$. The reason for combinations is due to the $\frac{\delta D}{2(w^A - \sum_j z^j r^j)}$ term. As an example, imagine that w^* is far below $\underline{U}$. Then Agent Assignment has far higher client coverage than say Deferred Acceptance. But as there are diminishing returns, at some point the additional client coverage is of little value. As a result, the firm would prefer at some point to transfer a client from DA to Agent Assignment, and using a combination of mechanisms becomes optimal.

weak substitutes and have diminishing returns. The client faces market prices $\overline{p}_{z1}$ and $\overline{p}_{z2}$. In the absence of the firm, the client choose z_1^* and z_2^* to maximize $U(z_1^*, z_2^*)$ and $\overline{p}_{z1}z_1 + \overline{p}_{z2}z_2 \leq y_i$. If the solution involves $z_i^* > 0$ for both i, the solution is characterized by the familiar first order condition

$$\frac{U_{z1}}{\overline{p}_{z1}} = \frac{U_{z2}}{\overline{p}_{z2}} \tag{27}$$

and $\overline{p}_{z1}z_1^* + \overline{p}_{z2}z_2^* = y_i$.

Now introduce our non-profit, and assume it gives covered clients 1 unit of z_2. Then the utility of the covered client becomes $U(z_1^{**}, z_2^{**} + 1)$, where z_i^{**} is the equilibrium choice given the firm's presence. To understand willingness to pay for the firm's contribution to z_2, there are three cases:

1. When $z_2^* \geq 1$ for all clients. Here $\frac{U_{z1}(z_1^{**}, z_2^{**})}{\overline{p}_{z1}} = \frac{U_{z2}(z_1^{**}, z_2^{**})}{\overline{p}_{z2}}$ and there is no difference in willingness to trade off good z_1 for good z_2 for richer and poorer clients. Giving the client a service at price p_j above (where $p_j \equiv p_{z2}$) is equivalent to increasing the client's income by $\overline{p}_{z2} - p_j$. If this is the case, then the solution is (27) subject to the client's income being $y_i + \overline{p}_{z2} - p_j$.

2. When $z_2^* < 1$ for all clients. This is when the client consumes less of z_2 without the firm. A leading case is where $z_2^* = 0$, which we describe here. Many goods provided non-profits are not chosen by clients in their absence. Then the price a client is willing to pay for the firm's service p_{z2}^* for a given y_i is given by $U(\frac{y_i - p_{z2}^*}{\overline{p}_{z1}}, 1) = U(\frac{y_i}{\overline{p}_{z1}}, 0)$. There are two reasons why willingness to pay varies by income. First, as all income is spent on z_1, higher income client consume more thereby reducing its marginal utility. Second, many services provided by non-profits are typically consumed by those with higher incomes: quality housing in safe neighborhoods, mental health therapy, tutoring for disadvantaged high school students, etc.. If these are normal goods, then willingness to pay by a high income client will be higher.

3. When $z_2^* > 1$ for wealthier clients, but $z_2^* < 1$ for poorer clients. Consider a case where clients are receiving food from a food pantry. Wealthier clients consume more food, and also purchase on the private market. Poorer clients do not. Then the marginal price paid by the wealthier clients is the market price, while the marginal price for the poorer clients is that provided by the firm. Then even if (27), it does so at different prices, leading to different willingness to pay.

In the second and third case, the willingness to pay by higher income clients is higher, as required above.

Unrestricted Choice and Fairness Tolerance: Unrestricted Choice overcame the Fairness Constraint by allowing a client's average income to change the support

of income. To address robustness, assume now that the income of the wealthiest and poorest client do not change as average income changes. Then under the Fairness Constraint above, higher average income does not make Unrestricted Choice more likely.

However, let there now be a tolerance for unfairness, where the firm is willing to accept that no more than a fraction Γ of clients can be denied coverage. Furthermore, a client's income be generated by a Binomial process where clients have n trials to generate income, with success rate ν. Each success gives income of Y. In this case, the highest earning client has income of $y_i = nY$ and the lowest earning client has $y_i = 0$. Increasing ν is equivalent to increasing average income. Under the strict Fairness Constraint, the firm cannot charge a positive price as some clients have no income.

Assume now that the firm has a relaxed Fairness Constraint with tolerance Γ. Unrestricted Choice arises if the analog of (8) holds for the m^* for which a fraction Γ are excluded from service. The CDF of the binomial is $G(k) = \sum_{i=0}^{k} \frac{n!}{(n-i)!}\nu^i(1-\nu)^{n-i}$. Let m^{**} be the marginal utility for the lowest earning covered client at which Γ are excluded. Note that $G(k)$ is decreasing in ν. As a result, $G(m^{**})$ is decreasing in ν, so that there exists a ν^* such that the firm can use Unrestricted Choice as $G(m^{**}) < \Gamma$. Hence the result that higher average income allows Unrestricted Choice generalizes to this case where the support of income does not change.[57]

6 Conclusion

For the last four decades, the gold standard of non-market allocations has been the market design literature on matching without prices. Yet one worker in ten is employed in a sector that ranks funding as its greatest concern, and raises half its revenue from clients, though does not use market clearing prices. We have proposed a larger set of optimal assignment mechanisms than in the previous literature, reflecting the fact that many social organizations are more concerned with raising funds and retaining and attracting staff. It interprets for instance why a firm would allow a bureaucrat with the "wrong" preferences unfettered control when something as apparently appealing as Deferred Acceptance is available.

One interpretation of this work is "in praise of bureaucracies". Non-profits use rules to allocate resources, through the use of queues, assignment by rule, or outcomes determined by the luck of the draw in terms of which bureaucrat handles your case. Funding issues are the source of such practices here. Unlike standard price theoretic logic, here when clients partially pay they may have less input to outcomes than when they do not pay at all. This is because payment raises the problem of Fairness. In this sense, rules that exclude client input again become part of the assignment matrix.

[57] This result does not rely on DARA. This is because the difference in income between the m^{**} percentile of income and the highest earner is decreasing in ν.

There are a number of key inputs. First, that understanding the social organization requires addressing that funding primarily derives from clients, but that existing pricing models do not relate well to their constraints. Second, that in the absence of allocation mechanisms that reveal client preferences, stakeholders can reasonably disagree. As an example, is it more important for a food bank that everyone be fed a sufficient amount, or that they only be fed nutritious food? Third, that the logic of the agency literature for distorting ex post allocations to improve ex ante incentives also applies to typically desired outcomes like *ex post* Pareto efficiency.

The paper is not intended to explain the actions of all social sector firms. We have assumed that the motives of the firm are "pure", only interested in client welfare. Some may have agency concerns. This assumption has been made to understand the complexity of the firm's tradeoffs even in the absence of agency concerns or profit motives. Most importantly, we do not address cases such as public school choice, where funding occurs through tax revenue. Given this, it is striking that funding concerns for core market design applications are minimal, occurring through taxation (school choice) or where the potential funds received are trivial compared to overall funding for the institutions (medical resident matching). Finally, many social organizations do not used the price system not because it would benefit the wealthy, but because there are externalities. For example, schools care about the composition of their student body. We have offered a view of one slice of the non-market sector, one focused on serving those in need.

For social organizations, pricing fails to fulfill their usual role in many ways. It stops the firm from trusting clients to choose, leading to Assigned Consumption. Efficient prices also identify client preferences, and disagreements about what client truly need are more common in their absence. Finally, prices pay the bills. Much of the paper has focused on how firms respond to these price failures, often by trading off the intensive margin of allocative efficiency for the extensive margin of expanding access to care. In this sense, one contribution of the paper may be to embed our understanding of the social sector in a more price (failure)-theoretic framework.

References

Abdulkadiroglu, Atila, Parag A. Pathak, Jonathan Schellenberg,

and Christopher R. Walters. "Do Parents Value School Effectiveness?", *American Economic Review*, 2020, 110(5): 1502–1539.

Abdulkadiroglu, Atila, Yeon-Koo Che and Yosuke Yasuda, "Resolving Conflicting Preferences in School Choice: The "Boston Mechanism" Reconsidered", *American Economic Review*, 2011, Vol. 101, No. 1, 399-410.

Adler P. S., and B. Borys, "Two types of bureaucracy: Enabling and coercive", *Administrative Science Quarterly*, 1996, 41(1), 61-89.

African Union, Africa CDC, et al., "Joint Statement on Dose Donations of COVID-19 Vaccines to African Countries", 2021. https://www.who.int/news/item/29-11-2021-joint-statement- on-dose-donations-of-covid-19-vaccines-to-african-countries.

Agarwal, Nikhil, and Paulo Somaini, "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism", *Econometrica*, 2018, 86 (2), 391-444.

Aghion, Philippe, and Jean Tirole, "Formal and Real Authority in Organizations", *Journal of Political Economy*, 1997, 105 (1), 1-29.

Andreoni, J., "An Experimental Test of the Public Goods Crowding-Out Hypothesis", *American Economic Review*, 1993, 83, 1317-27.

Andreoni, J. and A. Payne, "Is crowding out due entirely to fundraising? Evidence from a panel of charities", *Journal of Public Economics*, 2009, 95, 334-43.

Antecol, Heater and Janet Smith, "The Early Decision Option in College Admission and Its Impact on Student Diversity", *Journal of Law and Economics*, 2012, vol. 51, no. 1, 217–249.

Avery C., and J. Levin, "Early Admissions at Selective Colleges", *American Economic Review*, 2010, vol. 100, 2125-2156.

Balinski, M. and T. Sonmez, "A Tale of Two Mechanisms: Student Placement", *Journal of Economic Theory*, 1999, 84, 73-94.

Bansak, K., S. Lee, V. Manshadi, R. Niazadeh, and E. Paulson, "Dynamic Matching with Post-allocation Service and its Application to Refugee Resettlement", 2025, working paper.

Besley, Timothy, Stephen Coate, and Glenn Loury, "The Economics of Rotating Savings and Credit Associations", *American Economic Review*, 1993, 83, 792-810.

Bolton, G. and E. Katok, "An Experimental Test of the Crowding Out Hypothesis: The Nature of Beneficent Behavior", *Journal of Economic Behavior and Organization*, 1998, 37, 315-31.

Buchheit, S., and L. Parsons, "An experimental investigation of accounting information's influence on the individual giving process", *Journal of Accounting and Public Policy*, 2006, 25(6):666-686.

Budish, Eric, "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes", *Journal of Political Economy*, 2011, 119 (December): 1061-1103.

Budish, Eric, and and Estelle Cantillon, "The Multi-Unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard", *American Economic Review*, 2012, 102 (August): 2237-71.

Budish, Eric, Hannah Kettler, Scott Duke Kominers, Erik Osland,

Canice Prendergast and Andrew A Torkelson, "Distributing a billion vaccines: COVAX successes, challenges, and opportunities", 2022, *Oxford Review of Economic Policy*, 2022, 38(4), 941-74.

Bureau of Labor Statistics, "Celebrating National Nonprofit Day with BLS Data", August 15, 2023. https://www.bls.gov/blog/2023/celebrating-national-nonprofit-day-with-bls-data.htm.

Bureau of Labor Statistics, "Nonprofits: a look at national trends in establishment size and employment", January, 2024.

Calsamiglia, Caterian, Chao Fu, and Maia Guell, "Structural Estimation of a Model of School Choices: The Boston Mechanism Versus its Alternatives", *Journal of Political Economy*, 2019, vol. 128, issue 2, pages 642-80.

Cantarelli P., Belardinelli P., and Belle N., "A meta-analysis of job satisfaction correlates in the public administration literature", *Review of Public Personnel Administration,* 2016, 36, 115-144.

Carey, M., "It's a bit like being a robot or working in a factory": Does Braverman help explain the experiences of State Social Workers in Britain since 1971?", 2009, *Organization*, 16(4), pp. 505–27.

Che, Yeon-Koo, and Navin Kartik, "Opinions as Incentives", *Journal of Political Economy*, 2009, vol. 117, issue 5, pages 815-60.

Chaumba, J. and A. Locklear, "Strategies for Promoting Client Self-Determination: A Review of the Literature", *Journal of Social Work and Ethics*, 2021, 18(1), 60-71.

Chiappori, P.-A., and Paiella, M., "Relative risk aversion is constant: Evidence from panel data", *Journal of the European Economic Association*, 2011, 9(6), 1021–1052.

Cohen, N., Benish, A., and A. Shamriz-Ilous, "When the Clients Can Choose", *Social Service Review,* 2016, 90(4), 620-646.

Cook, J. , S. Shore, J. Burke-Miller , J. Jonikas, M. Hamilton, B Ruckdeschel,

W. Norris, A. Markowitz, M. Ferrara, and D. Bhaumik., "Mental Health Self-Directed Care Financing: Efficacy in Improving Outcomes and Controlling Costs for Adults With Serious Mental Illness", *Psychiatric Services*, 2019, 70(3), 191-201.

Dustin, D., "Skills and knowledge needed to practise as a care manager: Continuity and change", *Journal of Social Work*, 2006, 6(3), pp. 293–313.

Dworczak, P., "Inequality and Market Design", 2024, forthcoming, *ACM SIGecom Exchanges*.

Food Bank News, "Study Quantifies the Benefits of Client Choice", 2024. foodbanknews.org/study-quantifies-the-benefits-of-client-choice/.

Gai G, Zare H, Eisenberg MD, Polsky D, Anderson GF, "Comparison of Trends in Nonprofit Hospitals' Charity Care Eligibility Policies Between Medicaid Expansion States and Medicaid Nonexpansion States", *Medical Care Research and Review*, 2021, 79(3):458-468.

Goedde, J., "Pricing in Markets without Money: Theory and Evidence from Home Exchanges", 2024, Paris School of Economics.

Groves, J., Ho, W. and Siu, K., "Exploring the sacred-secular dialect in everyday social work practice: An analysis of religious responses to managerialism among outreach social workers in Hong Kong", 2016, *British Journal of Social Work*, 46(5), pp. 1411–28.

Guiso, L., and Paiella, M., "Risk aversion, wealth, and background risk", *Journal of the European Economic Association*, 2008, 108(12), 1109–1150.

Hansmann, Henry, "The Role of Nonprofit Enterprise", *The Yale Law Journal*, 1980, 89(5), 835-901.

Hirsch, B,, D. McPherson, and A. Preston, "Nonprofit Wages: Theory and Evidence", 2017, IZA Discussion Paper.

Huber, T., Jaspersen, J.G., Richter, A., and D. Strumpel , "On the change of risk aversion in wealth: a field experiment in a closed economic system, *Experimental Economics*, 26, 1–26 (2023). https://doi.org/10.1007/s10683-022-09762-x

Hylland, Aanund, and Richard Zeckhauser, "The Efficient Allocation of Individuals to Positions", *Journal of Political Economy*, 1979, Vol. 87, No. 2, 293-314.

Kapor, A., C. Neilson and S. Zimmerman, "Heterogeneous Beliefs and School Choice Mechanisms", *American Economic Review*, 2020, vol. 110, no. 5, 1274–1315.

Langer, Julie, Mary K. Feeney, and Sang Eun Lee, "Employee Fit and Job Satisfaction in Bureaucratic and Entrepreneurial Work Environments," *Review of Public Personnel Administration*, 2019, 39 (1), 135-155.

Leete, L., "Whither the nonprofit wage differential? Estimates from the 1990 Census", *Journal of Labor Economics*, 2001, 19 (1), 136-70.

Lipsky, Michael, *Street Level Bureaucracy: Dilemmas of the Individual in Public Services*, 1980, New York, Russell Sage Foundation.

Liu, Dong, Shu Zhang, T. W. Lee, "The effects of autonomy and empowerment on employee turnover: test of a multilevel model in teams.", *Journal of Applied Psychology*, 2011, 96(6), 1305-15.

MacEachan, E., Kosny, A., Ferrier, S., Lippel, K., Neilson, C.,

Franche, R. L., and Pugliese, D. "The ideal of consumer choice in social services: challenges with implementation in an Ontario injured worker vocational retraining programme.", *Disability and Rehabilitation*, 2013, 35(25), 2171–2179.

Murphy, James, "The Future of Fair Admissions", *Education Reform Now*, 2022.

Murphy, James, "Early Decision Benefits Rich Kids. That Doesn't Necessarily Mean That It's Evil", *Slate*, 2023, December 14.

National Council of Nonprofits, "Downloadable Charts and Figures", 2024. https://www.nonprofitimpactmatters.org/data/downloadable-charts/.

Nichols, Albert L., and Richard J. Zeckhauser, "Targeting Transfers through Restrictions on Recipients", *American Economic Review*, 1982, vol. 72(2), pages 372-377, May.

Pathak, Parag, and and Tayfun Sonmez, "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism", *American Economic Review*, 2008, Vol 98(4), 1636-52.

Parsons, L. M., "The impact of financial information and voluntary disclosures on contributions to not-for-profit organizations", *Behavioral Research in Accounting*, 2007, 19(1):179-196.

Pascoe, Katheryn, Bethany Waterhouse-Bradley, and Tony McGinn , "Social Workers' Experiences of Bureaucracy: A Systematic Synthesis of Qualitative Studies", *The British Journal of Social Work*, 2023, Volume 53, Issue 1, 513–533, https://doi.org/10.1093/bjsw/bcac106 .

Petersen, Anette, Lars Rune Christensen, and Thomas T. Hildebrandt , "The Role of Discretion in the Age of Automation", *Computer Supported Cooperative Work*, 2020, https://doi.org/10.1007/s10606-020-09371-3.

Prendergast, Canice, "The limits of bureaucratic efficiency", *Journal of Political Economy*, 2003, 111, no. 5: 929-958.

Prendergast, Canice, "The motivation and bias of bureaucrats", *American Economic Review*, 2007, 97, no. 1: 180-96.

Prendergast, Canice, "How Food Banks Use Markets to Feed the Poor", *Journal of Economic Perspectives,* 2017, 31 (4), 145-162.

Prendergast, Canice, "The Allocation of Food to Food Banks", *Journal of Political Economy*, 2022, 130 (8), 1993-2017.

Roberts, R., "A Positive Model of Private Charity and Public Transfers", *Journal of Political Economy*, 1984, 92, 138-48.

Roth, Al, "What Have We Learned From Market Design?", *Economic Journal*, 2008, Vol 118, 285-310.

Rose-Ackerman, Susan, "Charitable Giving and 'Excessive' Fundraising.", *Quarterly Journal of Economics*, 1982, 97, 193-212.

Ruhm, C., and C. Borkowski, "Compensation in the nonprofit sector", *Journal of Human Resources*, 2003, 38 (4), 992-1021.

Salamon, L., and S. W. Sokolowski, "Nonprofit organizations: New insights from QCEW data", *Monthly Labor Review*, 2005, September, 19-26.

Small Charities Data, "Trading and enterprise: how small charities earn their income", 2024. https://smallcharitiesdata.org/topic/trading-and-enterprise-how-small-charities-earn-their-income/

Statista, "Expenses of reporting 501(c)(3) nonprofit organizations in the United States from 1998 to 2020", 2024. https://www.statista.com/statistics/189257/expenses-of-reporting-non-profit-organizations-in-the-united-states-since-1998/.

US Census Bureau, "U.S. Volunteerism Rebounding After COVID-19 Pandemic", November 2024. https://www.census.gov/library/stories/2024/11/civic-engagement-and-volunteerism.html

USA Facts, "How many nonprofits are there in the US?", November 2023. https://usafacts.org/articles/how-many-nonprofits-are-there-in-the-us/

White. S., "Early Decision Provides A Big Bang For Your Buck At These Colleges". *Forbes*, July 15, 2024.

Wong, M.B., "Money and Barter in the Field: Evidence from the Life and Death of a Digital Currency", 2024, University of Hong Kong.

World Economic Forum, "The State of Social Enterprise: A Review of Global Data 2013–2023", 2024, Schwab Foundation for Social Entrepreneurship.

Wright B. E. and Davis B. S., "Job satisfaction in the public sector the role of the work environment". *The American Review of Public Administration*, 2003, 33, 70-90.

Yoruk, Baris, "Charity Ratings", *Journal of Economics and Management Strategy*, 2016, 25(1), 195-219.

Zhang, Yiran, Wen Xu, Susan Yoon, Weizhi Chen, and Sarah Parmenter, "Workplace support, job autonomy, and turnover intention among child welfare workers in China: The mediating role of job satisfaction", *Child Protection and Practice*, 2024, Volume 2, August 2024, 100027.

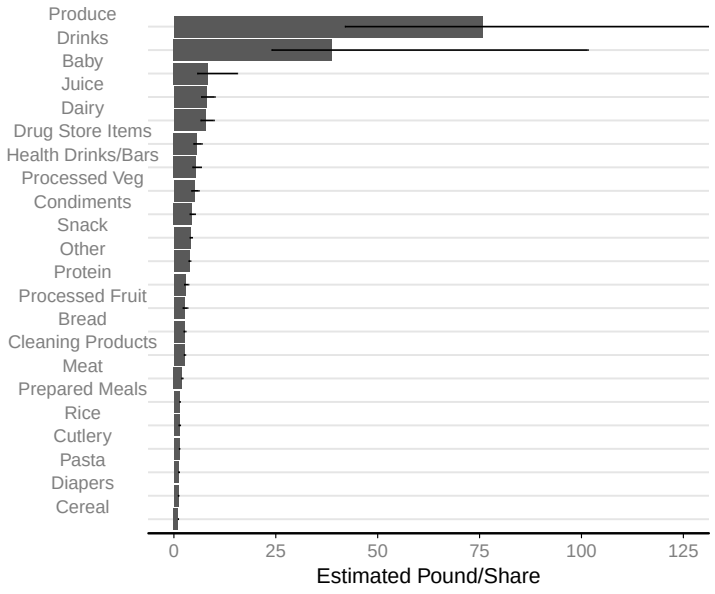


Figure 3: **Feeding America Price of Different Goods.**

7 Appendix

The Program: The firm's objective given A is to choose I_i^{*A}, P^A, and w^{*A} to maximize $S = \sum_{i=1}^{\infty}(v_i + T^A(t_i, \tau_i))I_i$, where $v_i = v - \delta i$ and $I_i \in \{0, 1\}$. If $w^A < p^A$, then $I^A = \sum_{i=1}^{\infty} I_i^A$ is defined by

$$I^A \leq \frac{\mathbb{D}}{w^A - r^A}. \tag{28}$$

S is maximized subject to (28), the Fairness Constraint and

1. Wages are given by (5).
2. $T(i, j|\underline{\phi}) \geq \underline{m} p_j^A$, in any equilibrium where type τ_i receives treatment $t = j$ with positive probability.
3. Client Reporting in A:
   - Let $H_{ij}^A(\hat{t}_l|\hat{t}_{-l})$ be the probability that client l of type $\tau = i$ receives $t = j$ after reporting a ranking $\hat{t}_l$ and others report $\hat{t}_{-l}$. In a Bayesian Nash equilibrium, client l reports $\hat{t}$ when others report $\hat{t}_{-l}$ if

     $$\sum_j H_{ij}^A(\hat{t}_l|\hat{t}_{-l})(T(i, j|\phi) - m_i p_j^A) \geq \sum_j H_{ij}^A(\hat{t'}_l|\hat{t}_{-l})(T(i, j|\phi) - m_i p_j^A), \forall m \in [\underline{m}, \overline{m}], \tag{29}$$

     where H_{ij}^A are generated by Bayes rule.

- If A involves Assigned Consumption and the allocation mechanism is strategy proof, this simplifies to

$$T(i,1|\underline{\phi}) - m_i p_1^A \geq T(i,0|\underline{\phi}) - m_i p_0^A, \forall m_i \in [\underline{m}, \overline{m}], \tag{30}$$

and

$$T(i,i|\overline{\phi}) - m_i p_j^A \geq T(i,j|\overline{\phi}) - m_i p_k^A, \forall m \in [\underline{m}, \overline{m}]. \tag{31}$$

- If A involves Choice where the client chooses $I_i^A = k_i^A \leq 1$ at unit price p_i^A

$$k_i^A(T(i,i|\underline{\phi}) - m_i p_i^A) \geq p_j^A(T(i,j|\underline{\phi}) - m_i p_j^A), \forall m \in [\underline{m}, \overline{m}]. \tag{32}$$

Proof of Lemma 1: Unrestricted Choice requires that for a client i of type $\tau_i = a$, $a^2 + \phi - m_i p_a \geq ab - m_i p_b$, for all a, b, ϕ and m_i. Under Client Choice, let $\hat{t}(\tau, \underline{\phi}, m)$ be the treatment chosen (meaning their first choice) by a client of type $\tau, \underline{\phi}$ with marginal utility m. Consider the case where a client misreports $\hat{t}^1 > \tau$: $\hat{t}(0, \underline{\phi}, m) = 0 \succ \hat{t}(0, \underline{\phi}, m) = 1$ if $\underline{\phi} - mp_0 \geq \mu - mp_1, \forall m$, or $p_1 - p_0 \geq \frac{\mu - \underline{\phi}}{\overline{m}}$, for $\frac{\underline{\phi}}{\underline{m}} \geq p_0 \geq 0$ and $p_1 \leq \frac{\mu + \underline{\phi}}{\underline{m}}$. This can be satisfied if $\overline{m} p_0^* \leq \mu - \underline{\phi} + \overline{m}\frac{\mu + \underline{\phi}}{\underline{m}}$. As $p_0 \geq 0$, the Fairness Constraint can only be satisfied if $\frac{\overline{m}}{\underline{m}} \geq \frac{\mu - \underline{\phi}}{\mu + \underline{\phi}}$. If this holds, $p_0^* = \min\{\frac{\underline{\phi} - \mu}{\overline{m}} + \frac{\mu + \underline{\phi}}{\underline{m}}, \frac{\underline{\phi}}{\underline{m}}\}$. If $\frac{\overline{m}}{\underline{m}} \geq 1 - \frac{\underline{\phi}}{\mu}$, $p_0^* = \frac{\underline{\phi}}{\underline{m}}$. If this holds for type $\underline{\phi}$, it also holds for type $\overline{\phi}$. With this contract, no client wishes to misreport downwards.

If $\frac{\overline{m}}{\underline{m}} < \frac{\mu - \underline{\phi}}{\mu + \underline{\phi}}$, Assigned Consumption is used. Agent Assignment and Bureaucratic Assignment do not depend on client rankings. As $q < 1$, for any assigned consumption mechanism, $prob(t = 0|\tau) > 0$ for all τ. Because $T(\tau, 0) = 0$ for $\tau = 1 \implies p_0 = 0$. As $min_\tau T(1, \tau) = \mu$, this implies that $p_1 \leq \frac{\mu}{\underline{m}}$. Letting this hold with equality implies that $P^{ac} = \{0, \frac{\mu}{\underline{m}}\}$. Let $\hat{t}(\tau, \phi, m)$ be the ranking of a client with types τ, ϕ and m. Fairness with Assigned Consumption implies $\hat{t}(\tau, \underline{\phi}, m) = (1 \succ 0)$, $\hat{t}(1, \overline{\phi}, m) = (1 \succ 0)$, and $\hat{t}(0, \overline{\phi}, m) = (0 \succ 1)$ for all m. Then $\hat{t}(\tau, \underline{\phi}, m) = (1 \succ 0)$ requires $\frac{\mu}{\underline{m}} - p_1^{da} \geq \frac{\underline{\phi}}{\underline{m}} - p_0^{da}$. This holds with equality so $p_1^{da} - p_0^{da} = \frac{\mu - \underline{\phi}}{\underline{m}}$. Because $T(\tau, 0) = 0$ for $\tau = 1 \implies P^{da} = \{0, \frac{\mu - \underline{\phi}}{\underline{m}}\}$ and $p^{da} = \frac{0 + \frac{\mu - \underline{\phi}}{\underline{m}}}{2} = \frac{\mu - \underline{\phi}}{2\underline{m}}$. □

Proof of Proposition 1: Surplus is $S^* = \frac{D}{w^A - r^A}\left(v + T^A - \frac{\delta D}{2(w^A - p^A)}\right)$. From Lemma 1, $p = p^{da}$. Let $\hat{\tau}^a$ be the agent's posterior. Then as $t \perp (\hat{\tau}^a | \tau^a) \implies \hat{U}^a = \overline{U}^a \implies w = \overline{w}$ and (5) holds. Then $I = \frac{D}{\overline{w} - p^{da}}$.

Clients $i = 1, 2, ...I^{da}$ report $\hat{t}$. The firm offers $I_i = 1 \forall i = 1, 2, ...\frac{D}{\overline{w} - p^{da}}$. $P^{da} = \{0, \frac{\mu - \underline{\phi}}{\underline{m}}\} \implies \hat{t}(\tau, \underline{\phi}, m) = (2 \succ 1 \succ 0)$, $\hat{t}(2, \overline{\phi}, m) = (2 \succ 1 \succ 0)$, $\hat{t}(1, \overline{\phi}, m) = (1 \succ 2 \succ 0)$, and $\hat{t}(0, \overline{\phi}, m) = (0 \succ 2 \succ 1)$ for all m. As a result

$$\hat{t} = \begin{cases} 1 \succ 0 & \text{with probability } 1 - \frac{h}{2} \\ 0 \succ 1 & \text{with probability } \frac{h}{2}. \end{cases} \tag{33}$$

1. $t = 1$ *if* $\tau^a = 1$ *and* $\hat{t} = (1 \succ 0)$. Consider all who report $\hat{t} = (1 \succ 0)$. Of these, $\frac{1}{2}$ are $\tau = 1$ and $\frac{(1-h)}{2}$ are $\tau = 0$. $Prob(\tau^a = 1|\hat{t} = (1 \succ 0)) = 1 - h(1-q)$. Let $O(t|\tau^a, \hat{t})$ be the probability that a client with $\hat{t}$ and signal τ^a receives an offer of treatment t. Step 1 of the algorithm $\implies O(1|1, (1 \succ 0)) = 1$.

2. $t = 1$ *if* $\tau^a = 0$ *and* $\hat{t} = (1 \succ 0)$ *until there are no more* $t = 1$ *slots.* There are then $\frac{h(1-q)}{2}$ remaining $t = 1$ slots available per capita, and $\frac{q(1-h)+(1-q)}{2}$) clients for whom $\hat{t} = (1 \succ 0)$ and $\tau^a = 0$. Step 2 $\implies$ $O(1|0, (1 \succ 0)) = r = \frac{h(1-q)}{q(1-h)+1-q} = \frac{h(1-q)}{1-hq}$.

3. $t = 0$ *for all other clients.*

Surplus is $S^{da} = \frac{D}{\overline{w}-p^{da}} \left(v + T^{da} - \frac{\delta D}{2(\overline{w}-p^{da})} \right)$ where

$$T^{da} = \frac{\mu}{2} + q\hat{\phi} + \frac{h(1-q)\overline{\phi}}{2} + \frac{r((1-q)\hat{\phi} - q(1-h)\underline{\phi})}{2}. \tag{34}$$

Proof or Proposition 3: An agent who receives signal τ_i^a, with probability λ, $t_i = \tau_i^a$. With probability $1 - \lambda$, $t_i = \tilde{\tau}_i$. Then $\hat{T}^a \approx \lambda q\hat{\phi} + (1-\lambda)\hat{\phi}$. The agent's belief of the surplus from a random alternative agent is

$$\overline{T}_i^a = \begin{cases} q\hat{\phi} & \text{with probability } \lambda^2 \\ \rho\hat{\phi} & \text{with probability } (1-\lambda)^2 \\ q\hat{\phi} & \text{with probability } \lambda(1-\lambda). \\ \frac{\hat{\phi}}{2} & \text{with probability } (1-\lambda)\lambda, \end{cases} \tag{35}$$

$\implies \hat{T}^a - \overline{T}^a = (1-\lambda)(1-\frac{\lambda}{2}-(1-\lambda)\rho)\hat{\phi}$. Then from (5), $w^* = \overline{w} - \alpha(1-\lambda)(1-\frac{\lambda}{2}-(1-\lambda)\rho)\hat{\phi}$ (we ignore the approximation). Then as prices are $P^{ac} = \{0, \frac{\mu}{\underline{m}}\}, I^a = \frac{D}{w^*-p^{ac}}$. The firm's belief of surplus is $T^a = \mu + \lambda q\hat{\phi} + (1-\lambda)\frac{\hat{\phi}}{2}$ and

$$S^a = \frac{D}{w^* - p^{ac}} \left(v + T^a - \frac{\delta D}{2(w^* - p^{ac})} \right). \square \tag{36}$$

Agent Deferred Acceptance: Clients report $\hat{t}$ at prices P^{da} as above. Pareto efficiency implies that $\hat{t} = 0 \succ 1 \implies t = 0$, as the agent's posterior is that $\tau = 0$ with probability 1. There is a fraction $\frac{1}{2} - q\frac{h}{2}$ of clients with $\tau_i^a = 0$ of whom $\tau_i^a \neq \tilde{\tau}_i^a$ with probability $(1-\lambda)$. There is a fraction $\frac{1}{2} - (1-q)\frac{h}{2}$ of clients with $\tau_i^a = 1$ of whom $\tau_i^a \neq \tilde{\tau}_i^a$ with probability $(1-\lambda)$. The agent has a fraction $1 - \frac{h}{2}$ clients for which

$$\tau_i^a = \begin{cases} 1 & \text{with probability } \frac{1}{2} - (1-q)\frac{h}{2} \\ 0 & \text{with probability } \frac{1}{2} - q\frac{h}{2}. \end{cases} \tag{37}$$

The agent has beliefs that

$$\tilde{\tau}_i^a = \begin{cases} 1 & \text{with probability } \lambda(\frac{1}{2} - (1-q)\frac{h}{2}) + (1-\lambda)(\frac{1}{2} - \frac{h}{4}) \\ 0 & \text{with probability } \lambda(\frac{1}{2} - q\frac{h}{2}) + (1-\lambda)(\frac{1}{2} - \frac{h}{4}). \end{cases} \tag{38}$$

Per client, there are $\frac{1}{2}$ slots for $t=1$, and $1-\frac{h}{2}$ slots for $t=0$. The agent assigns $t=1$ to all clients with $\tilde{\tau}_i^a = 1$. There are then $\frac{1}{2} - \lambda(\frac{1}{2} - (1-q)\frac{h}{2}) + (1-\lambda)(\frac{1}{2} - \frac{h}{4})$ remaining $t=1$ slots. He randomly offers those to the $\lambda(\frac{1}{2} - q\frac{h}{2})$ clients for whom he agrees with the $\tau_i^a = 0$ signal. This implies randomizing with probability r above. The remaining clients are assigned $t=0$. Then

$$\hat{T}^a = \mu + \frac{h}{2}\underline{\phi} + (\frac{1}{2} - (1-q)\frac{h}{2})\lambda q\hat{\phi} + (1-\frac{h}{2})(1-\lambda)\frac{\hat{\phi}+\underline{\phi}}{2} + \\ (\frac{1}{2} - \frac{qh}{2})\lambda q\underline{\phi} + \lambda r\left((1-q)\hat{\phi} - q\underline{\phi}\right), \tag{39}$$

while

$$\overline{T}^a = \mu + \frac{h}{2}\underline{\phi} + (\frac{1}{2} - (1-q)\frac{h}{2})\lambda^2 q\hat{\phi} + (\frac{1}{2} - (1-q)\frac{h}{2})\lambda(1-\lambda)\frac{\hat{\phi}}{2} + \\ (1-\lambda)^2(\frac{1}{2} - \frac{h}{4})\rho\hat{\phi} + (1-\lambda)\lambda(\frac{1}{2} - \frac{h}{4})q\hat{\phi} + \\ (1-\lambda)^2(\frac{1}{2} - \frac{h}{4})\rho\underline{\phi} + (1-\lambda)\lambda(\frac{1}{2} - \frac{h}{4})q\underline{\phi} + (\frac{1}{2} + q\frac{h}{2})\lambda^2 q\underline{\phi} + \\ (\frac{1}{2} + q\frac{h}{2})\lambda(1-\lambda)\frac{\phi}{2} + \lambda r\left((1-q)\hat{\phi} - q\underline{\phi}\right), \tag{40}$$

which simplifies to

$$\mu + \frac{h}{2}\underline{\phi} + (\frac{1}{2} - (1-q)\frac{h}{2})\lambda^2 q\hat{\phi} + +(\frac{1}{2} - (1-q)\frac{h}{2})\lambda(1-\lambda)\frac{\hat{\phi}}{2} + \\ (1-\lambda)^2(\frac{1}{2} - \frac{h}{4})(\rho + \frac{1-\rho}{2})(\hat{\phi} + \underline{\phi}) + (1-\lambda)\lambda(\frac{1}{2} - \frac{h}{4})q(\hat{\phi} + \underline{\phi}) + \\ + (\frac{1}{2} + q\frac{h}{2})\lambda^2 q\underline{\phi} + (\frac{1}{2} + q\frac{h}{2})\lambda(1-\lambda)\frac{\phi}{2} + \lambda r\left((1-q)\hat{\phi} - q\underline{\phi}\right). \tag{41}$$

The wage that must be paid is

$$w^{**} = \overline{w} - \\ (1-\lambda)\left(\frac{\lambda}{2}\left(\frac{1}{2} - (1-q)\frac{h}{2})\right)\hat{\phi} + (\frac{1}{2} + q\frac{h}{2})\underline{\phi}\right) + \frac{\hat{\phi}+\underline{\phi}}{2}(1-\frac{h}{2})(1-(1-\lambda)\rho)\right) > w^*. \tag{42}$$

□

Proof of Proposition 4: Let $\hat{t}$ now refer to the good chosen by the client. Unrestricted Choice requires $\hat{t} = \tau_i$ at a price p_i. As $p_0 \geq 0$, the Fairness Constraint requires $\frac{V'(\hat{y}+\overline{\eta})}{V'(\hat{y}+\underline{\eta})} \equiv \frac{\overline{m}}{\underline{m}} \geq \frac{\mu-\underline{\phi}}{\mu+\underline{\phi}}$. Note that $\frac{d\left(\frac{V'(\hat{y}+\overline{\eta})}{V'(\hat{y}+\underline{\eta})}\right)}{d\hat{y}} > 0$ if $\frac{V''(\hat{y}+\overline{\eta})}{V'(\hat{y}+\overline{\eta})} > \frac{V''(\hat{y}+\underline{\eta})}{V'(\hat{y}+\underline{\eta})}$ as occurs under Decreasing Absolute Risk Aversion. As $V'' \to 0$ as $y \to \infty$, there $\hat{y}^{uc}$ such that $\frac{V'(\hat{y}+\overline{\eta})}{V'(\hat{y}+\underline{\eta})} \geq \frac{\mu-\underline{\phi}}{\mu+\underline{\phi}}, \forall \hat{y} \geq \hat{y}^{uc}$. Lemma 1 implies $P^{uc} = \{p_0^{uc}, \frac{\mu+\underline{\phi}}{\underline{m}}\}$, for all $\hat{y} \geq \hat{y}^{uc}$. Straightforward calculations show that $p_0^* = \frac{\underline{\phi}}{\underline{m}}$ for $\hat{y} \geq \tilde{y}^{uc}$ where $\frac{V'(\tilde{y}^{uc}+\overline{\eta})}{V'(\tilde{y}^{uc}+\underline{\eta})} = 1 - \frac{\underline{\phi}}{\mu}$. Revenue per client is $p^{uc} = \frac{\mu+p_0^{uc}+\underline{\phi}}{2V'(y^{uc}-\underline{\eta})}$ for $y^{uc} \leq \hat{y} \leq \tilde{y}^{uc}$ and $p^{uc} = \frac{\frac{\mu}{2}+\underline{\phi}}{V'(y^{uc}-\underline{\eta})}$ for $y^{uc} > \tilde{y}^{uc}$. As $T^{uc} = T^* = \mu + \hat{\phi}$ and $I^{uc} = \frac{D}{\overline{w}-r^{uc}}$,

$$ES^{uc} = \frac{D}{\overline{w} - p^{uc}} \left(v + T^* - \frac{\delta D}{2(\overline{w} - p^{uc})} \right). \square \tag{43}$$

The Non-Profit Outcome: Let $\Delta p = p^{ac} - p^{da} = \frac{\underline{\phi}}{2\underline{m}}$. The firm chooses Bureaucratic Assignment for all $D \leq D_b^*$, where D_b^* is uniquely defined by

$$(T^{da} - T^b) D_b^* = \frac{D_b^* \Delta p}{(\overline{w} - p^{da})(\overline{w} - p^{ac})} \left(v + T^b - \frac{\delta(D_b^* + \frac{\Delta p}{2})}{2(\overline{w} - p^{ac})} \right). \tag{44}$$

It chooses Agent Assignment over Bureaucratic Assignment for all $D \leq D_a^*$, where D_a^* is uniquely defined by

$$(T^b - T^a) D_a^* = \frac{D_a^* \Delta w}{(w^* - p^{ac})(\overline{w} - p^{ac})} \left(v + T^a - \frac{\delta(D_a^* + \frac{\Delta w}{2})}{2(w^* - p^{ac})} \right). \tag{45}$$

Monopoly Pricing: Let k_i be the probability that the client receives the good. The firm chooses $p_i(k_i)$ to maximize S subject to $I_i = 1 \implies I_{i'} = 1$ for $i' < i$. If $h\overline{\phi} < \underline{\phi} \implies h(\overline{\phi} + \mu) < \underline{\phi} + \mu \implies p_0(1) = \frac{\underline{\phi}}{\underline{m}}$ and $p_1(1) = \frac{\underline{\phi}+\mu}{\underline{m}}$ are the only prices charged. If $h\overline{\phi} > \underline{\phi}$ but $h(\overline{\phi} + \mu) < \underline{\phi} + \mu$, revenue is again maximized at $p_0(1) = \frac{\underline{\phi}}{\underline{m}}$ and $p_1(1) = \frac{\underline{\phi}+\mu}{\underline{m}}$, as the firm is required to exhaust all supply of $t = 0$. If $h(\overline{\phi} + \mu) > \underline{\phi} + \mu$, the firm may choose to price at $\tilde{p}_1(1) > \frac{\underline{\phi}+\mu}{\underline{m}}$ such that only those clients with $\tau = 1, \phi = \overline{\phi}$ will purchase $t = 1$ for sure. There is then residual supply of $\frac{1-h}{2}$ of $t = 1$ and supply of $\frac{1}{2}$ of $t = 0$. It chooses a random allocation where clients receive $t = 0$ with probability $\frac{1}{2-h}$ and $t = 1$ with probability $\frac{1-h}{2-h}$. It the maximum possible price of $\frac{1-h}{2-h}\mu$ for that. Incentive compatibility for the $\tau = 1, \overline{\phi}$ client is $\overline{\phi} + \mu - m_i p_1 \geq \frac{1-h}{2-h}\overline{\phi} \implies p_1 = \frac{\overline{\phi}+\mu-\frac{1-h}{2-h}\overline{\phi}}{\overline{m}} = \frac{\mu+\frac{1}{2-h}\overline{\phi}}{\overline{m}}$. For h high enough, this raises more revenue than in Proposition 4. The firm compares that extra revenue versus reduced surplus a from misallocated treatments. If surplus is higher from the extra revenue, the analog to (8) is $\frac{\overline{m}}{\underline{m}} \geq \frac{\mu-\underline{\phi}}{\mu+\frac{1}{2-h}\overline{\phi}}$. This retains the feature of Proposition 4 that choice is only possible for income high enough.

The Restricted Choice Mechanism:

1. The firm chooses quantities κ_1^* and unit prices $P^{rc} = (p_0^{rc}, p_1^{rc})$.
2. All clients $i \leq I^{rc}$ report $\hat{t}$ and receive their first choice.
3. The firm assigns $I_I = 1$ the remaining $t = 2$ at price r_2^{rc*} until there is no more remaining supply.

Proof of Proposition 6: *The firm chooses quantities* $t_i = (1, \kappa_1^*)$. Under Restricted Client Choice, let $\hat{t}(\tau, \underline{\phi}, m)$ be the good chosen by a client of type $\tau, \underline{\phi}$ with income m. Then $\hat{t}(0, \underline{\phi}, m) = 0 \succ \hat{t}(0, \underline{\phi}, m) = 1$ if $\underline{\phi} - m_i p_0 \geq \kappa_1(\mu - m_i p_1), \forall m$, for $\frac{\underline{\phi}}{\underline{m}} \geq p_0 \geq 0$ and $p_1 \leq \frac{\mu + \underline{\phi}}{\underline{m}}$. If the Fairness Constraint can be overcome $p_1^* = \frac{\mu+\underline{\phi}}{\underline{m}}$ and $\kappa_1 = 1$. For $\hat{y} < \tilde{y}^{uc}$, $\underline{\phi} + \overline{m} p_0^* = \kappa_1^*(\mu + \overline{m}\frac{\mu+\underline{\phi}}{\underline{m}})$ for $p_0^* \geq 0$. The benefits and costs of reducing p_0^* from $\underline{\phi}$ to 0 are linear so there are two relevant corner solutions as optima, where $p_0^* = \frac{\underline{\phi}}{\underline{m}}$ and $p_0^* = 0$. First, consider κ_1 at prices P^{uc}. Then (32) with $k_0 = 1$ implies that $(\hat{t}|\tau = 0) = 0 \succ (\hat{t}|\tau = 0) = 1$ iff $(1 - \frac{m}{\underline{m}})\underline{\phi} \geq \kappa_1(\mu - \frac{m}{\underline{m}}(\mu + \underline{\phi}))$ for all m. This requires $\kappa_1^* \leq \frac{1}{\frac{\mu}{\underline{\phi}} - \frac{\overline{m}\underline{\phi}}{\underline{m}(\underline{m}-\overline{m})}}$, as in the Proposition. There is remaining supply of $I^{rc}(\frac{1-\kappa_1^*}{2})$. This is assigned to all clients $i = I^{rc} + 1, ..., I^{rc}(1 + \frac{1-\kappa_1^*}{2})$ at a price of $p_1^{rc*} = \frac{\mu}{\underline{m}}$. Then revenue per client is $(1 - \frac{1-\kappa_1^*}{2})p^{uc} + \frac{1-\kappa_1^*}{2}p^{ac} = p^{rc}$. Surplus is

$$S^{rc} = I^{rc}\left(v + \mu + \hat{\phi}(1 - \frac{1-\kappa_1^*}{2}) + \frac{1-\kappa_1^*}{2}\frac{\hat{\phi}}{2}\right) - \delta\frac{I^{rc2}}{2}\left((1 - \frac{1-\kappa_1^*}{2}) + (1-\kappa_1^*)(1 + \frac{1-\kappa_1^*}{4})\right). \quad (46)$$

Now consider when $p_0 = 0$. Then $(\hat{t}|\tau = 0) = 0 \succ (\hat{t}|\tau = 0) = 1$ iff $\underline{\phi} \geq \kappa_1(\mu - \frac{m}{\underline{m}}(\mu + \underline{\phi}))$ for all m. This requires $\kappa_1^* \leq \frac{1}{\frac{\mu}{\underline{\phi}} - \frac{\overline{m}}{\underline{m}}\frac{\mu+\underline{\phi}}{\underline{\phi}}}$. The outcome is qualitatively identical to the other case.

The Artificial Currency Game: Each client $i = 1, 2, ...I^{rc}$ receives 1 unit of artificial currency. The firm offers supply of $t_0 = \frac{I^{rc}}{2}$ and $t_1 = \frac{\kappa_1^* I^{rc}}{2}$ at unit fiat money prices of $P = P^*$. Each client i reports minimum quantities $\lambda_{ij}$of artificial currency for t_j at unit fiat money price of p_i. The equilibrium price λ_j^* is the minimum λ_j at which demand is equal to supply. Any currency not spent has no value. Any over-subscribed treatments are assigned equally to all paying the equilibrium artificial currency price.

Let the equilibrium quantities per client be $(\lambda_0^*, \lambda_1^*)$. Then $(\lambda_0^*, \lambda_1^*) = (1, \frac{1}{\kappa_1^*})$ is an equilibrium outcome. Consider a client i of type τ_{ik}, $k = 0, 1$. Then consider the following equilibrium strategy: $\lambda_k = 0$ and $\lambda_{-k} = 1$. With these strategies, there is

demand for each good of $\frac{I^{rc}}{2}$ at a minimum allocation of 0, so with equal rationing the equilibrium quantities of each good are $(t_0^*, t_1^*) = (1, \frac{1}{\kappa_1^*})$. Finally consider the equilibrium strategies of each agent. By the definition of κ_i^*, client i of type $\tau_i = k$ prefers the quantity κ_i^* at unit price p_i, where $\kappa_0^* = 1$ to any κ_{-i}^*. Hence this is an equilibrium.

Proof of Lemma 1a: Unrestricted Choice requires that for a client i of type $\tau_i = a, a^2 + \phi - mp_a \geq ab - mp_b$ for all a, b, ϕ and m. Under Client Choice, let $\hat{t}(\tau, \underline{\phi}, m)$ be the treatment chosen (meaning their first choice) by a client of type $\tau, \underline{\phi}$ with income m. Consider all cases where a client misreports $\hat{t}^1 > \tau$. This implies that $\hat{t}(\tau, \underline{\phi}, m) = \tau \succ \hat{t}(\tau, \underline{\phi}, m) = \tau'$ if $\tau\mu + \underline{\phi} - mp_0 \geq \tau'\mu - mp_1, \forall m$, or $p_{\tau'} - p_\tau \geq \frac{\tau'\mu - \underline{\phi}}{\tau\mu\overline{m}}$, for $\frac{\underline{\phi}}{\underline{m}} \geq p_\tau \geq 0$ and $p_{\tau'} \leq \frac{\tau\ \mu + \underline{\phi}}{\underline{m}}$. The most binding is $\tau = 0$ and $\tau' = n - 1$ and can be satisfied if $\overline{m}p_0^* \leq (n-1)\mu - \underline{\phi} + \overline{m}\frac{(n-1)\mu + \underline{\phi}}{\underline{m}}$. As $p_0 \geq 0$, the Fairness Constraint can only be satisfied if

$$\frac{\overline{m}}{\underline{m}} \geq \frac{(n-1)\mu - \underline{\phi}}{(n-1)\mu + \underline{\phi}}. \tag{47}$$

Then if (47) holds, $p_0^* = \min\{\frac{\underline{\phi} - (n-1)\mu}{\overline{m}} + \frac{(n-1)\mu + \underline{\phi}}{\underline{m}}, \frac{\underline{\phi}}{\underline{m}}\}$. Then $p_0^* = \frac{\underline{\phi}}{\underline{m}}$ if $\frac{\overline{m}}{\underline{m}} \geq 1 - \frac{\underline{\phi}}{(n-1)\mu}$. If this holds for type $\underline{\phi}$, it also holds for type $\overline{\phi}$. It is also straightforward to show that with this contract, no client wishes to misreport downwards.

Now consider Deferred Acceptance. Let $\hat{t}(\tau, \phi, m)$ be the ranking of a client with types τ, ϕ and m. The Fairness Constraint with Assigned Consumption implies that $\hat{t}(\tau, \underline{\phi}, m) = (n \succ (n-1) \succ .. \succ 0)$ and $\hat{t}(\tau, \overline{\phi}, m) = (\tau \succ n \succ (n-1)... \succ 0)$ for all m. For a client of type τ, $\frac{\mu}{\underline{m}} - p_{\tau+1} \geq \frac{\underline{\phi}}{\underline{m}} - p_\tau$.These hold with equality and $p_{\tau+1} - p_\tau = \frac{\mu - \underline{\phi}}{\underline{m}}$ and $p_0 = 0$. This implies $P^{da} = \{0, \frac{\mu - \underline{\phi}}{\underline{m}}, \frac{2(\mu - \underline{\phi}))}{\underline{m}}, ...\}$ and $p^{da} = \frac{(n-1)(\mu - \underline{\phi})}{2\underline{m}}$. For Bureaucratic and Agent Assignment, $p^{ac} = \frac{(n-1)\mu}{2\underline{m}}$.

We now show the Deferred Acceptance outcome. We show this for h low, but the conceptual results generalize to all values of h. The mechanism is as follows:

1. The firm chooses $P^{da} = (p_0, p_1, ..., p_{n-1})$ and all clients $i < I^{da}$ report $\hat{t}_i = (\hat{t}_1, \hat{t}_2, ..\hat{t}_{n-1})$.

2. The firm *offers* $t_i \in \{0, 1, ...n-1\}$ to all clients $i < I^{da}$ for which $\tau_i^a = n - 1$. If there are remaining $t = n - 1$ slots, it randomly assigns them to clients for which $\tau_i^a = n - 2$, then $\tau_i^a = n - 3$..., until there are no more $t = n - 1$ slots.

3. The firm *offers* $t_i \in \{0, 1, ...n-1\}$ to all clients $i < I^{da}$ for which $\tau_i^a = n - 1$. If there are remaining $t = n - 1$ slots, it randomly assigns them to clients for which $\tau_i^a = n - 1$, then $\tau_i^a = n - 1$..., until there are no more $t = n - 1$ slots.

4. The firm *offers* $t_i \in \{0, 1, ...n-2\}$ to all clients $i < I^{da}$ for which $\tau_i^a = n-2$. If there are remaining $t = n-1$ slots, it randomly assigns them to clients for which $\tau_i^a = n-3$, then $\tau_i^a = n-4$..., until there are no more $t = n-2$ slots.

5. .....

6. The firm *offers* $t_i \in \{0, 1\}$ to all clients $i < I^{da}$ for which $\tau_i^a = 1$. If there are remaining $t = 1$ slots, it randomly assigns them to clients for which $\tau_i^a = 0$ there are no more $t = 1$ slots.

7. The firm allocates $t = 0$ to all other clients.

The outcome is as follows for $h \leq h^*$ defined below:

1. The firm offers $t \in \{0, 1, .., n-1\}$ to all the $\frac{1}{n}$ clients for which $\tau^a = n-1$. A fraction $1 - \frac{h(1-q)}{n}$ choose $t = n-1$ and a fraction $\frac{h(1-q)}{n(n-2)}$ choose each $t_i, i = 1, ..., n-2$.

2. The firm offers $t = n-1$ randomly to clients for which $\tau^a = n-2$ until there is no more supply of $t = n-1$ slots. This arises only if there are enough clients with $\tau^a = n-2$ who accept the $\frac{h(1-q)}{n}, j = n-1$ slots. This arises if

$$\frac{h(1-q)}{n} \leq \frac{1-q}{n^2} + \frac{(1-h)(1-q)(n-2)}{n^2} + \frac{(1-h)q}{n^2}. \tag{48}$$

   Let $r_{n-1} = \frac{h(1-q)}{n}\frac{1-q}{n^2} + \frac{(1-h)(1-q)(n-2)}{n^2} + \frac{(1-h)q}{n^2}$ be the probability of a client with $\tau^a = n-2$ being offered $t = n-1$. This must be less than 1. This arises if or $h \leq h^* = \frac{2(n-1)+\frac{q}{1-q}}{n-1+\frac{q}{1-q}} \geq \frac{1}{2}$.

3. The firm offers $t \in \{0, 1, .., n-2\}$ to all the $\frac{1}{n}$ clients for which $\tau^a = n-2$. A fraction $1 - \frac{h(1-q))(n-1)}{n^2}$ choose $t = n-2$ and a fraction $\frac{h(1-q)}{n(n-1)}$ choose each $t_i, i = 1, ..., n-3$.

4. The firm offers $t = n-2$ randomly to clients for which $\tau^a = n-3$ until there is no more supply of $t = n-2$ slots. This arises only if there are enough clients with $\tau^a = n-3$ who accept or $r_{n-2} \leq 1$

5. This algorithm continues for the remaining treatments. r_{n-j} is simply computed to be

$$r_{n-j} = \frac{n-j}{\frac{q(1-h)}{(1-q)h} - j + 2 + \frac{n-1}{n}}. \tag{49}$$

   Note that r_{n-j} is decreasing in j so that if $r_{n-1} \leq 1$ this also holds for all r_{n-j}. Hence if $h \leq h^*$ this is the outcome of the Deferred Acceptance mechanism.

Now let $h > h^* \implies \frac{h(1-q)}{n} > \frac{1-q}{n^2} + \frac{(1-h)(1-q)(n-2)}{n^2} + \frac{(1-h)q}{n^2}$. A client with $t_i^a = n-2$ is $t_i = n-1$ with probability 1. Let $n-k_1$ be defined by the highest value of τ^a such that not all clients with $\tau_i^a = \tau_{n-k_1}$ receive an offer of $t_i = n-1$. The game then continues by allocating $t = n-2$ using the residual supply of clients until no supply remains.

Unbalanced Demand and Supply: Here we identify the key outcomes for the case of $\zeta > 0$. First, Bureaucratic and Agent Assignment have the same prices as above, as they are set at the minimum valuations of those goods. Surplus from Bureaucratic Assignment is given by

$$\tilde{T}^b = \mu + \frac{1-\zeta}{2} q\hat{\phi} + \frac{1}{2}\frac{q(1+\zeta)}{q(1+\zeta)+(1-q)(1-\zeta)}\hat{\phi} + (q(1+\zeta)+(1-q)(1-\zeta)-\frac{1}{2})\frac{(1-q)(1-\zeta)}{q(1+\zeta)+(1-q)(1-\zeta)}\hat{\phi}. \quad (50)$$

Surplus from Agent Assignment $\mu + \frac{1-\zeta}{2}(\lambda q + \frac{1-\lambda}{2})\hat{\phi} + \frac{\zeta}{2}(\lambda(1-q)\hat{\phi} + \frac{1-\lambda}{2})\hat{\phi} + (\frac{1-(1-\lambda)(1+\zeta)}{2})q\hat{\phi}$, as the agent always assigns those clients he is certain about to $\tau = t$. As above, $\tilde{T}^b > \tilde{T}^a$. The wage reduction from Agent Assignment is $\tilde{\Delta}w = \alpha\hat{\phi}(\lambda(1-\lambda)(q-\frac{1}{2}) + (1-\zeta)(1-\lambda)^2\rho + \frac{(1-\lambda)\lambda}{2} + (1+\zeta)(1-\lambda)^2\rho + (1+\zeta)(1-\lambda)\lambda(1-q))$. Unrestricted Choice now requires that $p_0 = 0$ as some $\tau = 1$ clients are assigned $t = 0$. Then the condition for Unrestricted Choice is (8) but the maximum revenue from Unrestricted Choice is now $\frac{\mu+\underline{\phi}}{2}$ rather than $\mu + \frac{\underline{\phi}}{2}$ in the balanced case. This still exceeds the revenue from any other mechanism. The κ_i^* for the case of Restricted Choice remain as above.

Finally, the Deferred Acceptance mechanism yields higher allocative efficiency than any other Assigned Consumption mechanisms. There is one potential change. In the balanced case, r represented the fraction of clients receiving $\tau^a = 0$ who are assigned $t = 1$. This qualitatively remains the case if $q\frac{1+\zeta}{2} + (1-h)(1-q)\frac{1-\zeta}{2} < \frac{1}{2}$ so that there are not enough $\tau^a = 1$ signals to fill the $\frac{1}{2}$ slots. In that case the probability that a client who reports $\hat{t} = 1 \succ 0$ but has $\tau^a = 0$ receives $t = 1$ is $\tilde{r} = \frac{1-q(1+\zeta)-(1-h)(1-q)(1-\zeta)}{(1-q)(1+\zeta)+(1-h)q(1-\zeta)}$. Treatment surplus per client is then given by

$$\tilde{T}^{da} = \mu + q\frac{1+\zeta}{2}\hat{\phi} + \tilde{r}\frac{(1-q)(1+\zeta)}{(1-q)(1+\zeta)+(1-h)q(1-\zeta)} + \frac{h}{2}\underline{\phi} + (1-\tilde{r})\frac{(1-h)q(1-\zeta)}{(1-q)(1+\zeta)+(1-h)q(1-\zeta)} > \tilde{T}^b. \quad (51)$$

If however, $q\frac{1+\zeta}{2} + (1-h)(1-q)\frac{1-\zeta}{2} > \frac{1}{2}$, then the firm randomly assigns $t = 1$ to clients with $\tau^a = 1$ and assigns the residual to $t = 0$. This occurs with probability $\overline{r} = 1 - q(1+\zeta) - (1-h)(1-q)(1-\zeta)$. $\tilde{T}^{da}$ is defined analogously and exceeds $\tilde{T}^b$. Hence all the features of the balanced model are retained.

Proof of Proposition 7:

1. Prices $p_j^A, j = 0, 1, 2$ are given by $T(i, j|\underline{\phi}) \geq \underline{m} p_j^A$, in any equilibrium where type τ_i receives treatment $t = j$ with positive probability. Revenue per client is $p^A = \frac{p_0^A + p_1^A + p_2^A}{3}$.

2. Client Reporting in A:

   - If A involves Assigned Consumption and the allocation mechanism is strategy proof, this simplifies to

$$T(i, 2|\underline{\phi}) - m_i p_2^A \geq T(i, j|\underline{\phi}) - m_i p_j^A, \forall m \in [\underline{m}, \overline{m}], \tag{52}$$

$$T(i, 1|\underline{\phi}) - m_i p_1^A \geq T(i, 0|\underline{\phi}) - m_i p_0^A, \forall m \in [\underline{m}, \overline{m}], \tag{53}$$

   and

$$T(i, i|\overline{\phi}) - m_i p_i^A \geq T(i, j|\overline{\phi}) - m_i p_j^A, \forall m \in [\underline{m}, \overline{m}]. \tag{54}$$

The firm then chooses the A that maximizes ES given these constraints.

We first show existence of a Partial Choice equilibrium where (i) all $\tau = 1$ clients choose $t = 1$, (ii) all $\underline{\phi}$ clients of type $\tau \neq 2$ do not, and (ii) revenue raised exceeds r^b. This equilibrium has the following form. All $\tau = 1$ receive $t = 1$. Then the firm allocates all $t_i = \tau_i^a$ with probability 1 and allocates $t_i = 2$ with probability $\frac{1}{2}$ and $t_i = 0$ with probability $\frac{1}{2}$ when $\tau_i^a = 1$. To show existence of this equilibrium, we consider incentive compatibility. Type $\tau = 2$ does not choose $t = 1$ if $\mu(1 - \frac{\overline{m}}{\underline{m}}) \leq (q + \frac{1-q}{4})(2\mu(1 - \frac{\overline{m}}{\underline{m}}) + \phi)$. which always holds. Type $\tau = 1$ chooses $t = 1$ if $\mu(1 - \frac{\overline{m}}{\underline{m}}) + \phi \geq (\frac{q}{2} + \frac{1-q}{2})2\mu(1 - \frac{\overline{m}}{\underline{m}})$, which always holds. Finally, a client with preferences $\underline{\phi}$ and $\tau = 0$ does not choose $t = 1$ only if $\mu(1 - \frac{\overline{m}}{\underline{m}}) \leq \frac{3(1-q)}{4} 2\mu(1 - \frac{\overline{m}}{\underline{m}}) + (q + \frac{1-q}{2})\underline{\phi}$, which holds only if

$$\frac{\overline{m}}{\underline{m}} \geq 1 - \frac{(1 + 3q)\underline{\phi}}{2\mu(3q - 1)}. \tag{55}$$

Note that $1 - \frac{(1+3q)\underline{\phi}}{2\mu(3q-1)} < 1 - \frac{\underline{\phi}}{\mu}$, which defined $\tilde{y}^{uc}$. The condition for Unrestricted Choice remains $\tilde{y}^{uc}$ with three treatments. If these conditions hold, then an equilibrium exists at prices P^b. But if the final constraint is slack, the firm can increase p_1 above $p_1 = \frac{\mu}{\underline{m}}$, thereby showing $p^{1st} \geq p^b$. Note also that $T^{1st} = \frac{2q\hat{\phi}}{3} + \frac{\hat{\phi}}{3} > q\hat{\phi} = T^b$.

Partial Choice entails that one $\underline{\phi}$ type chooses their allocation $t = \tau_i$ and the other two are assigned. We conclude by showing that the only Partial Choice equilibrium is where type $\tau = 1$ choose $t = 1$.

1. Type $\tau = 0$. In order for type $\tau = 0$ to choose $t = 0$ it must be that $\underline{\phi} - m_i p_0 \geq min\{\mu - m_i p_1, 2\mu - m_i p_2\}$ for all m. As $t = 1$ and $t = 2$ are assigned by DA, $p_1 \leq \frac{\mu}{\underline{m}}$ and $p_2 \leq \frac{2\mu}{\underline{m}}$ as services are misallocated with positive probability. Similarly, $p_1 \geq 0$. This implies that $\frac{\overline{m}}{\underline{m}} \geq min\{\frac{2\mu - \underline{\phi}}{2\mu}, \frac{\mu - \underline{\phi}}{\mu}\}$. Note however that

the firm can use Unrestricted Choice, which has higher payoff, if $\frac{\overline{m}}{\underline{m}} \geq \frac{2\mu-\underline{\phi}}{2\mu+\overline{\phi}}$. But $\frac{2\mu-\underline{\phi}}{2\mu+\overline{\phi}} < min\{\frac{2\mu-\underline{\phi}}{2\mu}, \frac{\mu-\underline{\phi}}{\mu}\}$ and hence the firm never uses Partial Choice where $\tau = 0$ chooses.

2. Type $\tau = 2$. In order for only type $\tau = 2$ to choose $t = 0$ it must be all $\tau = 0$ types do not also do so. But this is the condition for Unrestricted Choice, and so the firm never implements Partial Choice with $\tau = 2$ choosing. $\square$