The Structure of the Social Organization

Canice Prendergast*

University of Chicago

March 30, 2025

Abstract

We offer a theory of the social organization, allocating services to clients subject to the need to fund their mission. Without funding constraints, we replicate the outcomes of the market design literature. When funding depends on service provision, we show a much broader taxonomy of optimal allocation mechanisms, ranging from different forms of Assigned Consumption to (subsidized) Choice. These outcomes are related to the absence of three roles that prices normally play in the canonical economic transaction. We also identify a novel constraint on whether firms incorporate as non-profit rather than a for-profit Social Enterprise.

*I am grateful to Jacob Leshno and seminar participants at the 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, p.93.

The private social sector is enormous, with non-profits alone accounting for over 10% of US employment, generating revenues of $3 trillion per annum.[1] There are also estimated to be ten million for-profit Social Enterprises worldwide. Social organizations play a fundamental role in society - they provide medical care to the sick, feed the hungry, house those in need of shelter, and on and on. The majority of social organizations distribute goods and services to needy clients in areas such as social assistance, housing, education and health care.[2] Our focus is on how they should be structured to most effectively distribute those services.

The single central issue for studying any type of organization is understanding how goods or services are best transferred to its consumers. The relationship between allocation mechanisms, efficiency and profit has been perhaps the dominant theme of modern microeconomics.[3] In the social sector, the market design literature on matching without prices has addressed how to allocate students to schools, kidneys to patients, and so on. Our starting point is that neither represents the reality of most private social organizations. The first literature has shown how profit-maximizing firms optimally design allocation mechanisms to sell to high "willingness to pay" clients. For social organizations, who typically exist to serve those who cannot pay, this is anathema to their mission. Furthermore, non-profits consistently report lack of funding as their greatest concern, yet the market design literature does not address the need for funding. Importantly, social organizations typically fund themselves through how they distribute services, with non-profits receiving 49% of revenue from fees from clients. A further 22% comes from external donors.

The firm we study distributes services to clients, but needs funds to do so as clients cannot afford them. When client coverage does not depend on service provision, we replicate the outcomes of the market design literature. When it does, we show a far wider range of ways to distribute services. Different forms of *Assigned Consumption* optimally arise, where the firm chooses services obtained by the client. The gold standard of the market design literature, Deferred Acceptance, is one example. The other end of the assignment spectrum is where the firm simply subsidizes services and allows clients *Unrestricted Choice*. There are also intermediate outcomes. *Partial Choice* arises when a client can choose among under-subscribed service, but over-

[1]Its importance has grown, with non-profit employment increasing by 33% over the last fifteen years compared to 9% in the for-profit sector (Bureau of Labor Statistics, 2024). The data on the revenue of the non-profit sector are from USA Facts, 2023, and employment from BLS, 2023.

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

[3]This is through vast literatures on practices such as optimal auction design, non-linear pricing, quality distortions, bundling, platform competition, tie-ins, and so on.

subscribed services are assigned by the firm. Finally, *Restricted Choice* occurs where clients choose not between available services A and B, but only between A and a less desirable service B'. For each of these allocation mechanisms, we also allow the firm to incorporate as a non-profit or for-profit.

We address the social organization through two lenses. First, that its mission is limited by its capacity to raise revenue. Second, that existing models of allocation mechanisms and pricing are inappropriate for understanding how social organizations operate: the outcome here is not some version of "sell to the highest bidder". We propose an alternative pricing model: that firms are governed by a *Fairness Constraint*, where no client can be denied for lack of income. For many service providers, this is a lexicographic constraint that defines their mission of serving those in need. On the margin of allocating services to a given client, this is also the least distortionary mechanism. Despite its simplicity, this offers rich implications for the social organization.

It is useful to see the firm's objective in terms of extensive and intensive margins. The extensive margin of the number of clients served is limited by funding. The intensive margin is allocative efficiency, the accuracy with which services match client needs. A key outcome below is that these are likely substitutes for the social organization for three distinct conceptual reasons. The firm's objective is to maximize client welfare and there are no agency concerns. Clients are privately informed of their need, which varies in both ordinal and cardinal dimensions. They also vary in their ability to pay for services. Clients may be able to report their preferences. To allocate services, the firm hires an agent who obtains a noisy signal on the client's needs. The firm chooses an organizational form and an allocation mechanism - for example, a non-profit using Deferred Acceptance - where prices, donations and agent compensation potentially depend on both choices.

Assigned Consumption arises when the firm determines consumption. For this to arise, it must be that client choice does not yield efficient outcomes. This occurs here when no price menu exists that generates efficient choice by all clients. Among Assigned Consumption mechanisms, the traditional gold standard is *Deferred Acceptance*, where clients' strategy proof rankings implement stable outcomes. *Bureaucratic Assignment* arises where outcomes arise by rule where client rankings play no role. Such rigid practices are characteristic of bureaucracies. Finally, *Agent Assignment* allows a social service provider to have discretion over services. We show how Deferred Acceptance has the highest allocative efficiency of these mechanisms but the lowest level of client coverage. Agent Assignment has the worst allocative efficiency but greatest client coverage, while Bureaucratic Assignment is intermediate on each dimension.

There are two distinct reasons why allocative efficiency and funding are substitutes here. Deferred Acceptance has two key features: strategy proof preference rankings and stable outcomes. In our setting, Deferred Acceptance is also Pareto efficient. First, the feature that limits clients from choosing services - that fairness

may be violated - simultaneously limits clients from ranking them. We show that this problem can be overcome only by reducing prices paid by clients, thereby decreasing firm revenue. Second, disagreement among stakeholders is a characteristic of many mission-based organizations. In the absence of disconfirming information, agents are assumed to disagree with the firm over client needs. As a result, Agent Assignment reduces allocative efficiency. However, it also reduces compensation as the marginal impact of the agent is higher. Put simply, algorithmic mechanisms may generate more efficient ex post outcomes, but are not much of an inducement for agents who are motivated by making a difference. We show that this logic also implies that Pareto efficiency is costly, even though it is common knowledge that client welfare is higher and all are motivated by altruism. As evidence of the potential importance of this channel, we describe an extensive empirical literature on the importance of worker autonomy for the recruitment and retention of social service providers.

Prices allow for an alternative distribution mechanism, namely client choice at subsidized prices. The only distortion here is that clients cannot afford the cost of service provision. *Unrestricted Choice* allows all relevant client information to be revealed through choice of subsidized prices. This has an obvious appeal in overcoming the "how shall I choose better than you?" component of the opening quote. It also raises more revenue. We show that if clients exhibit Decreasing Absolute Risk Aversion, (subsidized) Unrestricted Choice can screen efficiently if and only if average client income is sufficiently high. This is because two-dimensional screening reduces to a single dimension.

Many services assigned by social organizations are vertically differentiated. Families prefer larger homes to smaller ones, patients prefer more tests, and so on. Without prices, it is particularly hard to allocate these efficiently. Progress can be made when services are divisible. *Restricted Choice* offers 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 even with low client income. This outcome can be implemented by a common market design, namely allowing clients access to artificial currency. The problem with Restricted Choice is poor targeting of high priority clients. We use two recent settings, Feeding America and Covax, to illustrate this tradeoff.

Sufficient client income is needed for Unconstrained Choice. Many social organizations allow clients to choose from undersubscribed services but oversubscribed ones are assigned by the firm. We show this *Partial Choice* outcome can be implemented through a non-strategy proof First Choice mechanism at lower levels of client income, allowing a form of choice for poorer clients. We show two further outcomes in that instance. The first is well known: First Choice mechanisms can offer higher allocative efficiency than Assigned Consumption. The second is not: Partial Choice raises more client revenue.

Viewing the social organization through the lens of funding leads to a variety of novel implications. First, firms that serve low income clients use Assigned Consump-

tion. Those with the lowest donations allow the agent to choose services, while those who are somewhat better funded allocate through Bureaucratic Assignment. Only those with sufficient funding use Deferred Acceptance. Yet even in this instance, the higher revenue from Partial Choice may result in its implementation over Deferred Acceptance. All marginal impacts of funding reduce the likelihood of Deferred Acceptance. Note that the remarkable success of Deferred Acceptance in school choice, kidney exchange and resident matching are all in settings where funding though the allocation mechanism is not central.

This approach also offers a pecking order for non-profits, where the firm first prioritizes employees interests, then imposes formal bureaucratic processes, and only then incorporates client feedback. This outcome addresses organizations serving low income clients: better resourced clients instead can have subsidized choice. This offers an additional implication, where although there are no increasing returns to scale in the technology, they occur endogenously. Non-profits aimed at clients who are less poor not only gain more revenue as those clients can pay more, but additionally can allow Choice, which itself results in additional revenue from giving clients more efficient outcomes. In this way, we highlight a particular problem facing non-profits serving the poorest sections of society.

So far, we have described strategies for the non-profit. Yet there are 10 million for-profit Social Enterprises worldwide, which account for 3% of global employment (World Economic Forum, 2024). The key characteristic of the non-profit is the Non-distribution Constraint, where residual income cannot be taken by the firm (Hansmann, 1980). The existing literature has focused on harmful firm incentives that arises from the for-profit social organization. Reflecting recent empirical evidence, we instead consider a setting where there is a value to the firm acting as residual claimant, such that - all else equal - the optimal organizational form is for-profit.

Despite this, we show that due to the behavior of donors (the other major source of income for the social organization), the sector is distorted towards too many non-profits. Our starting point here is that to a first approximation, small private donors do not give to for-profit enterprises. Yet many have a clear social mission, and a large fraction of expenditures are on that mission. Why then do Social Enterprises not attract undirected donor funding? We argue here that it arises because of the classic distinction between margins and averages. Donors overwhelmingly report that their primary consideration in giving is the impact of their gift. A donor gives holding the expected donations of others constant. As a result, she evaluates her potential gift at her belief of the use of the marginal dollar received by the firm. If the marginal dollar goes to profit, she will not give as her donation has no impact on the mission. This arises even if the vast majority of infra-marginal funds are used on the mission. This also means that the non-profit firm has to live "hand-to-mouth" to attract donors, never sure of how well it can cover its mission.

As a result, an organizational form that increases allocative efficiency deters donors, such that the Social Enterprise can only exist if it is fully client-funded.

This becomes the final conceptual reason why allocative efficiency and funding are substitutes, and offers a non-behavioral reason why a firm that predominantly uses resources for social objectives cannot attract donors. We show this in two settings where residual claimant status increase allocative efficiency: where the firm offers financial security and where incentive contracting is needed.

We begin in Section 1 by describing the model. Section 2 considers the optimal form of the non-profit, while we address the Social Enterprise in Section 3. Section 2.5 addresses robustness and extensions while Section 5 provides an agency interpretation of the results. Section 6 concludes.

1 Model

A firm provides treatments $t_i \in \{0, 1, 2\}$ with intensity $I_i \in [0, 1]$ to clients i, where $i = 1, 2, 3, ..., \infty$. The return to treatment depends on a client's privately known type $\tau \in \{0, 1, 2\}$. 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.[4] We assume that the supply of types $\tau \in \{0, 1, 2\}$ and the distribution of treatments $t = \{0, 1, 2\}$ is balanced and uniform, where there are $\frac{1}{3}$ of all types and treatments.[5] We also assume that the number of clients is large to avoid sampling error. Surplus to the firm 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 $v_i = v - \delta i$. Clients are served in order, $i = 1, 2, ...$[6]

To allocate services to clients, the firm chooses an assignment mechanism A described below. The firm charges prices $R = \{r_0, r_1, r_2\}$, where $r_j \geq 0$ is the unit charge for treatment t_j.

[4]There are many examples in the social sphere that reflect this combination of vertical and horizontal differentiation. 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.

[5]Below we discuss the impact of unbalanced demand and supply.

[6]There is a cost below to serving a client given by the wage cost of the agent. We interpret this cost as a diagnosis cost incurred to determine the appropriate t to offer the client.

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

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

There is uncertainty over the client's capacity to pay for services, where m_i is privately known and defined over a support $[\overline{m}, \underline{m}]$. The parameter m_i derives from a client's indirect utility function $V(y)$, where $y \geq 0$ is her income. We assume that $V'(0) = \infty, V^{'} > 0, V^{''} < 0$, so $m_i = V'(y_i)$. Let the poorest clients have marginal utility $\underline{m}$ and the richest $\overline{m} < \underline{m}$. Income and type τ are independent. (We interpret the client's preferences below.)

The problem is only of interest if client preferences do not lead to trivial outcomes.[7] 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 simply ask them what they want. We choose preferences such that client rankings 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.[8]

Assumption 1: $\overline{\phi} > 2\mu$ and $\underline{\phi} < \mu$.

Assumption 1 implies that absent pricing those with $\phi = \underline{\phi}$ prefer $t = 2$, while those with $\overline{\phi}$ prefer τ_i.[9]

Agent: The firm hires an agent to implement t_i. Agent a receives a (contractible) signal τ_i^a on each served client's type:

$$\tau_i^a = \begin{cases} \tau_i & \text{with probability } q \\ \tau_j \neq \tau_i & \text{with probability } \frac{1-q}{2}, \end{cases} \tag{4}$$

[7]The model reflects the concerns of the market design literature. There is a return to matching ϕ but where vertical differentiation μ implies that clients may not choose the efficient outcome, $\tau_i = t_i$. The firm has a priority ranking over clients given by v_i.

[8]Note that client preferences have been chosen to reflect the two approaches to market design in the social sector. First, there is variation in the ordinal preferences of clients, where a fraction $1 - \frac{2h}{3}$ clients have $t = 2$ as there utility maximizing treatment, with $\frac{h}{3}$ for each of the other two. This reflects the market design literature on matching. Second, there is variation in the intensity of preferences. For instance, a client of type $\tau = 2$ values $t = 2$ over $t = 1$ at $\mu + \phi$ whereas a client of type $\tau = 1$ values $t = 2$ over $t = 1$ at $\mu - \phi$.

[9]This assumption implies that for clients with type $\underline{\phi}$, $t = 2 \succ t = 1 \succ t = 0$. Their vertical preferences are for a"higher" treatment. We do not address the intermediate case where $\mu < \underline{\phi} < 2\mu$. In this case, there is only one instance where client preferences are distorted relative to firm preferences, which is where type $\tau = 0$ and $\underline{\phi}$ prefers $t = 2$ to $t = 0$. Otherwise the firm can delegate choice to the client. This case leads to a greater region where client choice is used - as they can be trusted more - but is otherwise conceptually identical to the outcomes below.

where $\frac{1}{3} \leq q < 1$. Signals are uncorrelated across clients conditional on τ_i. The agent has preferences over wages and client welfare:

$$U^a = \sum_{i=1}^{\infty} (w + \alpha \hat{U}_i^c) I_i. \tag{5}$$

where α reflects altruism towards the client, and $\hat{U}_j^c$ is his belief about client i's expected utility. The agent has a reservation utility $\overline{U}$ per unit of I.[10]

The agent is motivated by the impact that he has on client welfare. A random agent is offered the position at a wage of w per unit of I. If he accepts the position, let $\hat{U}^{ca}$ be his belief about average client welfare, so his utility is then $U^a = w + \alpha \hat{U}^{ca}$. if he does not, let $U^a = \overline{U} + \alpha \overline{U}^{ca}$ be his utility. Then the agent accepts any offer where

$$w \geq \overline{U} + \alpha(\overline{U}^{ca} - \hat{U}^{ca}). \tag{6}$$

This is a formalization of a commonly stated preference by workers in the social sector, namely the desire to "make a difference". Note that if the allocation mechanism implies that the identity of the agent makes no difference, $\overline{U}^{ca} = \hat{U}^{ca}$ and the agent is paid his reservation wage.

The role of the agent is that he can disagree with others on appropriate actions: this is where the difference between $\overline{U}^{ca}$ and $\hat{U}^{ca}$ arise. We model this as differences in priors over client need. For the firm, τ_i^a is pivotal for choosing t_i. The agent holds a similar prior 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 a given agent holding a belief $\tilde{\tau}$, there is a probability $\rho \geq \frac{1}{3}$ of two randomly chosen agents agreeing.

The second component of funding is by external donors. Let b be donor belief of the probability that their donation is spent on I. Donations are given by $\mathbb{D}(b) + \epsilon$, where ϵ is a mean zero error term defined over a support of $[\underline{\epsilon}, \overline{\epsilon}]$ according to a distribution function F, with density $f > 0$. We assume that $\mathbb{D}(0) = 0$, $\mathbb{D}'(b) \geq 0$, $\mathbb{D}(1) \equiv D$. When the firm incorporates as a non-profit, $b = 1$ and donations are given by D.

Client Coverage: Client coverage $I = \sum I_i$ depends on funding. Let r be average revenue per client. The budget constraint depends on two cases. First, when $w > r$, donor-funding limits client coverage. Let π be the net profit of the firm (described below). Then

$$I = \frac{\mathbb{D} - \pi + \epsilon}{w - r}. \tag{7}$$

[10]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.

When $r \geq w$, the firm is self-funded without donors.

Profit Status: Let $\Pi \in \{0, 1\}$, where $\Pi = 0$ implies non-profit status and $\Pi = 0$ allows profit to be taken. Let π be the profit taken by the firm, where $E\pi = 0$ (this is simply to distinguish the firm from being a donor). We assume

$$\pi(\tilde{I}) \begin{cases} = 0 & \text{if } \Pi = 0 \\ \leq \mathbb{D} + \epsilon & \text{if } \Pi = 1. \end{cases} \tag{8}$$

The firm cannot take any revenue if is a non-profit.[11] If $\Pi = 1$, the only constraint is budget balance.

Pricing: Each client is defined by two attributes that affect her willingness to pay in (3), her value $T(\tau_i, t_i)$ and her capacity to pay m_i. A defining characteristic of many social organizations is that no client should be denied treatment for lack of money. To reflect this, we assume that treatment is independent of m_i:

Fairness Constraint: $t_i(\tau_i|m_i) = t_i(\tau_i), I_i(\tau_i|m_i) = I_i(\tau_i)$, for all m_i.

Client willingness to pay and firm objectives are misaligned because of differences in endowments. We have chosen a pricing benchmark that is least distortionary, in that client income can neither exclude her from service, nor affect it.

Strategies and timing:

1. The firm chooses A, R^A, w^A, π^A and Π, which is observed by clients and agents.
2. The firm offers the position to one agent subject to (6).
3. The firm and agent observes τ_i^a for all i.
4. ϵ is revealed.
5. A is implemented, r_i^A paid by clients, and w^A and π^A are paid.

Given this timing, the objective of the firm is to choose $A, w^A, \Pi, \pi^A(\tilde{I})$ and R^A to maximize ES subject to (3), (5), (6), (7), $E\pi = 0$, the Fairness Constraint, and Assumption 1. The equilibrium concept used is Bayesian Nash. Throughout, we restrict

[11] Assuming $\pi = 0$ is an exaggeration of the legal standard. This is discussed below.

attention to pure strategies.[12] The program is formally defined in the Appendix.[13]

Prices play three roles in the canonical economic transaction. First, efficient pricing allows the firm to delegate choice to clients. Second, prices serve as information to coordinate the actions of suppliers. In the famous words of Hayek, they "act to co-ordinate the separate actions of different people". Finally, they pay the firm's bills. However, social organizations typically exist as (at least some) clients cannot pay enough to cover service costs. All three roles for prices fail in different ways below. The firm's problem then becomes the quote that begins the paper: how best to allocate an (endogenously) scarce resource to clients who know more than the firm does.

A Benchmark: Without describing A, the modeling assumptions allow a simple characterization of surplus. Let T^A be average per capita treatment surplus from assignment mechanism A, let r^A be average revenue per client, and let w^A be the wage paid per unit of I. Then for a given realization of ϵ and firm profit π, $I^A = \frac{\mathbb{D}-\pi+\epsilon}{w^A-r^A}$ is the number of services (treatments) distributed by the donor constrained firm. The first best level of treatment for the firm is $\frac{v+T^A}{\delta}$. To avoid trivial outcomes where the firm discards donations, we assume $\mathbb{D}+\overline{\epsilon} < \frac{v+T^A}{\delta}$. As returns are linear in $I_i \in (0,1)$, optimal service is $I_i \in \{0,1\}$, where $I_i = 1$ for all $i \leq I^A$ and 0 otherwise. As $v_i = v - \delta i$, the average value of a client with realization ϵ and firm profit π is $v+T^A-\frac{\delta(\mathbb{D}-\pi+\epsilon)}{2(w^A-r^A)}$. As a result, expected surplus is $ES^* = \int_{\underline{\epsilon}}^{\underline{\epsilon}} \frac{\mathbb{D}-\pi+\epsilon}{w^A-r^A}\left(v+T^A-\frac{\delta(\mathbb{D}-\pi+\epsilon)f(x)dx}{2(w^A-r^A)}\right)$, or

$$ES^* = \frac{\mathbb{D}}{w^A-r^A}\left(v+T^A-\frac{\delta\mathbb{D}}{2(w^A-r^A)}\right) - \frac{\delta\int_{\underline{\epsilon}}^{\overline{\epsilon}}(-\pi+x)^2}{2(w^A-r^A)^2}f(x)dx, \tag{9}$$

This characterization of the problem focuses the analysis on five endogenous variables: (i) wages w^A, (ii) client payments r^A, (iii) assignment surplus T^A, (iv) donations $\mathbb{D}$, and (v) firm profits π.[14]

Much of the analysis deals with the interaction between funding and allocative efficiency. Note two components of first best allocative efficiency. Let $\hat{\phi} = h\overline{\phi} + (1-h)\underline{\phi}$. First, efficient assignment entails $\tau_i = t_i$, with expected surplus $T^* = \mu + \hat{\phi}$. Let

[12] We do not consider mixed strategies. One of the central concerns of the market design literature is the need to offer simple choices to clients, with its focus on strategy proof mechanisms. We exclude mixing as social organizations are unlikely to impose these costs on their clients.

[13] The client has three pieces of private information, m, τ and ϕ. In general, this could render the problem complex. However, the modeling assumptions imply that it is much simplified. Although ability to pay varies between $\underline{m}$ and $\overline{m}$, the Fairness Constraint merely means that we need to check outcomes at the boundaries. Second, Assumption 1 implies that type $\overline{\phi}$ always wants $\tau = t$ as its preferred outcome. As a result, for most of the analysis below we merely need to check for the preferences of type $\underline{\phi}$ for different τ.

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

$\tilde{I} = \frac{D+\epsilon}{w-r}$. Second, the firm can eliminate uncertainty in funding by choosing $\pi(\tilde{I}) = \epsilon$. First best surplus conditional on $\mathbb{D}, w^A$ and r^A is then

$$ES^{**} = \frac{\mathbb{D}}{w^A - r^A}\left(v + \mu + \hat{\phi} - \frac{\delta \mathbb{D}}{2(w^A - r^A)}\right). \tag{10}$$

1.1 Outcomes and Assignment Mechanisms

The central issue in the paper is who makes the decision over the allocation of the service. There are three parties: the client, the firm, and the agent. Key to outcomes is the role played by the client. There are three possible outcomes: (i) when the client reveals τ_i, (ii) when the client reveals an ordinal ranking of $T(\tau_i, t_i)$, and (iii) when the client reveals nothing. We now address different outcomes and assignment mechanisms A associated with these:

1. Assigned Consumption arises the firm distributes services. We consider three cases:

   - *Deferred Acceptance*: When the client reveals an ordinal ranking of $T(\tau_i, t_i)$ that is used in a standard DA mechanism.[15]
   - *Bureaucratic Assignment*: The agent allocates based solely on the firm's signal.
   - *Agent Assignment*: The agent allocates based solely on his beliefs.

2. Choice arises when the client reveals τ through his actions:

   - *Unrestricted Choice*: The client receives $I_i = 1$ of her highest ranked preference.
   - *Restricted Choice*: The client 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.

For each of these, the firm decides whether to operate as a non-profit or as a social enterprise.

[15] We are bundling Top Trading Cycles with Deferred Acceptance here. In our setting, there is no difference between Top Trading Cycles and Deferred Acceptance.

2 The Non-Profit

We begin by considering the optimal structure of the non-profit. The non-profit spends all funding on the mission. This implies that $\pi = 0$ and $\mathbb{D} = D$. As a result, expected surplus in (9) becomes

$$ES^* = \frac{D}{w^A - r^A}\left(v + T^A - \frac{\delta D}{2(w^A - r^A)}\right) - \frac{\delta \sigma_\epsilon^2}{2(w^A - r^A)^2}, \tag{11}$$

where $\sigma_\epsilon^2 = \int_{\underline{\epsilon}}^{\epsilon} x^2 f(x)dx$. This characterization of the problem allows us to evaluate the non-profit in terms of three endogenous variables, T^A, r^A, and w^A. In this section, we show the link between T^A and the other two variables.

A central issue for optimal assignment is the information that can be provided by clients on their needs. These include (i) complete revelation, (ii) the revelation of the ranking of T, and (iii) no information. The only case where (ii) arises is with Deferred Acceptance. As a result, we describe that mechanism before showing the cost of information revelation.

2.1 Assigned Consumption

Assigned Consumption arises when the the firm determines all services received. This arises when wither (i) capacity of the firm to price does not change the ranking of clients compared to a no-pricing regime. or (ii) the allocation is independent of $\hat{t}$.

2.1.1 Deferred Acceptance

Deferred Acceptance is the gold standard for matching without prices (Roth, 2008, Pathak and Sonmez, 2008, Abdulkadiroglu, 2005, etc). Clients report strategy proof rankings of $T(\tau_i, t_i)$, and the firm uses these to match services. Let $\hat{t}_i = (t_i^1 \succ t_i^2 \succ t_i^3)$ be the preference ranking of client i, where t_i^j is her jth preferred treatment.

The Deferred Acceptance mechanism is as follows. From the timing assumptions, the firm knows ϵ and hence I^{da} when allocating t_i. It offers $I_i = 1$ to all $i \leq I^{da}$ and 0 to all other clients.

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

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

There are two senses of priority for the firm. First, they prioritize across clients, given by v_i. Second, their priority ranking for treatments is given by τ_i^a. The algorithm works as follows. All those with $\tau_i^a = 2$ have highest priority over treatments, followed by those with $\tau_i^a = 1$. The firm begins the algorithm by matching those who report $t = 2$ as their first choice with those who have $\tau_i^a = 2$. All offers are accepted. If there are remaining $t = 2$ slots, these are offered randomly to those agents who report $t = 2$ as their first choice with those who have $\tau_i^a = 1$. The firm allocates $t = 1$ is a similar way, and the remaining clients receive $t = 0$.

We now show how fairness affects revenue raised. First, it requires that all clients are willing to pay for assigned treatment. In any equilibrium where type $\tau = k$ receives treatment $t = j$ with positive probability, this implies that

$$r_j \leq min_k \frac{T(k,j)}{\underline{m}}. \tag{12}$$

When all types of clients receive all services, $r_j \leq min_k \frac{T(k,j)}{\underline{m}}$ over all k. We show below that this implies that maximum possible revenue is $R^{ac} = \{0, \frac{\mu}{\underline{m}}, \frac{2\mu}{\underline{m}}\}$ with revenue per client $r^{ac} = \frac{\mu}{\underline{m}}$.

Second, Fairness requires that in any equilibrium where t_i depends on $\hat{t}_i$, $\hat{t}_i$ must be independent of y_i. This is the client's *Incentive Compatibility* constraint. Lemma 1 shows the impact of incentive compatibility, where Unrestricted Choice implies that type τ chooses $t = \tau$.

Lemma 1:

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

$$\frac{\overline{m}}{\underline{m}} \geq \frac{2\mu - \underline{\phi}}{2\mu + \underline{\phi}}. \tag{13}$$

  In that case, maximum prices are $R^* = \{r_0^*, \frac{\underline{\phi}+\mu}{\underline{m}}, \frac{\underline{\phi}+2\mu}{\underline{m}}\}$, *where* $r_0^* = \min\{\frac{\underline{\phi}-2\mu}{\overline{m}} + \frac{2\mu+\underline{\phi}}{\underline{m}}, \frac{\underline{\phi}}{\underline{m}}\}$.

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

First notice the importance of uncertainty over m_i from (13). The model requires a reason why Choice is not always the appropriate assignment mechanism. Without

uncertainty about m_i, $\frac{\overline{m}}{\underline{m}} = 1$ and the firm can always achieve the efficient outcome though Choice.[16] Uncertainty about capacity to pay is therefore necessary for Assigned Consumption. Also notice that vertical differentiation μ makes this condition harder to satisfy but differentiated preferences via $\underline{\phi}$ renders it easier.

Unrestricted Choice has greater allocative efficiency and raises more revenue than Deferred Acceptance. As a result, Deferred Acceptance is only potentially used when $\frac{\overline{m}}{\underline{m}} < \frac{2\mu-\underline{\phi}}{2\mu+\underline{\phi}}$, in which case it results in lower revenue than r^{ac}. In words, the allocative benefit of Deferred Acceptance comes at a revenue cost. The striking feature here is that revenue here is decreasing in client idiosyncratic preferences $\underline{\phi}$. Put differently, the feature that renders Deferred Acceptance valuable - namely horizontal differentiation of preferences - simultaneously reduces its revenue.

A simple example illustrates the logic. Imagine a housing association is assigning clients to three cities, Chicago ($t = 0$), Boston ($t = 1$), and New York ($t = 2$), where $\mu = 5$ and $\underline{\phi} = 2$. Then $T(k, j)$ for clients with $\phi = \underline{\phi}$ is given below:

$$\begin{array}{cc} \tau = & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{array}{c} Chi \\ Bos \\ NY \end{array} & \left[\begin{array}{ccc} 2 & 0 & 0 \\ 5 & 7 & 5 \\ 10 & 10 & 12 \end{array}\right] \end{array}$$

First, from (13) the firm can use Unconstrained Choice if $\frac{\overline{m}}{\underline{m}} \geq \frac{2}{3}$, and can charge prices of $r_0 = \frac{2}{\underline{m}}, r_1 = \frac{7}{\underline{m}}, r_2 = \frac{12}{\underline{m}}$ when $\frac{\overline{m}}{\underline{m}} \geq \frac{4}{5}$. This yields revenue of $\frac{\mu+\underline{\phi}}{\underline{m}} = \frac{7}{\underline{m}}$. Here, idiosyncratic preferences increase revenue. As $q < 1$, mistakes are made under Assigned Consumption so Individual Rationality implies that Chicago at most costs 0, Boston $\frac{5}{\underline{m}}$ and New York $\frac{10}{\underline{m}}$, with maximum revenue per client of $\frac{\mu}{\underline{m}} = \frac{5}{\underline{m}}$, independent of $\underline{\phi}$. However, all clients of type $\underline{\phi}$ report $\hat{t} = \{NY \succ Bos \succ Chi\}$ under Deferred Acceptance, which is the ranking in the matrix above. Then as $\tau = 1$ must prefer New York to Boston, their price difference cannot be larger than $\frac{10}{\underline{m}} - \frac{7}{\underline{m}} = \frac{3}{\underline{m}}$. Similarly, for type $\tau = 0$ the price difference between Boston and Chicago cannot be larger than $\frac{3}{\underline{m}}$. A a result, maximum DA prices are Chicago = 0, Boston = $\frac{3}{\underline{m}}$ and New York = $\frac{6}{\underline{m}}$, with expected revenue of $\frac{\mu-\underline{\phi}}{\underline{m}} = \frac{3}{\underline{m}}$. More generally, $min_k \frac{T(k,j)}{\underline{m}} - min_k \frac{T(k,j')}{\underline{m}} = \mu$ for $j > j'$ but $min_k \frac{T(k,j)-T(k,j')}{\underline{m}} = \mu - \underline{\phi}$ and the loss in revenue is $\underline{\phi}$. Hence, the source of the value of Deferred Acceptance, namely horizontal differentiation, simultaneously limits revenue.[17] (Below we discuss the robustness of this outcome with other client

[16] For example, with common marginal utility m, the firm generates efficient outcomes with prices $r_0 \in (0, \frac{\underline{\phi}}{\underline{m}}], r_1 \in (\frac{\mu}{\underline{m}}, \frac{\mu+\underline{\phi}}{\underline{m}}]$ and $r_2 \in (\frac{2\mu}{\underline{m}}, \frac{2\mu+\underline{\phi}}{\underline{m}}]$.

[17] Deferred Acceptance requires clients to rank outcomes. Without prices, in the example above all types rank *New York* $\succ$ *Boston* $\succ$ *Chicago*. However, once prices are used, some low income clients may say either Boston or Chicago as the cost to them is large, but their wealthier counterparts may stick with New York. This is the same violation of Fairness that limits Unrestricted Choice.

preferences.)

Having determined r^{da}, we can now illustrate the outcome with Deferred Acceptance. Proposition 1 provides the outcome for $h \leq \frac{2}{3-q}$ for brevity. The case of $h > \frac{2}{3-q}$ is given in the Appendix and is conceptually similar.

Proposition 1 *Let $h \leq \frac{2}{3-q}$. Expected surplus from Deferred Acceptance is*

$$ES^{da} = \frac{D}{\overline{U} - r^{da}} \left(v + T^{da} - \frac{\delta D}{2(\overline{U} - r^{da})} \right) - \frac{\delta \sigma_\epsilon^2}{2(\overline{U} - \underline{r})^2}. \tag{14}$$

where

$$T^{da} = \mu + \frac{1}{3} \left(\hat{\phi}(q + p_2 \frac{1-q}{2}) + \overline{\phi} h (q + \frac{1-q}{2}) + (1-h)\underline{\phi} q (1-p_2) + \hat{\phi} \frac{1-q}{2} p_1 \right) + \\ + \frac{1}{3} \left(h\overline{\phi} + q(1-p_1)(1-h)\underline{\phi} \right), \tag{15}$$

$p_2 = \frac{h(1-q)}{q(1-h)+(2-h)\frac{1-q}{2}}$, *and* $p_1 = \frac{\frac{h(1-q)}{2}}{(1-h)q+(1-q)}$.

Expected surplus in (11) depends on w^A, r^A and T^A. Revenue is r^{da}. Allocations are independent of the agent's prior so $\hat{U}^{ca} = \overline{U}^{ca}$ and $w^{da} = \overline{U}$ from (6). This implies that expected client coverage is $EI^{da} = \frac{D}{\overline{U} - r^{da}}$.

Now consider T^{da}. The outcome is stable as there is no agent who envies another's assignment, for which he has higher priority with the firm. Preferences are acyclic, so Deferred Acceptance outcomes are also Pareto efficient (Ergin, 2002). Surplus depends on q, the quality of the signal, and the p_i terms, which reflect the utilitarian benefits of Pareto efficiency, accruing from the $\frac{2h}{3}$ clients that do not report $t = 2$ as their first choice. Note that the p_i are decreasing in q: said more simply, the input of clients is less important when the firm is better informed.[18]

2.1.2 Bureaucratic Allocation

Bureaucratic allocation arises when outcomes depend only on the firm's signal. The firm knows I^b when allocating t. It offers $I_i = 1$ to all $i \leq I^b$.

[18] One firm allocates all services here, unlike the traditional market design setting where many firms make offers. The outcome in (14) is analogous to a Non-Random Serial Dictatorship algorithm as if each treatment t makes independent offers. After the firm observes its signal, it offers a choice of $t = \{0, 1, 2\}$ to all clients who receive a $\tau_i^a = 2$ signal, as if each treatment was offering its treatment to the highest ranked clients.Those who do not choose $t = 2$ split equally between $t = 1$ and $t = 0$. The firm then offers a choice of $t = \{0, 1, 2\}$ to those with $\tau_i^a = 1$ until the remaining $t = 2$ slots are filled. It does similarly for $t = 1$ slots among those who have not chosen. Note that when $q = \frac{1}{3}$, Deferred Acceptance is identical to Random Serial Dictatorship.

1. The firm chooses R^b.

2. The agent observes τ_i^a and allocates $t_i = \tau_i^a$ for each client $i < I^a$.

Note that the mechanism does not ask clients to report preferences.

Proposition 2 *Expected surplus from Bureaucratic Allocation is*

$$ES^b = \frac{D}{\overline{U} - r^b}\left(v + T^b - \frac{\delta D}{2(\overline{U} - r^b)}\right) - \frac{\delta\sigma_\epsilon^2}{2(\overline{U} - r^b)^2}, \tag{16}$$

where $r^b = \frac{\mu}{\underline{m}}$, *and* $T^b = \mu + q\hat{\phi}$.

This outcome follows automatically because the allocations do not depend on information provided by the client. As a result, the firm need only satisfy (12) for all k, yielding revenue per client of $r^b = \frac{\mu}{\underline{m}}$. Bureaucratic allocation has worse allocative efficiency than Deferred Acceptance but higher client coverage.[19]

2.1.3 Agent Allocation

Agent Allocation offers autonomy to the agent by allowing him to assign t with no constraints. The firm knows I^a when allocating t. It offers $I_i = 1$ to all $i \leq I^a$.

1. The firm chooses R^a.

2. The agent observes τ_i^a and allocates t_i for each client $i < I^a$.

Proposition 3 *Expected surplus from Agent Allocation is*

$$ES^a = \frac{D}{w^* - r^b}\left(v + T^a - \frac{\delta D}{2(w^* - r^b)}\right) - \frac{\delta\sigma_\epsilon^2}{2(w^* - r^b)^2}. \tag{17}$$

where

$$T^a = \mu + (\lambda q + \frac{1-\lambda}{3})\hat{\phi}, \tag{18}$$

and $w^* = \overline{U} - \alpha(1-\lambda)(1 - \frac{\lambda}{3} - (1-\lambda)\rho)\hat{\phi} < \overline{U}$.

Agent Allocation has lower allocative efficiency than Bureaucratic Allocation $T^b - T^a = (1-\lambda)\hat{\phi}(q - \frac{1}{3})$. Yet it has higher client coverage because $w^* < \overline{U}$, where $\overline{U} - w^* \equiv \Delta w^* = \alpha(1-\lambda)(1 - \frac{\lambda}{3} - (1-\lambda)\rho)\hat{\phi}$. He accepts a lower wage because with autonomy, he believes that he "makes a difference". Intuitively, the wage discount is higher when the extent of disagreement $(1-\lambda)$ is greater, when the agent cares

[19] $T^{da} - T^b = \frac{1}{3}\left(\hat{\phi}p_2\frac{1-q}{2} + h(1-q)\overline{\phi} - p_1 q(1-h)\underline{\phi} - h\overline{\phi}\frac{1-q}{2} + \hat{\phi}\frac{1-q}{2}p_1 - (1-h)\underline{\phi}p_2 q\right) > 0$ from the definitions of p_i.

more about the client (α), and when the client believes that an alternative agent is less likely to share his views (ρ lower). (Note that the assumption of the agent's prior being tight (≈ 1) is not important for this outcome.[20]) Notice also that the outcome depends not only on the agent potentially disagreeing with the firm (with probability $1-\lambda$), but also with an alternative hire ($1-\rho$). Below we describe empirical evidence on the importance of autonomy for workers in social organizations.

Agent Assignment is not Pareto efficient: there are pairs of clients that would swap allocations and increase ex post allocative efficiency. Pareto Efficient Agent Assignment can be easily implemented, by a very simple extension: by allowing clients to report their ordinal preferences. Those clients who report a first choice of $t \neq 2$ are then known by the agent to that type with probability 1, and so Pareto efficient outcome are implemented.[21] This increases allocative efficiency. Despite this, it reduces firm funding (see Appendix). Client welfare is known by all to be higher under Pareto Efficient Agent Assignment than in Proposition 3. Despite this, the agent demands a higher wage when outcomes are Pareto efficient. This is because agent compensation in (6) depends on his *marginal* impact on client welfare, not client welfare itself. Pareto efficiency reduces that marginal impact as in effect some clients choose their own allocation. Said more simply, algorithmic allocation mechanisms may increase allocative efficiency, they may do little to recruit employees who feel that any agent can implement a rule. Instead, agents are motivated by their marginal impact: "making a difference".

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.[22] It is also common in agency problems to distort ex post outcomes to improve ex ante incentives. Yet here the agent is only motivated by altruism towards the client. Despite this, a relevant antecedent here is Aghion and Tirole, 1997, where control is delegated to an agent to induce more effort. Yet this is not generated by private benefits, but by the classic economic distinction between margins and averages.

To summarize, among Assigned Consumption mechanisms, Deferred Acceptance has highest allocative efficiency for the usual reasons. But it covers fewest clients for two reasons. First, the feature that leads Choice to be infeasible here - that outcomes would benefit wealthier clients - also limits information from Deferred Acceptance unless revenue is reduced. Second, the use of an algorithm implies that the agent no

[20]The agent holds a prior $\tilde{\tau}$ where he is almost certain of the client's type with probability $1-\lambda$. This is for simplicity. Consider an alternative where with probability $1-\lambda$, the agent holds a prior where with probability q^a the client is type $\tilde{\tau}$ where $q^a \sim U[\frac{1}{3}, 1]$. Then the agent's prior is pivotal after observing τ_i^a with probability $\frac{3(1-q)}{2}$ and $\Delta w^* = \frac{3(1-q)}{2}\alpha(1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}$. The qualitative results remain.

[21]An alternative is to run a Deferred Acceptance algorithm using the agent's posterior rather than the firms. It offers a similar outcome.

[22]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, 2016).

longer "makes a difference". While this increases a covered client's welfare, it reduces the agent's marginal impact and hence his willingness to work for lower wages, even though he is motivated by altruism. More generally, unlike the standard for-profit firm, allocative efficiency and funding are substitutes here.

Empirical Research: The paper is meant to be anchored in the empirical concerns of the social organization. Its primary deviation from the existing literature is that funding is a first order concern for social organizations. Empirical evidence consistently point 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).[23] Because of this, we also have focused on r^A and D as the primary sources of funds for the social organization in (11), where non-profits receive 49% of their revenues from fees to clients, with another 22% coming from donors (National Council of Nonprofits, 2024).

Second, a key ingredient above is that what matters to employees is the impact that they have on clients. This is reflected in the autonomy that they have to treat clients. Put simply, if an algorithm determines outcomes, anyone could do that job. Social work is an occupation where autonomy is typically seen as an important component of serving clients.[24] This is typically empirically measured when firms choose to 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 these transitions on social workers. The response is overwhelmingly negative. 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). Langer et al, 2019, study a large set of public sector workers and conclude that "when autonomy is reduced because of high levels of routineness and centralization in the workplace negative outcomes such as decreased job satisfaction result". For more studies corroborating this outcome, see Adler and Borys, 1996, Cartarelli et al, 2016, and Wright and Davis, 2003.

As a result of these concerns, it should not be surprising that the primary focus of many social organizations is to raise funds, an important part of which is keeping

[23]Concern about funding is even higher for small non-profits, and the size distribution of non-profits heavily skews towards small firms, with 67% of non-profits having annual expenditures less than $500,000 (Faulk et al, 2021).

[24]For example, Petersen et al, 2020, describe how "social workers have traditionally been given a great deal of autonomy to use discretion, but their decisions have increasingly been accused of being subjective and random". The classic reference is Lipsky, 1980.

the job attractive to their employees rather than some of the concerns of the market design literature.[25] Related to this, 29% of non-profits report "difficulty recruiting and retaining quality staff" as their biggest challenge, likely as workers are paid below market wages in the sector.[26]

2.2 Choice

Many social organizations offer subsidized choice to their clients.[27] We characterize Unrestricted Choice arising when each client's highest ranked service is $\hat{t}_i = \tau_i$ for any m_i, and all receive that choice. The key is that is must satisfy Fairness.

So far, we have simply described m_i as parametric, without offering any foundation for when (13) is likely to arise. We now do so by assuming Non-Increasing Absolute Risk Aversion, where $-\frac{V''(y)}{V'(y)}$ is non-increasing in y.[28] We also assume that $y_i = \hat{y} + \eta_i \geq 0$ and $E\eta = 0$. Average income is $\hat{y}$. Finally, let $y \sim (\hat{y} + \underline{\eta}, \hat{y} + \overline{\eta})$. With Decreasing Absolute Risk Aversion, Unrestricted Choice arises for average income above $\hat{y}^{uc}$, where

$$\frac{V'(\hat{y}^{uc} + \overline{\eta})}{V'(\hat{y}^{uc} + \underline{\eta})} \equiv \frac{\overline{m}^{uc}}{\underline{m}^{uc}} = \frac{2\mu - \underline{\phi}}{2\mu + \underline{\phi}}. \tag{19}$$

This outcome reflects the following simple intuition: when clients have low average income, small differences in income translate into large differences in utility. This declines as average income rises, which allows efficient screening here.[29] With Constant Absolute Risk Aversion, Choice can be used for any income level or for none.[30]

[25]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 skilled labor, where existing employees understand the technology and their impact on outcomes. 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 that he is offered on the job (for examples, see Liu at al., 2011, and Zhang et al., 2024).

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

[27]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-packet basket). See Food Bank News, 2024, for empirical evidence on how clients value choice. 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).

[28]How risk preferences vary with wealth has been extensively used in finance. See Chiappori and Paiella, 2011, Guiso and Paiella, 2008, and Huber et al., 2023, for empirical evidence supporting DARA

[29]With logarithmic utility $V(y) = ln(y)$, $\hat{y}^{uc} = \overline{\eta}(\frac{\mu}{\underline{\phi}} - \frac{1}{2}) - \underline{\eta}(\frac{\mu}{\underline{\phi}} + \frac{1}{2})$.

[30]We model m_i as variation in client income. This could alternatively be interpreted as savings,

2.2.1 Unrestricted Choice

Unrestricted choice arises when the equilibrium entails all clients of type τ receiving $I_i = 1$ of $t_i = \tau_i$. The mechanism for Unrestricted Choice is as follows. The firm knows I^{uc} when allocating services to all clients and offers $I_i = 1$ to all $i \leq I^{uc}$. The mechanism is:

1. The firm chooses $R^{uc} = \{r_0^{uc}, r_1^{uc}, r_2^{uc}\}$.
2. Each client $i \leq I^{uc}$ reports $\hat{t}_i$.
3. The firm allocates all clients their first choice if i is undersubscribed.
4. For all $t = i'$ that are oversubscribed, the firm implements a Deferred Acceptance Algorithm among those where $\hat{t}^k = i' \succ ...$

The final stage of the mechanism is only used out-of-equilibrium as all agents receive $t_i = \tau_i$ in equilibrium. Given Lemma 1, Proposition 4 follows.

Proposition 4 *Unrestricted Choice is used only if $\hat{y} \geq \hat{y}^{uc}$. Surplus from Unrestricted Choice is*

$$ES^{uc} = \frac{D}{\overline{U} - r^*}\left(v + T^* - \frac{\delta D}{2(\overline{U} - r^*)}\right) - \frac{\delta \sigma_\epsilon^2}{2(\overline{U} - r^*)^2}, \tag{20}$$

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

Wage are given by $\overline{U}$ as $\hat{U}^{ca} = \overline{U}^{ca}$. Prices are $R^* = \{r_0^*, \frac{\mu+\phi}{\underline{m}}, \frac{2\mu+\phi}{\underline{m}}\}$ from Lemma 1. The firm can induce efficient choice at $\hat{y}^{uc}$ with $r_0 = 0$. As average income rises above $\hat{y}^{uc}$, the firm can increase the price of the $t = 0$ service, until it reaches its maximum of $\frac{\phi}{\underline{m}}$ above average income of $\tilde{y}^{uc} > \hat{y}^{uc}$, where $\frac{V'(\tilde{y}^{uc}+\overline{\eta})}{V'(\tilde{y}^{uc}+\underline{\eta})} = 1 - \frac{\phi}{2\mu}$. Proposition 4 illustrates how Choice not only yields efficient assignment but also raises more revenue than either Bureaucratic Assignment or Deferred Acceptance.

Note the simplicity of the implications of the mechanisms for revenue. The technology has two parameters, μ and ϕ. Under Unrestricted Choice, revenue (normalized by $\underline{m}$) is $\mu + \underline{\phi}$ beyond $\tilde{y}^{uc}$, with Bureaucratic and Agent Assignment it is μ and under Deferred Acceptance, it is $\mu - \underline{\phi}$.

The purpose of the assumption of Decreasing Absolute Risk Aversion is to offer a dichotomy between those non-profits that serve poorer clients compared to wealthier ones. Those serving wealthier clients not only get more revenue - as they are willing to pay more - but it allows the firm to transition to Unrestricted Choice, which in turn raises more revenue as clients get what the want. By contrast, the capacity of

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.

those serving the poor is often so low that the firm further compromises allocative efficiency below that in Deferred Acceptance to raise funds. In this sense, a model that has no technological increasing returns to scale endogenously creates them.

2.2.2 Restricted Choice

When $\hat{y} < \hat{y}^{uc}$, Unrestricted Choice is not possible. We now show how Choice can arise for lower income levels. For illustrative purposes, first consider a case where the firm can degrade services, where instead of offering $I_j = 1$, it chooses $\kappa_j \in (0, 1)$ at no cost. Client utility becomes $U_i^c = (T(\tau_i, t_j) - m_i r_j)\kappa_j$. Consider the κ_j that are consistent with Choice for any given income level, where $I_0 = 1$, $I_1 = \kappa_1 < 1$, and $I_2 = \kappa_2 < \kappa_1$. We show in the Appendix that for $\hat{y} < \hat{y}^{uc}$ with per unit quality prices R^*, Choice satisfies the Fairness Constraint for any $\kappa_i \leq \kappa_i^*$ where

$$\kappa_i^* = \frac{\underline{\phi}}{i\mu - \frac{\overline{m}\underline{\phi}}{\underline{m}-\overline{m}}}, i = 1, 2. \tag{21}$$

Choice is between more of a less desired good and less of a more desired one, but where greater client income increases the κ_i^*. Expected treatment surplus in this setting is given by $\tilde{T}^{rc} = \frac{\hat{\phi}+\kappa_1^*(\hat{y})(\hat{\phi}+\mu)+\kappa_2^*(\hat{y})(\hat{\phi}+2\mu)}{3}$. This shows how the firm can implement $\tau_i = t_i$ but at the cost of distorting services.

In reality, market designs with divisible goods (described below) do not degrade services, but instead change quantities. Consider a case where the client chooses between *quantities* $I_0 = 1$, $I_1 = \kappa_1 < 1$, and $I_2 = \kappa_2 < \kappa_1$. The κ_i^* are identical. The only difference is that there is now residual supply of services $t = 1$ and $t = 2$ beyond client I^{rc} to be allocated. After client $i = I^{rc}$ there is no more remaining supply of $t = 0$, but there is $I^{rc}(1 - \kappa_i)$ of treatment i. The only extension to the example above is that we assume that firm uses a Deferred Acceptance mechanism to allocate the remaining $t = 1$ and $t = 2$.[31] The mechanism for distributing divisible quantities is described in the Appendix and the outcome is given in Proposition 5.

Proposition 5 *Let $\hat{y} < y^{uc}$ and $I^{rc} = \frac{D}{\overline{U}-r^{rc}}$, where r^{uc} is defined in the Appendix. Assume that $R = R^*$ for all clients $i \leq I^{rc}$. Surplus from Restricted Choice is*

$$S^{rc} = I^{rc}[(v + (\frac{1+\kappa_1^*+\kappa_2^*}{3})\left(T^* - \frac{\delta I^{rc}}{2}\right) + (\frac{1-\kappa_1^*}{3} + \frac{1-\kappa_2^*}{3})\left(\tilde{T}^{da} - \frac{\delta I^{rc}(1 + \frac{1-\kappa_1^*}{3} + \frac{1-\kappa_2^*}{3})}{2}\right)] - \frac{\delta\sigma_\epsilon^2}{2(\overline{U} - r^{rc})^2}, \tag{22}$$

[31] This could be replaced with any alternative without changing the conceptual outcomes.

where $\tilde{T}^{da}$ is the surplus from the Deferred Acceptance component, defined in the Appendix.

The outcome has two parts. All clients $i \leq I^{rc}$ choose efficiently $t_i = \tau_i$ and pay unit prices R^*. The remaining supply is allocated by Deferred Acceptance. Surplus and revenue are a combination of these two groups.[32]

This interpretation adds a new dimension to the problem: *client targeting*. All other mechanisms entail $I_i^A = 1$ for all $i \leq I^A$. Here all clients $i = 1, ..., I^{rc}$ receive the efficient type of treatment $t = \tau$, but $\tau \in \{1, 2\}$ receive lower than optimal quantities. While the additional supply of treatments $1 - \kappa_i$ is consumed by another client, these are lower priority clients. This is what we mean below by poor targeting of clients. These is the final part of the bracketed term in (22). We give empirical examples of such targeting issues below.

Restricted Choice screens on marginal rates of substitution. Given this, it should not be surprising that there is a relationship to competitive equilibrium. The outcome above implements competitive equilibrium with equal incomes (Foley, 1967). We show this in the Appendix, where the firm offers supply of $t_0 = \frac{I^{rc}}{3}$, $t_1 = \frac{\kappa_1^* I^{rc}}{3}$ and $t_2 = \frac{\kappa_2^* I^{rc}}{3}$ 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 $I_0 = 1$, $I_1 = \kappa_i^*$, and $I_2 = \kappa_2^*$. The reason for this extension is that competitive equilibrium with artificial currency is the most common market design alternative to Deferred Acceptance (Budish, 2011, Budish and Cantillon, 2012, Prendergast, 2017, 2022, Goedde, 2024, and Wong, 2024). This section offers a link between the two existing mechanisms typically used.

2.3 Partial Choice

Consider a setting where $\hat{y} < \hat{y}^{uc}$ and $I_i \in \{0, 1\}$, so neither Unconstrained nor Restricted Choice can be used. There is an alternative mechanism that involves choice, where the client can choose one t but where the alternative is an assigned mechanism over the other two services. This opens up the possibility of a region of Partial Choice. In the Appendix, we show that there is one other possibility for indivisible treatments involving Choice, where any client can choose the intermediate treatment $t = 1$. A Partial Choice outcome arises when only type $\tau = 1$ does so.

To illustrate this, we consider the case where the optimal form of Assigned Consumption is Bureaucratic Assignment.[33] We then compare that benchmark to one

[32]Proposition 5 assumes that the firm charges R^*. 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 r_0. This implies that there is one other possible equilibrium outcome, where $r_0 = 0$. This is done to increase κ_i above the levels in (21). This does not change the tradeoffs, and is described in the Appendix.

[33]This is for simplicity. The same calculus as above would more generally be used to determine this, but the qualitative results would not change.

where any client can choose an undersubscribed service, but oversubscribed ones are allocated by Bureaucratic Assignment. This can be implemented by a First Choice mechanism, which we now describe.

The firm knows I^{1st} when allocating t_i. It offers $I_i = 1$ to all $i \leq I^{1st}$. The First Choice Mechanism is:

1. The firm chooses $R^{1st} = \{r_0^{1st}, r_1^{1st}, r_2^{1st}\}$.
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 $\hat{t}^k = i' \succ ...$

Proposition 6 shows the outcome.

Proposition 6 *A Partial Choice equilibrium exists with $T^{1st} > T^b$ and $r^{1st} > r^b$ if $\hat{y} \geq \hat{y}^{1st}$, where $\hat{y}^{1st} < \hat{y}^{uc}$. $\hat{y}^{1st}$ is increasing in q.*

There are four parts to Proposition 6. First, a Partial Choice equilibrium exists for lower client income than is needed for Unconstrained Choice. 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. Second, allocative surplus is higher than with Bureaucratic Assignment. The higher welfare from the first choice mechanism arises as *all* type $\tau = 1$ clients choose $t = 1$. That non-strategy proof First Choice mechanisms can result in higher welfare is well known. 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. Our objective is not to reiterate this point.

Third, the First Choice mechanism raises more funding than Bureaucratic Assignment, so $EI^{1st} = \frac{D}{\overline{U} - r^{1st}} > EI^b$. This result is novel. In the Appendix we show that the outcome with higher revenue exists if

$$\frac{\overline{m}}{\underline{m}} > 1 - \frac{(1+3q)\underline{\phi}}{2\mu(3q-1)}. \tag{23}$$

Finally, this implies that required income depends on the quality of the firm's information. When $q = 1$ this is identical to the definition of $\tilde{y}^{uc}$ above, but otherwise is lower. For $q < q^* = \frac{2\mu+\underline{\phi}}{3(2\mu-\underline{\phi})}$, the Partial Choice outcome exists for any income level.[34]

[34]The relationship to q is intuitive. When q is high, type $\tau = 0$ clients almost surely will be assigned $t = 0$. In that instance, that client will be more likely to choose $t = 1$, which can only be deterred though higher prices.

2.4 The Non-Profit Outcome

We now aggregate our results. Outcomes vary in three ways: allocative efficiency, wages and service charges. First consider the comparison between Deferred Acceptance and Bureaucratic Assignment. Let $\Delta r = r^b - r^{da} = \frac{\phi}{\underline{m}}$ be the additional per client revenue from Bureaucratic Assignment. The firm chooses Bureaucratic Assignment for all $D \leq D_b^*$, where D_b^* is uniquely defined by

$$(T^{da} - T^b)D_b^* = \frac{D_b^* \Delta r}{(\overline{U} - r^{da})(\overline{U} - r^b)} \left(v + T^b - \frac{\delta(D_b^* + \frac{\Delta r}{2})}{2(\overline{U} - r^b)} \right) - \frac{\delta[(\overline{U} - r^b)^2 - (\overline{U} - r^d)^2]\sigma_\epsilon^2}{2(\overline{U} - r^b)^2(\overline{U} - r^d)^2}. \quad (24)$$

Similarly, the firm prefers Agent Assignment over Bureaucratic Assignment for all $D \leq D_a^*$, where D_a^* is uniquely defined by

$$(T^b - T^a)D_a^* = \frac{D_a^* \Delta w}{(w^* - r^b)(\overline{U} - r^b)} \left(v + T^a - \frac{\delta(D_a^* + \frac{\Delta w}{2})}{2(w^* - r^b)} \right) - \frac{\delta[(\overline{U} - r^b)^2 - (w^* - r^b)^2]\sigma_\epsilon^2}{2(w^* - r^b)^2(\overline{U} - r^b)^2}. \quad (25)$$

This shows how the firm eschews allocative efficiency if it is sufficiently short of resources, given by D_i^*.[35]

The other key variable for optimal Assigned Consumption is q, the quality of the firm's information. Let D^* be defined by $D^* + \underline{\epsilon} = \frac{v + T^{da}}{\delta}$. This defines the maximum amount of donations that the firm would ever use. (Assume that there is free disposal for any donations beyond the optimal level for a given assignment mechanism A.) Then Figure 1 offers a taxonomy of Assigned Consumption. Two features beyond (24) and (25) give rise to Figure 1. First, for q low enough, the firm always prefers Agent Assignment over Bureaucratic Assignment. The reason is that Agent Assignment raises more funds, but at the cost of using worse information to assign treatments. As q falls, the value of better information declines, and goes to 0 as $q \to \frac{1}{3}$ (the D_a^* line in Figure 1). Hence, Agent Assignment is preferred to Bureaucratic Assignment for q low enough in Figure 1. Second, for q high enough, the firm always prefers Bureaucratic Assignment over Deferred Acceptance. The value of Deferred Acceptance here is to incorporate client preferences that are not reflected in the firm's signal. Remember from (15) that the value of Deferred Acceptance (the p_i terms) were more valuable when the firm is less well informed. As q increases, the

[35] These two outcomes show that lack of resources lead to assignment mechanisms that do not seek client input. We have not determined that $D_i^* > 0$ so it may be that Deferred Acceptance is the preferred option. This would arise for instance if $\alpha \to 0$ in which case wage savings from Agent Assignment are minimal, or when δ is large enough so the additional revenue from Bureaucratic Assignment is of little value. However, it illustrates a ranking of mechanisms that seek client input over those that do not.

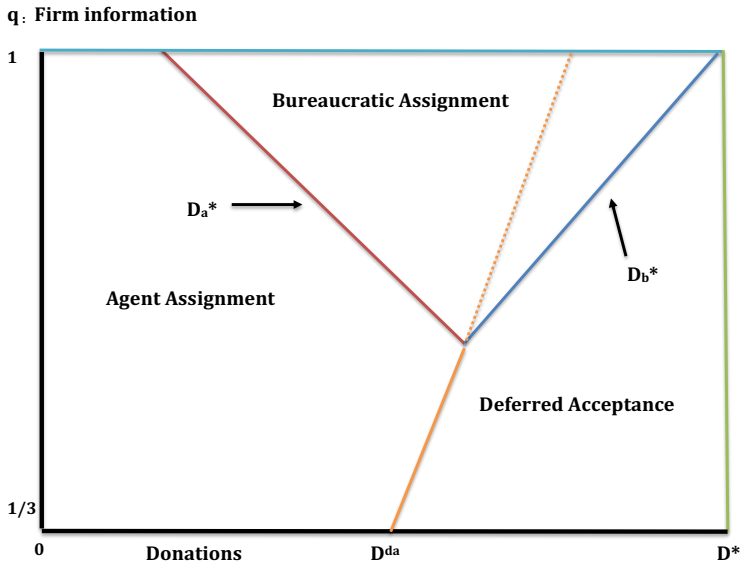


Figure 1: **Optimal Assigned Consumption.**

value of the additional information provided by client rankings falls and goes to 0 as $q \to 1$ (the D_b^* line in Figure 1). Hence, Bureaucratic Assignment is preferred to Agent Assignment for q high enough in Figure 1.[36]

More generally, this shows how poorly resourced non-profits allocate resources in ways that are allocatively inefficient. This is not through incompetence, malfeasance or increasing returns to scale, but because their primary concern is funds. It also shows how Deferred Acceptance arises among Assigned Consumption mechanisms only when (i) funding is not a significant concern, $D \geq D_b^*$, (ii) the firm is poorly informed (q low), and (iii) where clients have sufficient heterogeneity in preferences. Given (i), it is striking that funding concerns for core market design settings are minimal, occurring through taxation (school choice, refugee assignment) or where the potential funds received are trivial compared to overall funding for the institutions (medical resident matching and kidney exchange).

Funding for client services has two components: service charges and compensation. Many social organizations provide services to clients who cannot pay. Consider the case where $\underline{m} = \infty$, so pricing plays no role. In this case, the firm never uses Bureaucratic Allocation, whose only benefit over Deferred Acceptance is more client revenue. However, it remains the case that the firm will choose Agent Assignment over Deferred Acceptance for low funding levels, as (25) is merely evaluated at $r^b = 0$. As a result, the insights here do not depend on pricing services to clients. The outcome in

[36] An even starker outcome arises as h varies, measuring the extent to which client preferences potentially offer valuable information for allocations. Specifically, as h gets small, the region of Figure 1 where Deferred Acceptance is used gets smaller, vanishing as $h \to 0$.

this instance is given by the hashed line in Figure 2, where the firm chooses between Agent Assignment and Deferred Acceptance.[37]

Figure 2 assumes that the firm uses Assigned Consumption. Choice is the alternative. If $\hat{y} \geq \hat{y}^{uc}$ the firm chooses Unrestricted Choice over either Bureaucratic Assignment or Deferred Acceptance. This has the implication that social organizations that are targeted at less poor clients benefit in three ways: not only can they charge more because clients are better off, but they can allow choice, which in turn further allows more revenue generation and eliminates allocative error. This further limits the use of Deferred Acceptance to settings where it is impossible to design prices to screen clients, which arises with Decreasing Absolute Risk Aversion only for low income clients. For wage reductions large enough, it may continue to prefer Agent Assignment if $\underline{m}$ is high enough, as funds may be low.

We have not offered a rubric to compare the other two mechanisms. With Restricted Choice, this is because the comparison occurs along a different dimension, client targeting. We have also not compared Assigned Consumption to Partial Choice because we have not addressed the costs explored in the market design literature. Specifically, the mechanism is not strategy proof.[38] For this reason, we feel that a comparison of Partial Choice is inappropriate. Instead, our purpose there is to show how shortness of funding may lead firms to offer non-strategy proof mechanisms despite these problems.

At the outset of the paper, we described the two canonical ways of transferring services to clients as matching without prices and markets with prices. Here we have addressed matching with (fair) prices. Depending on donations, client income, firm information and diminishing returns, a broad range of outcomes involves combinations of the the assignment ethos of the market design literature and the choice component of markets. Yet he link between allocative efficiency and funding nullifies the priority typically offered to Deferred Acceptance, offering a wider set of mechanisms consistent with the concerns of the social organization.

2.5 Observations on the Non-Profit

Here we discuss some modeling choices and their robustness.

[37] This allows us to relate our work to the existing market design literature, where Choice may be inefficient because the firm has different objectives over t_i. For instance, schools care about the composition of classes that cannot be resolved by client willingness to pay. This replicates the outcome of the preferences for Deferred Acceptance in the absence of choice for a firm that is well enough funded.

[38] Above we described the market design literature as prioritizing Deferred Acceptance. A smaller literature addresses the merits of First Choice mechanisms, where similar effects on efficiency have been found. 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.

More General Technology: The ability of the client to reveal information was characterized by two results above: (i) Unrestricted Choice can only be used for income high enough, and (ii) Deferred Acceptance reduces revenue. This was shown using preferences in (1). These preferences distinguished between a fraction h of types with horizontal preferences $\overline{\phi}$ and $1-h$ with vertical preferences $\underline{\phi}$. The more general analog to this is where $T(k,j)$ for the horizontal types is $\overline{\mu}_{kj}$, and $\underline{\mu}_{kj}$ for the vertical types. Then the analog to Assumption 1 is then (i) $\overline{\mu}_{kk} \geq \overline{\mu}_{kj}$, and (ii) $\underline{\mu}_{k2} \geq \underline{\mu}_{k1} \geq \underline{\mu}_{k0}$.

It remains the case that Deferred Acceptance has higher allocative efficiency than Bureaucratic Assignment or Agent Assignment. The more relevant issue is how this affects the capacity of clients to reveal information. Screening satisfies the Fairness Constraint in the analog above for Unrestricted Choice with prices $R^* = \{\frac{\underline{\phi}}{\underline{m}}, \frac{\underline{\mu}+\underline{\phi}}{\underline{m}}, \frac{2\underline{\mu}+\underline{\phi}}{\underline{m}}\}$ if

$$\frac{\overline{m}}{\underline{m}} \geq max\{\frac{\underline{\mu}_{12}-\underline{\mu}_{11}}{\underline{\mu}_{22}-\underline{\mu}_{11}}, \frac{\underline{\mu}_{02}-\underline{\mu}_{00}}{\underline{\mu}_{22}-\underline{\mu}_{00}}, \frac{\underline{\mu}_{01}-\underline{\mu}_{00}}{\underline{\mu}_{11}-\underline{\mu}_{00}}\}. \tag{26}$$

With the parameterization of preferences in (1), the binding constraint was $\frac{\underline{\mu}_{02}-\underline{\mu}_{00}}{\underline{\mu}_{22}-\underline{\mu}_{00}}$, where the "least deserving" $\tau = 0$ clients benefited most from deviating to claim $t = 2$.[39] In general, the outcome remains that Choice depends on client income being high enough,

Second, consider the impact of Deferred Acceptance on revenue relative to Bureaucratic Assignment. Individual rationality requires that maximum revenue is $r_j = min_i \underline{\mu}_{ij}$ so $r^{ac} = \sum_j min_i \underline{\mu}_{ij}$. Let $r_0 = min_i \underline{\mu}_{0i}$. Then in order for a client to rank $2 \succ 1 \succ 0$, incentive compatibility implies that the maximum prices for $t = 1$ is $r_1 = min_i(\underline{\mu}_{i1} - \underline{\mu}_{i0}) - r_0$ and $r_2 = min_i(\underline{\mu}_{i2} - \underline{\mu}_{i1}) - r_1$. For Deferred Acceptance to not reduce income, it must be then that $min_k \underline{\mu}_{k,j} - min_k \underline{\mu}_{k,j'} = min_k(\underline{\mu}_{k,j} - \underline{\mu}_{k,j'})$ can only arise when (i) μ_{ij} is increasing in j for all i, j, and (ii) $\underline{\mu}_{(i+1)j} - \underline{\mu}_{ij}$ is increasing in j.

This implies that for Deferred Acceptance not to reduce income, it must be that $\tau = 0$ has both the lowest value of $t = 0$ and the lowest marginal return to higher t for all t. If this arises, the firm never uses Bureaucratic Assignment. Yet the assumption is strong: in the example earlier, it implies that everyone prefers New York to Boston to Chicago, and the type with the lowest valuation for Chicago also has the smallest increase in utility from another other city.[40]

[39] It is straightforward to construct examples where the binding constraint is for another agent. For example if $T(k,j) = (k+1)(t+1)$, the binding constraint is for agent $\tau = 1$ reporting $t = 2$

[40] It is possible to generate examples where preference revelation does not reduce revenue. $T(t,\tau) = (\tau + k)(t + k')$, $k, k' \geq 0$ satisfies these conditions. However, if all clients have such preferences, there is no horizontal differentiation and all clients rank $t = 2$ as their first choice, and Deferred Acceptance has no value in revealing client preferences.

Uncertainty over Ability to Pay: We have interpreted ability to pay in terms of income, but it could alternatively represent variation in client assets, access to credit card debt, or capacity to receive a loan from friends or family. A small number of large non-profits have access to the financial position of clients, primarily colleges which require tax forms and information on assets from the families of applicants. The ability to monitor income income and assets reduces funding concerns, where $\underline{m}$ above now reflects the average lower bound of client capacity to pay given all observable data. Yet few non-profits have the capacity to both receive and process such information into price gradients by income, as they are typically too small. For example, 88% of non-profits have annual expenses of less than $500,000.[41] Even those that do means testing for clients typically only use if as a maximum income exclusion for high earning potential clients. For example, Gai et al., 2021, document how non-profit hospitals use income data only to determine eligibility caps for discounted care, typically around 400% of the Federal Poverty Level.

Targeting and Restricted Choice: Above we described the importance of client targeting under Restricted Choice. We now describe two examples. First, Feeding America allocates food using an artificial currency that food banks use to bid. They 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.

The targeting issue is that under the old allocation mechanism which entailed queueing, a food bank received 1 pound of cereal. Now it gets $\frac{1}{50}$ of a pound. Where does the other $\frac{49}{50}$ go, or more generally how does the Restricted Choice mechanism reallocate food across the national network? During the redesign at Feeding America, the designers cared greatly about who would benefit in terms of more pounds of food and who would get less. There was clearly a desire that the amount of food would be reoriented towards smaller, poorer food banks.[42] 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

[41] Two additional issues arises in the college example. First, our focus is on settings where the failure to use Choice is lack of client income. This is not the case with colleges, which are closer to the school choice environment. Second, many of the concerns about fairness are mitigated in this setting, because the supply of colleges is so large, where if an applicant is not accepted to one college, there are many other close substitutes where the applicant may be accepted. Finally, it is worth reiterating that very few non-profits have the capacity to monitor income and assets of clients as they are too small.

[42] 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.

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 over 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 are close to expiry while $t = 2$ vaccines were far from their expiry date. Countries varied in their capacity to quickly distribute vaccines. As an example, Turkey's health care system was far more capable of responding than that of the typical West African nation. The logic of economic efficiency suggests 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.[43]

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 mechanisms are 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.

There is one 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. By definition, targeting is then harmed by Restricted Choice, the final two terms in (22). In reality, political considerations implied that both Feeding America and Covax had to offer equal access to all food banks or to all participating countries. Yet Feeding America's priorities were to reallocate food to poorer food banks, and Covax wished to vaccinate the poorest countries (and to be perceived to do so). This is why targeting improved for Feeding America. The purpose of this section is not

[43]This led to well-publicized complaints from the Organization of African States and the Africa CDC Chief that "the majority of the donations have been ad hoc, provided with little notice and short shelf lives" (Al Jazeera, 2022, Organization of African States, 2021). Anecdotes abound: South Sudan destroying 60,000 vaccines, or the Democratic Republic of Congo returning 1m vaccines close to its expiration. It is also reported that 681m vaccines remained unused in low income countries (Al Jazeera, 2022).

to match the outcome tightly to the model but instead to illustrate how impacts on targeting have a critical influence on whether the mechanism is implemented.

The Role of the Fairness Constraint: The outcomes above are premised on the assumption that no client can be denied coverage for lack of income. This has two components: (i) whether the client is served, and (ii) what they are served. Relaxing the Fairness Constraint affects each of these differently.

First, some social organizations may feel that they have no choice but to exclude some low income clients in order to serve more clients.[44] Consider a Relaxed Fairness Constraint where it treats all clients that it serves equally, but that it only covers clients whose income exceeds $\hat{y}+\eta^* \geq \hat{y}+\underline{\eta}$. This changes (i) above without changing (ii). Let $V'(\hat{y}+\eta^*) = m^*$, where the firm chooses m^*, the lowest income client that it covers. Assume that a fraction $G(m^*)$ clients have income between $\underline{m}$ and m^*, where $G(\underline{m}) = 0$ and $G' \geq 0$.[45] For illustrative purposes, consider the case where the firm uses Unrestricted Choice to allocate services. If the firm chooses $m^* > \underline{m}$, it charges prices of $\tilde{R}^* = \{\tilde{r}_0^*, \frac{\underline{\phi}+\mu}{m^*}, \frac{\underline{\phi}+2\mu}{m^*}\}$, where $\tilde{r}_0^* = \min\{\frac{\underline{\phi}-2\mu}{m^*} + \frac{2\mu+\underline{\phi}}{m^*}, \frac{\underline{\phi}}{m^*}\}$ with revenue of $\tilde{r}^{uc} = \frac{\tilde{r}_0^* + \frac{\underline{\phi}+\mu}{m^*} + \frac{\underline{\phi}+2\mu}{m^*}}{3}$. Surplus is

$$ES^{uc} = \frac{D}{\overline{U}-\tilde{r}^{uc}}\left(v + T^* - \frac{\delta\frac{D}{1-G(m^*)}}{2(\overline{U}-\tilde{r}^{uc})}\right) - \frac{\frac{\delta\sigma_\varepsilon^2}{1-G(m^*)}}{2(\overline{U}-\tilde{r}^{uc})^2}, \tag{27}$$

This illustrates the firm's tradeoff. Excluding some clients increases total coverage, by $\frac{D(\tilde{r}^{uc}-r^{uc})}{(\overline{U}-\tilde{r}^{uc})(\overline{U}-r^{uc})}$. The problem is that it covers the wrong clients. With probability $G(m^*)$ a client is skipped in the priority ranking, so that while $\frac{D}{\overline{U}-\tilde{r}^{uc}}$ clients are covered, their average value is $v + T^{da} - \frac{\delta\frac{D}{1-G(m^*)}}{2(\overline{U}-\tilde{r}^{uc})}$, lower than when clients are not excluded. With the usual Monotone Likelihood Ratio property on G this becomes a variant of the traditional monopoly problem where here higher revenue is traded against lower coverage. For δ high enough, the firm will cover its highest priority clients over expanding coverage, which is the premise that underlies much of the analysis here. Note that this extension does not change the conceptual outcomes above, as fairness must hold for all those who are covered.

Second, a useful benchmark is where there is *No Fairness Constraint*. Here the firm can deviate from both (i) and (ii) above. We consider this in a two step fashion, where the firm first chooses all clients $m \leq m^*$ that it serves, as above, but now it can offer treatments to clients that are income dependent. The choice of m^* is as in (27) except that the surplus from the allocation mechanism changes and is given by $\tilde{T}^A$.

[44] As an example, private housing associations will not rent units to those who cannot pay at all, but instead have a minimum rent. Antecol and Smith, 2012, and Murphy, 2023, also show that the use of Early Decision in university admissions excludes lower income students.

[45] As there are an infinite number of potential clients, this includes the revenue maximizing choice by the firm, where it only serves clients of type $\overline{m}$.

The choice of prices and assignment mechanisms is simplified. It is straightforward to show that under Assigned Consumption, optimal prices are $\tilde{R}^{AC} = \{\tilde{r}_0^*, \frac{\underline{\phi}}{m^*}, \frac{\underline{\phi}}{m^*}\}$ under Assigned Consumption and $\tilde{R}^* = \{\tilde{r}_0^*, \frac{\underline{\phi}+\mu}{m^*}, \frac{\underline{\phi}+2\mu}{m^*}\}$ under Unrestricted Choice. As above, the firm chooses Unrestricted Choice for income above $\tilde{y}^{uc}$. However, the key difference is that the firm never chooses Bureaucratic Assignment. This is because both revenue and allocative efficiency are higher under Deferred Acceptance.

Finally consider the case when *Fairness is not Lexicographic*. Let the firm increase r_1 and r_2 from $\frac{\mu-\underline{\phi}}{\underline{m}}$ and $\frac{2(\mu-\underline{\phi})}{\underline{m}}$ respectively to prices no higher than $\frac{\mu}{\underline{m}}$ and $\frac{2\mu}{\underline{m}}$ (as these bounds are necessary to satisfy Individual Rationality). For $r_1 \in (\frac{\mu-\underline{\phi}}{\underline{m}}, \frac{\mu}{\underline{m}}]$ and $r_2 \in (\frac{2(\mu-\underline{\phi})}{\underline{m}}, \frac{2\mu}{\underline{m}}]$, outcomes are unfair. $1 - G(m_i^*)$ of clients with type $\tau = i$ report a first choice ranking of $t = i$ and the remainder report a first choice ranking of $t = 2$, where $m_1^* = \frac{\mu-\underline{\phi}}{r_1}$ and $m_2^* = \frac{2\mu-\underline{\phi}}{r_2}$. The cost of unfairness is now assumed to be $\Gamma(1 - G(m_1^*), 1 - G(m_2^*))$, where Γ is not lexicographic at 0, but continuously increasing in each argument. We show in the Appendix that (i) the set of prices that maximizes allocative efficiency is $r_1 = \frac{\mu}{\underline{m}}$ and $r_2 = \frac{2\mu}{\underline{m}}$, but (ii) the firm does not choose these for Γ' high enough. As a result, the idea that Deferred Acceptance reduces client revenue continues to hold with non-lexicographic preferences.

Non-Strategy Proof Assignments: Formal first Choice mechanisms are not common. However, we argue that social organizations routinely offer non-strategy proof choices akin to the mechanism above. This arises through presenting a client with a non-strategy proof choice between a less-than-optimal option for sure and a random alternative generated by the allocation process. We give two examples.

First, social housing is typically allocated through waiting lists, where a client is matched to a house as it becomes available.[46] The house may not be the client's most desired type or in her preferred location. She can turn the offer down. The offer typically comes with a cost: if the client turns down a "reasonable" offer, she is placed at the lower on waiting list going forward.[47] As a result, the client needs to strategize about whether to accept. Second, the process for applying to universities in the US has two distinct options. Students can apply through Regular Decision, where the student is not committed to a school if accepted. Students also have an earlier option to identify a single school to apply to via Early Decision. If the student is accepted by that school, she is obliged to accept the offer.[48] Among elite colleges, probabilities

[46]See Leshno, 2022, for a price-theoretic approach to addressing housing allocations through waiting lists.

[47]With repeated turn-downs, she may be be taken off the list, but this does not happen automatically. For example, the Citizens Advice Bureau for housing in Britain advises "refusing could also make your situation worse, for example if your local council reduces your points or removes you from their priority list" (https://www.citizensadvice.org.uk/).

[48]This is not a legally binding commitment, but is treated as binding by most students. Avery and Levin, 2011, have shown that Early Decision can be used to screen students on their enthusiasm

of acceptance are often four times higher than with Regular Decision and the use of ED has increased significantly over the last decade (Murphy, 2022). Despite not being designed as a First Choice mechanism above, students choose between a more certain option and a random alternative.

The conceptual novelty here is not that Partial Choice mechanisms may be more efficient, but that they can generate more funding. It is well known that Early Decision students pay higher average tuition (Antecol and Smith, 2012, Murphy, 2023). Raising revenue is through Partial Choice is more important when firms are short of resources. White, 2024, provides data that colleges with lower endowments (the analog to D in the model) fill a larger fraction of their class through Early Decision, consistent with the model.

3 The Social Enterprise

So far we have described outcomes conditional on non-profit status. Many social organizations instead operate as for-profit Social Enterprises.[49] This ability to act as residual claimant is valuable here. Each of the surplus calculations above has a term of the form $\frac{\delta\sigma_\epsilon^2}{2(w^A-r^A)^2}$, representing the cost of randomness in funding. The calculation of efficient outcomes in (10) eliminates this, by allowing the firm to be residual claimant $\pi(\epsilon)=\epsilon$, thereby providing insurance. As $\pi=0$ for the non-profit, this cost could not be avoided up to now. Then if $\mathbb{D}$ is independent of Π, a firm motivated by client welfare should organize as a for-profit.

We now show that with reasonable preferences for donors, the Social Enterprise can only exist if it is fully client-funded. This is because donors do not give to for-profit firms, even those who overwhelmingly use funds on mission related activities.

Donors: Any realistic model of donors must take account of their limited information.[50] We make three assumptions. First, donors do not observe assignment mechanism or contracts nor the actions of other donors. Instead, all they observe is Π the profit status of the firm. Second, donations are given by $\mathbb{D}(b)$, where b is a donor's belief of the probability that *their* donation is spent on I, the mission of the firm. Finally, we assume that donor beliefs about b are correct in equilibrium.

Let $\pi(\epsilon)\equiv\pi(\tilde{I}|\mathbb{D}(b),w^A,r^A)$ be the net profit of the firm given $\epsilon,\mathbb{D}(b),w^A$ and r^A, and let $\pi^*(\epsilon)$ be the equilibrium contract, where $\int_{\underline{\epsilon}}^{\overline{\epsilon}}\pi^*(x)f(x)dx=0$. All donors give holding the donations of others constant. As a result, they evaluate their donation at

for a school in the same way as described above.

[49]For instance, there are 1,300 for-profit hospitals and 1,100 for-profit charter schools in the United States (American Hospital Association, 2024, Network for Public Education, 2023). More generally, the World Economic Forum has estimated that there are 10 million such enterprises worldwide, accounting for 3% of world employment (World Economic Forum, 2024).

[50]This is manifested in the literature on whether donations are responsive to firm efficiency, where the outcome is ambiguous. See Buchheit and Parsons, 2006, Parson, 2007, and Yoruk, 2013.

the use of the marginal dollar received by the firm. Then for a realization ϵ, a fraction $1-\pi^{*'}(\epsilon)$ of the marginal dollar is spend on I rather than on π. As b is assumed to be correct in equilibrium, Lemma 4 follows.

Lemma 4: $b = \int_{\underline{\epsilon}}^{\overline{\epsilon}} (1-\hat{\pi}^{*'}(x)) f(x)dx.$

Now consider the outcome for an allocation mechanism A. Above we distinguished between donor driven organizations where $I^A = \frac{\mathbb{D}-\pi+\epsilon}{w^A-r^A}$ and self-funded organization, where $w^A \leq r^A$. In that latter case, client coverage is not constrained by funding, and $I^A = I^{A*} = \frac{v+T^A}{\delta}$. If $\Pi = 0$, by definition $\pi = 0$, so $b = 0$ and the outcome is given by (9). If $\Pi = 1$, the firm chooses π^*, but does so holding donations fixed as donors cannot observe contracts. Surplus is

$$ES = \int_{\underline{\epsilon}}^{\overline{\epsilon}} \frac{(\mathbb{D}(b|\Pi) - \pi^A(x) + x)}{w^A - r^A}(v + T^A - \frac{\delta(\mathbb{D}(b|\Pi) - \pi^A(x) + x)}{2(w^A - r^A)})f(x)dx. \quad (28)$$

This simplifies to

$$ES = \frac{\mathbb{D}(b|\Pi)}{w^A - r^A}(v + T^A - \frac{\delta\mathbb{D}(b|\Pi)}{2(w^A - r^A)}) - \int_{\underline{\epsilon}}^{\overline{\epsilon}} \frac{\delta(-\pi(x)+x)^2}{w^A - r^A} f(x)dx. \quad (29)$$

Proposition 7 immediately follows.

Proposition 7 *The Social Enterprise chooses* $\pi^{*'}(\epsilon) = 1$. *As a result,* $b = 0$ *and* $\mathbb{D}(0) = 0$.

Proposition 7 shows starkly how the Social Enterprise cannot attract donations.[51] As donors do not observe the contract, the firm provides full insurance to the mission. However, as b is correctly inferred in equilibrium, $(b|\Pi = 1) = \int_{\underline{\epsilon}}^{\overline{\epsilon}} (1-\hat{\pi}^{*'}(x)) f(x)dx = 0$. The central conclusion of Proposition 7 is then the impossibility of a for-profit charity.[52] Note that viewing the non-profit through this lens offers an additional implication, namely, that it has to live "hand to mouth", never sure of its ability to cover its mission. Being "hand to mouth" is necessary to encourage donations, as a donor must know that her donation affects the mission.[53]

This describes the hurdle that the firm must overcome to become a Social Enterprise, despite the efficiency of residual claimant status. Non-profits rely on donors for 22% of funding. Donors overwhelmingly report that their motivation is the impact

[51] While Social Enterprises are typically funded by clients, it is not unusual for foundations or states to offer seed money to allow these firms to get off the ground.

[52] Posner and Malani, 2006, have argued the merits of the for-profit charity.

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

that their gift has on the mission of the institution.[54] This is more pronounced for small non-profits, where 30% of income comes from donors (Tomasko et al., 2021). The result above shows that the Social Enterprise must make up that difference.[55] At the outset of the paper, we described a complex relationship between funding and efficiency in the social sector. Providing financial certainty to the firm increases expected surplus, but eliminates the funding that comes from donors. Once again, efficiency and funding are substitutes.

The central message of this section is that donors are unlikely to give if their gift is used for purposes other than the mission of the firm. This has long been a concern in the charitable sector. We have assumed that $\pi' = 0$ in order to qualify as a non-profit. In reality, the legal standard is not so strict, where some degree of incentive pay is allowed.[56] Compensation among fund raisers is governed by ethical standards set down by the Association of Fundraising Professionals. That ethical standard "prohibits members from working for percentage based compensation" as it "undermines the integrity of the voluntary sector". Specifically, "donor attitudes can be unalterably damaged in reaction to... the awareness that a commission will be paid for a fundraiser from his or her gift, thus compromising the trust on which charity relies". (American Association of Fundraising Professionals, 2016). In a similar vein, Sandberg, 2006, proposes appropriate legal transparency rules for pay-for-performance for charities. His concern is "that most donors would see a high fundraising commission as unreasonable and illogical irrespective of its economic rationale. If fully informed, many donors would probably refuse to contribute". As a result, it is widely believed in the sector that using donations as marginal compensation deters donors.[57]

There is a simpler behavioral interpretation of this outcome. Imagine that donors simply do not give to for-profit firms as they fear that their donation will be used to enrich the owners of the firm. That would offer a similar outcome. The purpose of this section is that (i) rational donors should value their gift at the marginal rather than average dollar, (ii) they correctly choose not to give if $\pi^{*'}(\epsilon) = 1$, and (ii) there

[54]This is called pure altruism in the literature. Ottoni-Wilhelm, et al., 2017, argue for the importance of such altruism for charitable giving. Survey evidence suggests that "97% of donors cite the impact of their gift as their main reason for giving" (CCS, 2024). This section extends the logic of Prendergast, 2023.

[55]In that case, coverage is at its efficient level, $I^* = \frac{v+T^A}{\delta}$. In settings where $\frac{r^A}{m} \geq w^a$, the model is silent on equilibrium pricing. The firm above only cares about surplus, and so is indifferent about any prices between w^a and $\frac{r^A}{m}$.

[56]The law on the non-profit allows "reasonable" pay-for-performance for their employees (Founding Church of Scientology v. United States).

[57]Here the donor makes an undirected donation. It is notable that social enterprises often attempt to attract donors through crowdfunding campaigns for a specific initiative. These serve to both target a mission-related initiative and to offer the prospect of the donor seeing herself as marginal. Though success rates are low, see Farhoud et al., 2021, for examples of some successes.

is a simple reason why the firm will indeed choose $\pi^{*'}(\epsilon) = 1$.

3.1 The Non-Profit Choice

Firms choose whether to incorporate as non-profit or for-profit. Assume that the firm can self-fund with mechanism A. Then the firm chooses a non-profit with mechanism A' over a self-funded one with A if

$$ES^{A'} \geq \frac{(v + T^A)^2}{2\delta}. \tag{30}$$

As $I^{A*} = \frac{v+T^A}{\delta}$ surplus from a self-funded firm is $\frac{(v+T^A)^2}{2\delta}$. Yet it may be that the only way to self-fund is with an inefficient assignment mechanism A. For example, it could be that only with Agent Assignment is $w^* < \underline{r}$. We now use (30) to address differences in how non-profits and Social Enterprises allocate services.

Proposition 8 *Assume that $w^* > \underline{r}$. Then if*

$$\frac{\mu - \underline{\phi}}{V'(\hat{y}^{uc} + \underline{\eta})} < \overline{U} < \frac{\mu + \frac{r_0^* + 2\underline{\phi}}{3}}{V'(\hat{y}^{uc} + \underline{\eta})} \tag{31}$$

all Social Enterprises use Unrestricted Choice and no Assigned Consumption firm does.

This follows trivially from the definition of $\hat{y}^{uc}$, when the firm can transition to Unrestricted Choice. Client payments increase discretely from $\frac{\mu - \underline{\phi}}{V'(\hat{y}^{uc} + \underline{\eta})}$ to $\frac{\mu + \frac{r_0^* + 2\underline{\phi}}{3}}{V'(\hat{y}^{uc} + \underline{\eta})}$. Then if the extra revenue allows the firm to cover its marginal cost $\overline{U}$, the Social Enterprise allows Choice but as $w^* > \underline{r}$, no Assigned Choice non-profit does.

People often complain about bureaucracies being unresponsive to their needs. This is contrasted with the "customer is always right" ethos of the typical for-profit setting. Proposition 8 provides a condition where this arises. More generally, the ability to self-fund is a comparison between the cost of the agent w^A and the ability to raise funds r^A. Let $\tilde{m}^A$ be the marginal utility of the poorest client such that $w^A = r^A$. This is given by the arrowed line in Figure 2, where we trace outcomes based on where that line lies. In general, the firm can be self-funded for any assignment mechanism if $\overline{U}$ is low enough: (31) shows the case where Choice is only observed among Social Enterprises. In Proposition 10 in the Appendix we show that while for-profit firms can exist with any allocation method, it more likely when client choice is involved. In this sense, it is a generalization of (31) and the view that non-profit firms are inherently more likely to make allocative errors.

4 Extensions

Diminishing Returns: We have assumed that diminishing returns are linear: $v_i = v - \delta i$ to allow the impact of uncertainty in funding to be separable from funding levels. If instead we do not make this assumption, the analog to (9) is

$$ES^* = \frac{D}{w^A - r^A} T^A + \sum_{i=1}^{\frac{D}{w^A - r^A}} v_i - \int_{\underline{\epsilon}}^{\overline{\epsilon}} \frac{(x - \pi(x))(v_{D+x-\pi(x)} - v_D)}{w^A - r^A} f(x)dx. \tag{32}$$

This changes the outcome to one where the cost of uncertainty depends on client coverage, but does not change any conceptual results above.

Uncertainty: The support of ability to pay when we addressed Unconstrained Choice does not change with average income. This led to the definition of $\hat{y}^{uc}$. More generally, $\overline{\eta}$ and $\underline{\eta}$ may depend on $\hat{y}$. Let η 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 $\hat{y}$. First, empirical evidence suggests that among low earners, dispersion of earnings is decreasing in average income (Farahan, 2015). If this translates to the support of outcomes, Decreasing Absolute Risk Aversion is sufficient but not necessary. Second, when randomness scales in a linear way with $\hat{y}$, $\eta(\hat{y}) = \eta\hat{y}$, the outcomes is analogous to finance settings that scale with investor wealth. Then our outcome arises 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.[58]

Interpreting the Client's Utility Function: The client's utility in (3) models preferences for t as orthogonal to the marginal utility of the outside composite good, m. The interpretation here is that clients do not consume t. This could be due to an income effect in preferences or there is an indivisibility in consuming t which stops the client from equating the marginal utility of consumption for all goods. Assume that the client would only choose to consume the treatment if $y \geq y^* > \hat{y} + \overline{\eta}$. The interpretation of τ is then the treatment that the client would choose in his consumption bundle for income higher than y^*.

A Managerial Implication: The logic of Agent Assignment is that the firm caters to worker preferences as a conduit to increasing client coverage. Consider instead a welfare weight approach to this problem where the firm's objective is to maximize

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

$\beta EU_i^c + (1-\beta)U^a$, subject to the worker being paid according to (6). How should the firm prioritize the interests of its employees relative to the average client? The implication of the analysis above is that for $\hat{y} < y^*$, the firm should choose $\beta = 0$ and for $\hat{y} \geq y^*$, the firm should choose $\beta = 1$. This proposes a managerial policy where small social organizations prioritize employee interests but larger firms reorient towards client interests.

Balanced Demand and Supply: We have assumed that the supply of services matches the distribution of types. Consider an example where the distribution of types $\{\tau_0, \tau_1, \tau_2\}$ is not uniform but instead $\{\frac{1-\zeta_1-\zeta_2}{3}, \frac{1+\zeta_1}{3}, \frac{1+\zeta_2}{3}\}$ for $\zeta_i > 0$. Then Unrestricted Choice does not clear the market, as there are more type $t = 0$ slots than there are matching clients. This has two implications. First, there is an uninteresting form of assignment when demand and supply are not balanced, namely that the firm must ration under Choice, where a client who "chooses" $t = 1$ or $t = 2$ would not receive it for sure, simply as there are too many clients of those types. This would entail some version of "first come, first served" or random rationing. Our interest in Assigned Consumption is in outcomes that incorporate the information or preferences of the firm (or agent). By considering balanced outcomes, we avoid this uninteresting form of assignment. Second, Choice above is strategy proof. This may not be the case when demand and supply and unbalanced. For example, let $\zeta_2 = \frac{1}{3}$ and $\zeta_1 = 0$. Then a client of type $\tau = 2$ who "chooses" $t = 2$ only receives it with probability $\frac{1}{2}$, and may prefer to choose $t = 1$ instead as she can get that for sure. These issues are secondary to our concerns so have been ignored.

Price Discrimination: Note one restriction when the Fairness Constraint does not hold. We have not allowed for price discrimination across clients. Consider an alternative where for any equilibrium above, the firm now offers uncovered clients services at higher prices than in the equilibrium. The appeal of this is that the firm covers more clients, but at the price of unfairness that some rich clients are served while poorer ones with higher priority are not. More generally, the firm could offer high priority clients service at lower prices than low priority clients. We have ruled this out.

Combinations of Mechanisms: We have assumed that the firm can only use a single mechanism for all clients. This is for the sake of realism where it is likely to be difficult to communicate to both clients and agents if combinations of mechanisms are used. However, if such communication is possible, it maybe be desirable to use a distribution of mechanisms. (This need not entail mixing for a given client as each client could be informed of the deterministic mechanism being used for her.) 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 adapted from (9) to

$$ES^* = \frac{D}{w^A - \sum_j z^j r^j}\left(v + \sum_j z^j T^j - \frac{\delta D}{2(w^A - \sum_j z^j r^j)}\right) - \frac{\delta \sigma_\epsilon^2}{2(w^A - \sum_j z^j r^j)^2}. \quad (33)$$

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.

Ex Ante and Ex Post Fairness: Note one feature of the Restricted Choice outcome. We have described this as relevant for settings where services are divisible. This is because the Fairness Constraint is *ex post*. An alternative is to characterize Fairness as ex ante. In this case, with indivisible goods the κ_i^* would be interpreted as client stating probabilities of receiving the service. Screening can then arise though single crossing in probabilities. This approach would mimic Hylland and Zeckhauser, 1979. We have ignored it here as 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.

Pricing: We have ignored one possibility for the firm to raise funds. The firm could potentially charge a fixed fee to clients before treatments are allocated: $R = \{\hat{r}, r_0, r_1, r_2\}$, where $\hat{r}$ is a charge before allocations are made. This allows the firm to raise more funds, which we describe in the Appendix. The results do not change in any conceptually substantive way. We have ignored it here because any additional implications rely on clients making inferences about the equilibrium, as the fee that the firm can charge depends on their beliefs of likely outcomes. Many clients of social services are not well informed about likely outcomes when they encounter the system, so we have ignored it here.

We have also assumed $r_i \geq 0$. This is for the usual reason that when clients are paid for receiving services, the supply elasticity of undeserving candidates only interested in the money may be large. Without that assumption of non-negativity, the cost to the firm is that by paying those who receive $t = 0$ (to generate efficient screening), there are less resources to increase the scope of the mission I. It does not change the essential tradeoff.

Donor Information and Preferences: We have assumed that donations are exogenously given by D. In reality, donors likely know the kind of clients who will be benefit, so D may be indexed by $\hat{y}$. All results continue to hold if $D(\hat{y}) + \underline{r}$ 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). It is also straightforward to provide a micro-foundation for donations. Assume that there are N potential donors, where donor l has marginal utility of income m_l^d. Assume $m^d \sim (0, \infty)$ according to a CDF H. Each donor is considering donating \$1 to the firm. He has preferences $U^d = \alpha^d ES^* - m_l^d \mathbb{I}_l^d$, where α^d represents altruism towards a client and $\mathbb{I}_l^d \in \{0, 1\}$ and equals 1 if donor l gives. Donors observe Π. All donor beliefs are that donations of others are $\hat{D}$, allocative efficiency is $\tilde{T}$, wages are $\tilde{w}$ and revenue is $\tilde{r}$. Donor l gives if $\frac{1}{\tilde{w}-\tilde{r}} \int_{\underline{\epsilon}}^{\overline{\epsilon}} (1 - \hat{\pi}'(x))(v + \tilde{T} - \frac{\delta(\hat{D}+x)}{2(\tilde{w}-\tilde{r})}) f(x)dx \geq \frac{1}{m_l^d}$. Total donations are $\mathbb{D}(b) \equiv NH((\frac{1}{\tilde{w}-\tilde{r}} \int_{\underline{\epsilon}}^{\overline{\epsilon}} (1 - \hat{\pi}'(x))(v + \tilde{T} - \frac{\delta(\hat{D}+x)}{2(\tilde{w}-\tilde{r})}) f(x)dx)^{-1})$. Then as Π is observed, donations for $\Pi = 1$ are 0 as $\pi^{*'}(\epsilon) = 1$, while $D \equiv NH((\frac{1}{\tilde{w}-\tilde{r}} \int_{\underline{\epsilon}}^{\overline{\epsilon}} (v + \tilde{T} - \frac{\delta(\hat{D}+x)}{2(\tilde{w}-\tilde{r})}) f(x)dx)^{-1})$ when $\Pi = 0$ as $\pi = 0$.

5 An Agency Interpretation

The value of allowing the firm to be residual claimant above is to offer financial security to the mission. The literature on non-profit status has instead focused on agency problems (Hansmann, 1980, Glaeser and Schleifer, 2001). We now interpret the choice of profit status through an agency lens. We do so for two reasons. First, to show a novel agency interpretation of the for-profit choice. Second, to show that outcomes need not be so stark as to eliminate all donations when the firm operates as a Social Enterprise.

The literature on the for-profit choice has studied a setting where a firm controls unobserved quality, and cuts corners on unobservable dimensions of quality in order to increase its profits at the expense of clients. The firm then potentially commits to non-profit status to eliminate that incentive. There are many examples of this behavior (Gertner, 2022). However, evidence for the most important non-profit sectors - schooling and medical care - find no evidence of lower quality among for-profits (Joynt et al, 2014, McClellan and Staiger, 1999, Dynarski et al., 2018, Cohodes and Parham, 2021, Taylor et al., 2023). As a result, we take a more positive approach to incentive contracting.[59]

We also assume that (i) the firm invests in unobservable quality, and (ii) when clients pay for treatment they do not see quality but instead make an equilibrium inference. Unlike the existing literature, the firm's objective is to maximize client welfare. Because quality is costly, there is a tradeoff between quality and clients

[59]This is not just to reflect the empirical evidence above, but because there remain many areas - fund raising, cost cutting by firing workers - where it remains difficult to motivate the management of non-profits. For example, Angrist et al, 2013, and Campos and Kearns, 2024, identify that the strongest predictors of charter school quality are increased student suspension rates and attendance rates, which school administrators are often reluctant to implement.

covered. This changes the firm's incentives, in the opposite direction of the existing literature. Here the non-profit firm covers more clients by reducing quality below the efficient level. As clients infer quality in equilibrium, this backfires. The value of for-profit status is that by suitable contracting, this incentive can be corrected. However, the optimal contract renders the firm residual claimant on the firm's revenues. As above, donors realize that the firm gets at least some of the donor's marginal contribution, and donations fall or are eliminated.

The firm now chooses non-contractible quality e at cost $C(e)$ per unit of I, where $C(0) = 0, C' > 0, C'' > 0$. We assume that $\delta = 0$ to isolate agency concerns. Clients value quality at $U_i^c = eT(\tau_i, t)I_i - V(y_i - r_j I_i)$. Quality is not observed by clients when they purchase, but is inferred in equilibrium and for perceived quality $\hat{e}$ will pay expected revenue of $r^A\hat{e}$. Firm utility is $\pi + S$ where $S = \sum_{i=1}^{N}(v_i + eT(\tau_i, t_i))I_i$. As above, we assume that $E\pi = 0$. Uncertainty is no longer given by ϵ but by $I(\epsilon; e)$ where $I(\epsilon; e) \sim F(\epsilon), f(\epsilon)$ over a support $(0, \overline{I})$, with mean

$$\tilde{I} = \frac{eD}{w^A + C(e^*) - r^A e}. \tag{34}$$

We assume that f_e is increasing in ϵ, and $\frac{f_e(x)}{f(x)}$ is increasing in x. The timing of the model is adapted to allow the choice of quality in *italics*:

1. The firm chooses A R, w, π and Π, which is observed by clients and the agent.
2. *The firm chooses* e and observes τ^a, the agent observes τ^a, and clients report $\hat{t}$.
3. $\tilde{I}$ *is revealed.*
4. I_i is implemented, r_i paid by clients, and w and π are paid.

To offer the best case for incentive contracting, we restrict attention to the surplus maximizing Bayes Nash equilibrium.

The analog to the donor-funded organization is $r^A\hat{e} - C(e) < w^A$. For a given allocation mechanism A, consider the first best choice of e, which would arise if e was contractible. The cost of each client served is now $w^A - r^A e + C(e)$. Expected surplus is now given by $ES = \frac{eD}{w^A + C(e) - r^A e}(v + T^A)$.[60] Note the comparison to (11). As $\delta = 0$ there is no longer a diminishing returns cost to uncertainty. However, quality scales both the value of surplus and the potential revenue from clients. This yields a first order condition for optimal quality

$$\frac{e^*(C'(e^*) - r^A)}{w^A + C(e^*) - r^A e^*} = 1. \tag{35}$$

Now consider the non-profit firm's incentives where $\pi = 0$, taking account of the fact that e is not observable to clients. Then $\tilde{e}$ is set to maximize $ES = \frac{\tilde{e}D}{w^A + C(\tilde{e}) - r^A\hat{e}}\left(v + T^A\right)$,

[60] We have not indexed e by the allocation mechanism A for notational simplicity.

where $\hat{e}$ is expected quality. Here the first order condition is $\frac{e^*C'(\tilde{e})}{w^A+C(\tilde{e})-r^A\hat{e}} = 1$, or as beliefs are correct in equilibrium,

$$\frac{\tilde{e}C'(\tilde{e})}{w^A + C(\tilde{e}) - r^A\tilde{e}} = 1. \tag{36}$$

Comparing (36) to (35) shows that the non-profit chooses quality too low, as it does not internalize the impact of quality on revenue.

Note one feature. This is not an agency setting in the usual way, where the actor receives a private benefit at the cost of some efficiency metric. Here the firm is only interested in the welfare of the client, and because of that preferences chooses quality that is too low, as it is not observed by the client. Contracting is used below to improve on this outcome.

Now consider the outcome for a for-profit firm in this setting. The contract offered is $\pi(\tilde{I}) \leq \mathbb{D}(b)$. Implemented treatments are now given by $I \leq \frac{e(\mathbb{D}(b)-\pi(\tilde{I}))}{w^a+C(e)-r}$, where $E\pi = 0$. The firm chooses e to maximize

$$\int_0^{\overline{I}} \pi(\tilde{I}(e))f(x)dx + e(v + T^A)\frac{\mathbb{D}(b)}{w^A + C(e) - r^A\hat{e}}, \tag{37}$$

The firm's choice of e is given by the first order condition

$$\int_0^{\overline{I}} \pi(x)f_e(x)dx + \frac{\mathbb{D}(b)(v + T^A)(w^A + C(e) - r^A\tilde{e} - eC'(e))}{(w^A + C(e) - r^A\hat{e})^2} = 0. \tag{38}$$

The firm wishes to implement e^* in (35). We consider Residual Claimant contracts, where the firm keeps all revenue above a certain level of $\tilde{I}$, which we call I^*.[61] Given risk neutrality and MLRP, the optimal contract (not observed by donors) has a simple form:

$$\pi(\tilde{I}(x)) = max\{\tilde{I}(x)) - I^*, 0\} - \overline{\pi} \tag{39}$$

where $\overline{\pi} = \int_{I^*}^{\overline{I}} \pi(\tilde{I}(x))f(x)dx$ to ensure that $E\pi = 0$. Then I^* is set such that the analog to (38) holds:

$$\int_{I^*}^{\overline{I}} (\tilde{I}(x) - I^*)f_e(x)dx =$$
$$\mathbb{D}(b)(v + T^A)\left(\frac{w^A + C(e^*) - r^Ae^* - e^*C'(e^*)}{(w^A + C(e^*) - r^Ae^*)^2} - \frac{w^A + C(\tilde{e}) - r^A\tilde{e} - \tilde{e}C'(\tilde{e})}{(w^A + C(\tilde{e}) - r^A\tilde{e})^2}\right). \tag{40}$$

[61] The relevant issue is what set of contracts are feasible within that class. One possibility would be to allow the firm to pay the firm D if $I \geq I^*$. These are likely to be infeasible. The theoretically most desirable contract is to offer the firm all revenue from the firm if I exceeds some level. We describe this contract in the Appendix. These contracts are empirically infeasible for two reasons. First, no clients will be served in these instances. Second, realistically the way in which this information becomes available is that at the end of the year, excess revenue is seen, and the firm can claim some of that. It cannot claim all revenue for the entire year.

If

$$\int_0^{\overline{I}} \tilde{I}(x) f_e(x) dx <$$
$$\mathbb{D}(b)(v + T^A) \left(\frac{w^A + C(e^*) - r^A e^* - e^* C'(e^*)}{(w^A + C(e^*) - r^A e^*)^2} - \frac{w^A + C(\tilde{e}) - r^A \tilde{e} - \tilde{e} C'(\tilde{e})}{(w^A + C(\tilde{e}) - r^A \tilde{e})^2} \right), \quad (41)$$

then full residual claimant status is not enough to induce efficient effort and efficiency is not feasible. I^* is set such that the expected marginal income balances the need for additional quality (the term in large brackets), applied to the surplus for all clients. Holding donations fixed, this optimally corrects the quality provision problem, improving allocative efficiency.

While this increases allocative efficiency, it deters donors. The firm chooses π^* holding d constant, as donors do not observe π.[62] However, in equilibrium the donor's contribution funds mission-related activities only with probability $F(I^*)$. Total donations are $\mathbb{D}(b) = \mathbb{D}(F(I^*))$. When $I^* = 0$, all donations are crowded out.[63]

This section has been provided both to illustrate the possibility of a for-profit firm continuing to attract donors ($I^* > 0$), but also to offer an agency based interpretation. Since Hansmann, 1980, the literature has argued that agency issues cause the firm to be non-profit, as harmful incentives causes the firm to cut corners and harm clients. This is reversed here: only if a firm can self-fund through clients can it be for-profit. From this perspective, the limit to for-profit social entrepreneurship is not quality concerns, but rather the simpler problem of whether the firm can pay the bills after scaring donors away.

6 Conclusion

Social organizations do not use market clearing prices to allocate their services. The academic literature has championed Deferred Acceptance as a preferred alternative to choice with prices. Its appeal is couched in terms of strategy proofness and stability rather than utilitarian gains (Li, 2017). The central premise here is that many social organizations are likely more concerned with raising funds to expand their mission, retaining and attracting staff, and exploring the possibility of client choice to avoid

[62] This analysis assumes that the optimal assignment mechanism A does not depend on the for-profit status of the firm. It can do so, of course, but this does not change the logic of the choice being between efficient quality and low donations, or less efficient quality with more donors.

[63] The firm then chooses between non-profit status with return S^n, or for-profit with return $ES^* = \frac{e^* F(I^*) D(\underline{m})}{\overline{U}^a + C(e^*) - r^A e^*} \left(v + T^A\right)$.When $I^* = 0$, the firm can only operate as a for-profit if it can self-fund: $w^a + C(e^*) \leq r^A e^*$. In that case, optimal quality is no longer given by (35), as the cost there is fewer clients served. With the self-funded firm, quality is instead constrained by the budget constraint, $w^a + C(e^*) \leq r^A e^*$. As the objective of the firm is to maximize surplus, the firm will choose the maximum possible e where $w^a + C(e^*) = r^A e^*$, and $C'(e^*) = r^A$.

assignment error. It interprets why a firm would allow a bureaucrat with the "wrong" preferences unfettered control when something as apparently appealing as Deferred Acceptance is available. We also showed how the cost for an organization serving very low income clients is not just that their funding is low, but also that allocations entail error, because choice can only offered to clients with more income.

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. It is for this reason that we have proposed the Fairness Constraint as the relevant obstacle for social organizations. 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 *ex post* Pareto efficiency. Finally, that donors who care about the impact of their gift are (rightly) suspicious if the firm can take their donation as profit.

The paper is not intended to explain the actions of all social sector firms. Most importantly, we do not address cases in the public sector such as school choice, where funding occurs through tax revenue. In addition, many social organizations do not used the price system because there are externalities to client choice. For example, schools care about the composition of their student body. Finally, we also do not consider instances where the objectives of the firm are a hybrid of both surplus and private returns. The main reason for this is to understand the complex tradeoffs that social sector firms face even when their objectives are "pure".

We began the paper by describing the importance of prices for allocative efficiency. 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 social organizations in a 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.

Al Jazeera, "Poorer Nations Forced to Dump Close-to-expiry Covid Vaccines", 2022. https://www. aljazeera.com/news/2022/1/13/poorer-nations-dump-close-to-expiry-covid-vaccines-unicef.

Alatas Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken,

Ririn Purnamasari and Matthew Wai-Poi "Ordeal Mechanisms In Targeting: Theory And Evidence From A Field Experiment In Indonesia", 2013, NBER Working Paper 19127.

American Association of Fundraising Professionals, "Professional Compensation: A Position Paper", 2016. https://afpglobal.org/sites/default/files/attachments/2018-10/2016ProfCompensationPositionPaper.pdf.

American Hospital Association, "Fast Facts on U.S. Hospitals, 2024", 2024. https://www.aha.org/statistics/fast-facts-us-hospitals.

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.

Angrist, Josh, , P. A. Pathak, and Chris Walters, "Explaining Charter School Effectiveness", *American Economic Journal: Applied Economics*, vol. 5, no. 4, October 2013.

Angrist, Josh, , Guthrie Gray-Lobe, Clemence M. Idoux and Parag A. Pathak, "Still Worth the Trip? School Busing Effects in Boston and New York", 2024, NBER Working Paper, 30308.

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.

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.

Bruhn, Jesse, Christopher Campos, Eric Chyn, and Anh Tran, "Who Benefits from Remote Schooling? Self-Selection and Match Effects", mimeo, July, 2024.

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.

Campos, Christopher and Caitlin Kearns, "The Impact of Public School Choice: Evidence from Los Angeles Zones of Choice", *Quarterly Journal of Economics*, 2024, 139, 1–43.

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.

CCS Fundraising, "2024 Philanthropic Landscape, 13th Edition", 2024.

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

Cheng, M., and Hinnant, L., "Rich Nations Dip into COVAX Supply While Poor Wait for Shots", Associated Press, 2021, https://apnews.com/article/joe-biden-middle-east-africa-europe-corona- virus-pandemic-5e57879c6cb22d96b942cbc973b9296c.

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.

Cohodes, Sarah R., and Katharine S. Parham, "Charter Schools' Effectiveness, Mechanisms, and Competitive Influence", 2018, NBER Working Paper No. 28477.

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.

Davis, Jonathan M.V. , Kyle Greenberg and Damon Jones, "An Experimental Evaluation of Deferred Acceptance: Evidence from Over 100 Army Officer Labor Markets", 2023, NBER Working Paper, 31612.

Dewatripont, Matthias, Ian Jewitt, and Jean Tirole, "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies", *Review of Economic Studies*, 1999, vol. 66, issue 1, pages 199–217.

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.

Dynarski, Susan, Daniel Hubbard, Brian Jacob and Silvia Robles, "Estimating the Effects of a Large For-Profit Charter School Operator", 2018, NBER Working Paper No. 24428.

Ergin, H., "Efficient Resource Allocation on the Basis of Priorities", *Econometrica*, 2002, 70, 2489–2497.

Farahan, Fatih, "Understanding Earnings Dispersion", 2015, Federal Reserve Bank of New York. https://libertystreeteconomics.newyorkfed.org/2015/11/understanding-earnings-dispersion/.

Farhoud, M., S. Shah, P. Stenholm, E. Kibler, M. Renko, and S. Terjesen, "Social enterprise crowdfunding in an acute crisis", *Journal of Business Venturing Insights*, 2021, e00211.

Faulk, Lewis, Mirae Kim, Teresa Derrick-Mills, Elizabeth T. Boris,

Laura Tomasko, Nora Hakizimana, Tianyu Chen, Minjung Kim, Layla Nath "Nonprofit Trends and Impacts 2021", 2021, *Urban Institute*.

Foley, Duncan, "Resource Allocation and the Public Sector", 1967, *Yale Economic Essays*, 7, 45-98.

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

Bai 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.

Gertner, Robert, "The Organization of Social Enterprises", *Annual Review of Economics*, 2023, 15:41–62.

Glaeser, E., and A. Shleifer, "Not-for-profit entrepreneurs", *Journal of Public Economics*, 2011, 81 (1), 99-116.

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

Goldhill, O., '"Naively Ambitious": How COVAX Failed on its Promise to Vaccinate the World", STAT News, 2021, https://www.statnews.com/2021/10/08/how-covax-failed-on-its-promise-to-vaccinate-the- world/.

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, 6(6), 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.

Joynt, Karen E, E John Orav, and Ashish K Jha, "Association between hospital conversions to for-profit status and clinical and economic outcomes", *Journal of the American Medical Association*, 2014, 16, 1644-52.

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.

Leshno, Jacob, "Dynamic Matching in Overloaded Waiting Lists",*American Economic Review*, 2022, vol. 112, no. 12, 3876–3910.

Li, Shengwu, "Ethics and Market Design", mimeo, 2017.

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.

McClellan, Mark and Doug Staiger, "The Quality of Health Care Providers", 1999, NBER Working Paper 7327.

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.

Neal, Derek and Joseph Root, "The Provision of Information and Incentives in School Assignment Mechanisms", 2024, BFI Working Paper 2024-55.

National Council of Nonprofits, "Downloadable Charts and Figures", 2024.

https://www.nonprofitimpactmatters.org/data/downloadable-charts/.

Network for Public Education, "Chartered for Profit", 2023. https://networkforpubliceducation.org/chartered-for-profit/.

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

Mark Ottoni-Wilhelm, Lise Vesterlund, and Huan Xie, "Why Do People Give? Testing Pure and Impure Altruism", *American Economic Review*, 2017, 107(11), 3617-3633.

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.

Posner, Eric, and Anup Malani, "The Case for For-Profit Charities", 2006, John M. Olin Program in Law and Economics Working Paper No. 304.

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, "Bureaucratic Responses", *Journal of Labor Economics*, 2016, 34 (S2), S183-215.

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.

Prendergast, Canice, "Organizational Design for Making a Difference", *Journal of Public Economics*, 2023, vol. 218, 104817.

Radford, R.A., "The Economic Organization of a P.O.W. Camp", *Economica*, 1945, 12, 189-201.

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.

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

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

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

Sandberg, Haim, "Where Do Your Pennies Go? Disclosing Commissions for Charitable Fundraising", *The International Journal of Not-for-Profit Law*, 2006, 8(4).

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/.

Taylor, Lauren A., Berkeley Franz, Amanda Zink, Andrew Fair, and

Cory E. Cronin, "Public perceptions of US for-profit, nonprofit, and public hospitals", *Health Affairs Scholar,* 2023, 1(4), 1-9.

Teamstage, "Volunteering Statistics for 2024", 2024, https://teamstage.io/volunteering-statistics/.

Terrier, Camille, Parag Pathak, and Kevin Ren, "From Immediate Acceptance to Deferred Acceptance: Effects on School Admissions and Achievement", 2021, NBER Working Paper 29600.

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.

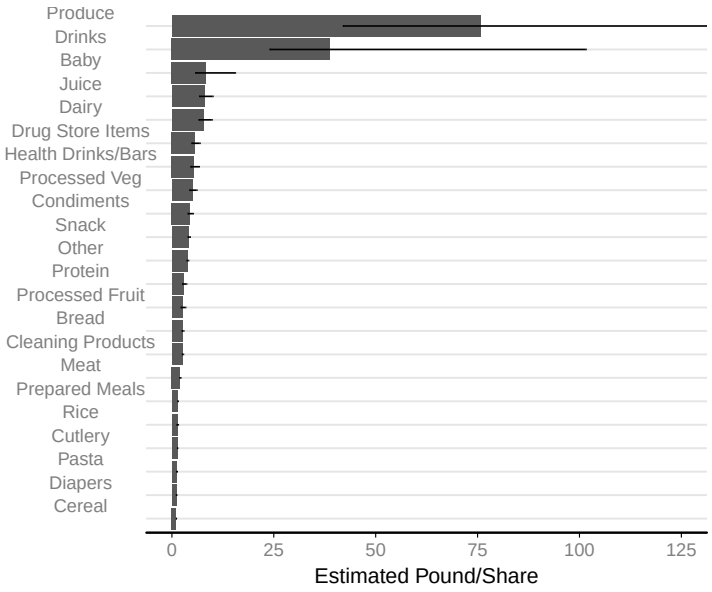


Figure 2: **Feeding America Price of Different Goods.**

7 Appendix

The Program: The firm begins the game by choosing A. The firm's objective given A is to choose $I_i^{*A}, \pi^{*A}, \Pi^{*A}$ and w^{*A} maximize

$$ES = \int_{\underline{\epsilon}}^{\overline{\epsilon}} \sum_{i=1}^{\infty} (v_i + T^A(t_i, \tau_i)) I_i(x) f(x) dx, \tag{42}$$

where $v_i = v - \delta i$ and $I_i \in [0,1]$. If $w^A < r^A$, then $I^A = \sum_{i=1}^{\infty} I_i^A$ is defined by

$$I^A \leq \frac{\mathbb{D}(b)^A - \pi^A + \epsilon}{w^A - r^A}, \tag{43}$$

where $\tilde{I}^A(\epsilon) = \frac{\mathbb{D}^A + \epsilon}{w^A - r^A}$. This is subject to the following constraints:

1. $E\pi^A = 0$. $\pi^A(\tilde{I}^A) = 0$ if $\Pi = 0$ and $\pi^A \leq \mathbb{D} + \epsilon$ if $\Pi = 1$.
2. Wages are given by (6).
3. Prices r_j^A are given by $T(i, j|\underline{\phi}) \geq \underline{m} r_j^A$, in any equilibrium where type τ_i receives treatment $t = j$ with positive probability. Revenue per client is $r^A = \frac{r_0^A + r_1^A + r_2^A}{3}$.
4. Client Reporting in A:
   - If A involves Assigned Consumption, let $P_{ij}^{1st}(\hat{t}_l | \hat{t}_{-l})$ 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
   $$\sum_j P_{ij}^{1st}(\hat{t}_l | \hat{t}_{-l})(T(i,j|\phi) - m r_j^{1st}) \geq \sum_j P_{ij}^{1st}(\hat{t'}_l | \hat{t}_{-l})(T(i,j|\phi) - m r_j^{1st}), \forall m \in [\underline{m}, \overline{m}], \tag{44}$$
   where P_{ij}^{1st} are generated by Bayes rule.
   - If A involves Assigned Consumption and the allocation mechanism is strategy proof, this simplifies to
   $$T(i, 2|\underline{\phi}) - m r_2^A \geq T(i, j|\underline{\phi}) - m r_j^A, \forall m \in [\underline{m}, \overline{m}], \tag{45}$$
   $$T(i, 1|\underline{\phi}) - m r_1^A \geq T(i, 0|\underline{\phi}) - m r_0^A, \forall m \in [\underline{m}, \overline{m}], \tag{46}$$
   and
   $$T(i, i|\overline{\phi}) - m r_i^A \geq T(i, j|\overline{\phi}) - m r_j^A, \forall m \in [\underline{m}, \overline{m}]. \tag{47}$$
   - If A involves Choice where the client choose $I_i^A = k_i^A \leq 1$ at unit price r_i^A
   $$k_i^A (T(i, i|\underline{\phi}) - m r_i^A) \geq k_j^A (T(i, j|\underline{\phi}) - m r_j^A), \forall m \in [\underline{m}, \overline{m}]. \tag{48}$$
5. In Section 3, donations are $d(b)$ where $b = \int_{\underline{\epsilon}}^{\overline{\epsilon}} \pi^{*'A}(\tilde{I}(x)) f(x) dx$ in equilibrium.

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

Proof of Lemma 1: Unrestricted Choice requires that for a client i of type $\tau_i = a$

$$a^2 + \phi - mr_a \geq ab - mr_b, \tag{49}$$

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 $\hat{t}^1 > \tau$. There are three cases:

- $\hat{t}(0, \underline{\phi}, m) = 0 \succ \hat{t}(0, \underline{\phi}, m) = 1$ if $\underline{\phi} - mr_0 \geq \mu - mr_1, \forall m$, or $r_1 - r_0 \geq \frac{\mu - \underline{\phi}}{\overline{m}}$, for $\frac{\underline{\phi}}{\underline{m}} \geq r_0 \geq 0$ and $r_1 \leq \frac{\mu + \underline{\phi}}{\underline{m}}$.

- $\hat{t}(0, \underline{\phi}, m) = 1 \succ \hat{t}(0, \underline{\phi}, m) = 2$ if $\underline{\phi} + \mu - mr_1 \geq 2\mu - mr_2, \forall m$, or $r_2 - r_1 \geq \frac{\mu - \underline{\phi}}{\overline{m}}$, for $r_1 \leq \frac{\mu + \underline{\phi}}{\underline{m}}$ and $r_2 \leq \frac{2\mu + \underline{\phi}}{\underline{m}}$.

- $\hat{t}(0, \underline{\phi}, m) = 0 \succ \hat{t}(0, \underline{\phi}, m) = 2$ if $\underline{\phi} - mr_0 \geq 2\mu - mr_2$ for $r_0 \geq 0$ and $r_2 \leq \frac{2\mu + \underline{\phi}}{\underline{m}}$. If the Fairness Constraint can be overcome, where all three conditions hold, then $r_2^* = \frac{2\mu + \underline{\phi}}{\underline{m}}$, and $r_1^* = \frac{\mu + \underline{\phi}}{\underline{m}}$ as this maximizes revenue and the first two constraints. The final constraint is the most binding and can be satisfied if

  $$\overline{m} r_0^* \leq 2\mu - \underline{\phi} + \overline{m} \frac{2\mu + \underline{\phi}}{\underline{m}}, \tag{50}$$

  for $r_0 \geq 0$. As $r_0 \geq 0$, this implies that the Fairness Constraint can only be satisfied if

  $$\frac{\overline{m}}{\underline{m}} \geq \frac{2\mu - \underline{\phi}}{2\mu + \underline{\phi}}. \tag{51}$$

It also implies that if (51) holds,

$$r_0^* = \min\{\frac{\underline{\phi} - 2\mu}{\overline{m}} + \frac{2\mu + \underline{\phi}}{\underline{m}}, \frac{\underline{\phi}}{\underline{m}}\}. \tag{52}$$

Straightforward calculations show that $r_0^* = \frac{\underline{\phi}}{\underline{m}}$ if $\frac{\overline{m}}{\underline{m}} \geq 1 - \frac{\underline{\phi}}{2\mu}$. 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.

For the second part, there are two cases that need to be considered. First, if $\frac{\overline{m}}{\underline{m}} < 1 - \frac{\underline{\phi}}{\mu}$, no types can screen. 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 $\hat{t}(\tau, \underline{\phi}, m) = (2 \succ 1 \succ 0)$, $\hat{t}(1, \overline{\phi}, m) = (1 \succ 2 \succ 0)$, and $\hat{t}(0, \overline{\phi}, m) = (0 \succ 2 \succ 1)$ for all m. From (45) and (46), $\hat{t}(\tau, \underline{\phi}, m) = (2 \succ 1 \succ 0)$ requires

$$\frac{\mu}{\underline{m}} - r_1 \geq \frac{\underline{\phi}}{\underline{m}} - r_0, \tag{53}$$

and

$$\frac{2\mu}{\underline{m}} - r_2 \geq \frac{\underline{\phi} + \mu}{\underline{m}} - r_1. \tag{54}$$

These hold with equality and $r_1 - r_0 = \frac{\mu - \underline{\phi}}{\underline{m}}$ and $r_2 - r_1 = \frac{\mu - \underline{\phi}}{\underline{m}}$. We show below that for any assigned consumption mechanism, $prob(t = 0|\tau) > 0$ for all τ as $q < 1$. Because $T(\tau, 0) = 0$ for $\tau \in 1, 2 \implies r_0 \leq 0$. Letting this hold with equality implies $R^* = \{0, \frac{\mu - \underline{\phi}}{\underline{m}}, \frac{2(\mu - \underline{\phi}))}{\underline{m}}\}$ and $\underline{r} = \frac{0 + \frac{\mu - \underline{\phi}}{\underline{m}} + \frac{2(\mu - \underline{\phi}))}{\underline{m}}}{3} = \frac{\mu - \underline{\phi}}{\underline{m}}$.

Now let $1 - \frac{\underline{\phi}}{\mu} < \frac{\overline{m}}{\underline{m}} < 1 - \frac{\underline{\phi}}{2\mu}$. This offers the possibility of partial choice. As $\hat{t}(0, \underline{\phi}, m) = 2$ as shown above, the client cannot be allowed to choose $\hat{t} = 2$. However, it raises the possibility that the firm could assign the client to either $t = 2$ or $t \in \{0, 1\}$ where the client chooses between $t = 0$ and $t = 1$. We show that this is impossible when $\frac{\overline{m}}{\underline{m}} < 1 - \frac{\underline{\phi}}{2\mu}$. $\hat{t}(0, \underline{\phi}, m) = 0 \forall m$ if $m(r_1 - r_0) \geq \mu - \underline{\phi}$. $\hat{t}(1, \underline{\phi}, m) = 1 \forall m$ if $m(r_1 - r_0) \leq \mu + \underline{\phi}$. But as $q = 1$, clients of type τ are assigned to a choice between $t = 0$ and $t = 1$. As a result, $r_0 = 0 \implies mr_1 \geq \mu - \underline{\phi}$ and $mr_1 \leq \mu + \underline{\phi}$. The first condition is most binding for $m = \overline{m}$ and the second for $m = \overline{m}$. Then it must be the case that $\frac{\overline{m}}{\underline{m}} \geq \frac{\mu + \underline{\phi}}{\mu - \underline{\phi}} = \frac{1}{1 - \frac{2\mu}{\mu + \underline{\phi}}}$. However as we are considering the case where $1 - \frac{\underline{\phi}}{\mu} < \frac{\overline{m}}{\underline{m}} < 1 - \frac{\underline{\phi}}{2\mu}$, we have a contradiction, which implies that for $\frac{\overline{m}}{\underline{m}} < 1 - \frac{\underline{\phi}}{2\mu}$ there is Assigned Choice and $r = \underline{r}$. □

Proposition 9 *Expected surplus from Deferred Acceptance is*

$$ES^{da} = \frac{D}{\overline{U} - \underline{r}} \left(v + T^{da} - \frac{\delta D}{2(\overline{U} - \underline{r})} \right) - \frac{\delta \sigma_\epsilon^2}{2(\overline{U} - \underline{r})^2}. \tag{55}$$

where (i) if $h \leq \frac{2}{3-q}$,

$$T^{da} = \mu + \frac{1}{3} \left(\hat{\phi}(q + p_2 \frac{1-q}{2}) + \overline{\phi} h (q + \frac{1-q}{2}) + (1-h)(\underline{\phi} q (1 - p_2) + \hat{\phi} \frac{1-q}{2} p_1) \right) + \\ + \frac{1}{3} \left(h\overline{\phi} + q(1 - p_1)(1 - h)\underline{\phi} \right), \tag{56}$$

where $p_2 = \frac{h(1-q)}{q(1-h) + (2-h)\frac{1-q}{2}}$ *and* $p_1 = \frac{\frac{h(1-q)}{2}}{(1-h)q + (1-q)}$, *and*
(ii) if $h > \frac{2}{3-q}$,

$$T^{da} = \mu + \frac{1}{3} \left(\hat{\phi}(q + (1 + \tilde{p}_2 \frac{1-q}{2}) + \overline{\phi} h (q + \frac{1-q}{2}) + (1-h)\underline{\phi} \frac{1-q}{2} \tilde{p}_1 \right) + \\ + \frac{1}{3} \left(h\overline{\phi} + q(1 - \tilde{p}_1)(1 - h)\underline{\phi} \right), \tag{57}$$

where $\tilde{p}_2 = \frac{h(1-q) - (q(1-h) + (2-h)\frac{1-q}{2})}{q(1-h) + (2-h)\frac{1-q}{2}}$ *and* $\tilde{p}_1 = \frac{h(1-q)}{(1-\tilde{p}_2)(q(1-h) + (2-h)\frac{1-q}{2})}$

Proof of Proposition 9 (and 1): From Lemma 1, $r = \underline{r}$. Let $\hat{\tau}^a$ be the agent's posterior. Then as $t \perp (\hat{\tau}a|\tau^a) \implies w = \overline{U}$ so $\hat{U}^a = \overline{U}^a \implies w = \overline{U}$ and (6) holds. Then $I = \frac{D+\epsilon}{\overline{U}-\underline{r}}$.

The mechanism begins with clients $i = 1, 2$, ...reporting $\hat{t}$. The firm offers $t \forall i = 1, 2, ...\frac{D+\epsilon}{\overline{U}-\underline{r}}$. At prices $R^* = \{0, \frac{\mu-\underline{\phi}}{\underline{m}}, \frac{2(\mu-\underline{\phi}))}{\underline{m}}\} \implies \hat{t}(\tau, \underline{\phi}, m) = (2 \succ 1 \succ 0)$, $\hat{t}(2, \overline{\phi}, m) = (2 \succ 1 \succ 0)$, $\hat{t}(1, \overline{\phi}, m) = (1 \succ 2 \succ 0)$, and $\hat{t}(0, \overline{\phi}, m) = (0 \succ 2 \succ 1)$ for all m. As a result, (45) and (46) hold. Note that

$$\hat{t} = \begin{cases} 2 \succ 1 \succ 0 & \text{with probability } 1 - \frac{2h}{3} \\ 1 \succ 2 \succ 0 & \text{with probability } \frac{h}{3} \\ 0 \succ 2 \succ 1 & \text{with probability } \frac{h}{3}. \end{cases} \tag{58}$$

1. *The firm allocates $t = 2$ to all clients for which $\tau^a = 2$ and $\hat{t} = (2 \succ 1 \succ 0)$.* Consider all who report $\hat{t} = (2 \succ 1 \succ 0)$. Of these, $\frac{1}{3}$ are $\tau = 2$ and $1 - \frac{h}{3}$ are each $\tau = 1$ and $\tau = 0$. $Prob(\tau^a = 2|\hat{t} = (2 \succ 1 \succ 0)) = 1 - h(1-q)$. 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 $\implies O(2|2, (2 \succ 1 \succ 0)) = 1$. All offers are accepted.

2. *The firm randomly allocates $t = 2$ to clients for which $\tau^a = 1$ and $\hat{t} = (2 \succ 1 \succ 0)$ until there are no more $t = 2$ slots.* There are then $h(1-q)$ remaining $t = 2$ slots available per capita, and $(q(1-h) + (2-h)\frac{1-q}{2})$ clients for whom $\hat{t} = (2 \succ 1 \succ 0)$ and $\tau^a = 1$. Note that $h(1-q) \leq (q(1-h) + (2-h)\frac{1-q}{2})$ if $h \leq \frac{2}{3-q}$. If this is the case then Step 2 $\implies O(2|2, (2 \succ 1 \succ 0)) = p_2 = \frac{h(1-q)}{q(1-h)+(2-h)\frac{1-q}{2}}$. These offers are all accepted.

3. *The firm allocates $t = 1$ to all remaining clients for which $\tau^a = 1$ and $\hat{t} = (2 \succ 1 \succ 0)$ and to all clients with $\tau^a = 1$ and $\hat{t} \neq (0 \succ 2 \succ 1)$.* There are $(1 - p_2)(q(1-h) + (2-h)\frac{1-q}{2})$ remaining clients per capita with for $\tau^a = 1$ and $\hat{t} = (2 \succ 1 \succ 0)$. There are $hq + \frac{h(1-q)}{2}$ for whom $\tau^a = 1$ or $\tau^a = 2$ and $\hat{t} = (1 \succ 2 \succ 0)$. Step 3 $\implies O(1|2, (2 \succ 1 \succ 0)) = O(1|1, (1 \succ 2 \succ 0)) = 1$. All are accepted.

4. *The firm randomly allocates $t = 1$ to clients for which $\tau^a = 0$ and $\hat{t} \neq (0 \succ 2 \succ 1)$ until there are no more $t = 1$ slots.* There are then $\frac{h(1-q)}{2}$ unfilled $t = 1$ slots, and $(1-h)q + (1-q)$ clients where $t^f = 0$ and $\hat{t} \neq (0 \succ 2 \succ 1)$. Step 4 $\implies O(1|0, (2 \succ 1 \succ 0)) = O(1|0, (1 \succ 2 \succ 0)) = p_1 = \frac{\frac{h(1-q)}{2}}{(1-h)q+(1-q)}$. All offers are accepted.

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

Expected surplus per client from treatment when $h \geq \frac{2}{3-q}$ is

$$T^{da} = \mu + \frac{1}{3}\left(\hat{\phi}(q + p_2\frac{1-q}{2}) + \overline{\phi}h(q + \frac{1-q}{2}) + (1-h)\underline{\phi}q(1-p_2) + \hat{\phi}\frac{1-q}{2}p_1\right) + \\ + \frac{1}{3}\left(h\overline{\phi} + q(1-p_1)(1-h)\underline{\phi}\right) \quad (59)$$

so expected surplus is

$$ES^{da} = \frac{D+\epsilon}{\overline{U}-\underline{r}}\left(v + T^{da} - \frac{\delta(D+\epsilon)}{2(\overline{U}-\underline{r})}\right), \quad (60)$$

or

$$ES^{da} = \frac{D}{\overline{U}-\underline{r}}\left(v + T^{da} - \frac{\delta D}{2(\overline{U}-\underline{r})}\right) - \frac{\delta\sigma_\epsilon^2}{2(\overline{U}-r)^2}. \quad (61)$$

This addressed the case where $h \leq \frac{2}{3-q}$. Now assume $h > \frac{2}{3-q}$. Then Step 2 $\implies$ $O(2|2,(2 \succ 1 \succ 0)) = 1$. These offers are all accepted. There are then $h(1-q) - (q(1-h) + (2-h)\frac{1-q}{2})$ remaining $t=2$ slots, and $(1-p_2)(q(1-h) + (2-h)\frac{1-q}{2})$ remaining clients per capita with for $\tau^a = 0$ and $\hat{t} = (2 \succ 1 \succ 0)$. For $h < 1$, $h(1-q) \geq 2(q(1-h) + (2-h)\frac{1-q}{2})$, so Step 2 also $\implies$ $O(2|0,(2 \succ 1 \succ 0)) = \tilde{p}_2 = \frac{h(1-q) - (q(1-h)+(2-h)\frac{1-q}{2})}{q(1-h)+(2-h)\frac{1-q}{2}}$. There are then $(1-\tilde{p}_2)(q(1-h)+(2-h)\frac{1-q}{2})$ clients for which $\hat{t} = (2 \succ 1 \succ 0))$ and $\tau^a = 0$ but who were not offered $t=2$ in the previous step. Step 3 $\implies$ $O(2|1,(1 \succ 2 \succ 0)) = 1$, so there are $h(1-q)$ unfilled $t=1$ slots. Step 4 $\implies$ $O(1|0,(2 \succ 1 \succ 0)) = \tilde{p}_1 = \frac{h(1-q)}{(1-\tilde{p}_2)(q(1-h)+(2-h)\frac{1-q}{2})}$. All offers are accepted. All other clients receive $t=0$. Expected surplus per client from treatment when $h \geq \frac{2}{3-q}$ is given by

$$\tilde{T}^{da} = \mu + \frac{1}{3}\left(\hat{\phi}(q + (1+\tilde{p}_2)\frac{1-q}{2}) + \overline{\phi}h(q + \frac{1-q}{2}) + (1-h)\underline{\phi}\frac{1-q}{2}\tilde{p}_1\right) + \\ + \frac{1}{3}\left(h\overline{\phi} + q(1-\tilde{p}_1)(1-h)\underline{\phi}\right), \quad (62)$$

so

$$ES^{da} = \frac{D}{\overline{U}-\underline{r}}\left(v + \tilde{T}^{da} - \frac{\delta D(\underline{m})}{2(\overline{U}-\underline{r})}\right) - \frac{\delta\sigma_\epsilon^2}{2(\overline{U}-r)^2}.\square \quad (63)$$

Proof or Proposition 3: Consider an agent who receives a signal τ^a on a client. With probability λ, $t_i = \tau_i^a$. With probability $1-\lambda$, $t_i = \tilde{\tau}_i$. These are offered at prices $R^* = \{0, \frac{\mu-\underline{\phi}}{\underline{m}}, \frac{2(\mu-\underline{\phi}))}{\underline{m}}\}$. The expected value of the treatment from the perspective of the agent is then $\hat{U}^{ca} \approx \lambda q\hat{\phi} + (1-\lambda)\hat{\phi}$. The firm's belief of surplus is $\lambda q\hat{\phi} + (1-\lambda)\frac{\hat{\phi}}{3}$ as

$\tilde{\tau}$ is uniformly distributed. The agent's belief of the surplus from a random alternative agent is

$$\overline{U}_i^{ca} = \begin{cases} q\hat{\phi} & \text{with probability } \lambda^2 \\ \rho\hat{\phi} & \text{with probability } (1-\lambda)^2 \\ q\hat{\phi} & \text{with probability } \lambda(1-\lambda). \\ \frac{\hat{\phi}}{3} & \text{with probability } (1-\lambda)\lambda, \end{cases} \tag{64}$$

where $\hat{U}^{ca} - \overline{U}^{ca} = (1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}$. Then from (6), $w^* = \overline{U} - \alpha(1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}$ (we ignore the approximation). Then $I^a = \frac{D+\epsilon}{w^*-r^b}$. The surplus from treatment is then

$$T^a = \mu + \lambda q\hat{\phi} + (1-\lambda)\frac{\hat{\phi}}{3} < T^{da}. \tag{65}$$

and

$$ES^a = \frac{D}{w^*-\underline{r}}\left(v + T^a - \frac{\delta D}{2(w^*-\underline{r})}\right) - \frac{\delta\sigma_\epsilon^2}{2(w^*-\underline{r})^2}. \tag{66}$$

Then

$$E(I^a - I^{da}) = \Delta I^a = \frac{D\alpha(1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}}{(w^*-\underline{r})(\overline{U}-r)}, \tag{67}$$

and $ES^a - ES^{da} > 0$ if

$$I^{da} \leq \frac{\Delta I^a(v + T^a - \frac{\delta\Delta w^*}{2})}{T^{da} - T^a + \delta\Delta I^a}, \tag{68}$$

where $\Delta w^* = \alpha(1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}$.□

Pareto Efficient Agent Assignment: We now allow the client to report $\hat{t}$ at stage 1 of the mechanism and set prices at the level at which $\hat{t}(\tau, \underline{\phi}, m) = (2 \succ 1 \succ 0), \hat{t}(2, \overline{\phi}, m) = (2 \succ 1 \succ 0), \hat{t}(1, \overline{\phi}, m) = (1 \succ 2 \succ 0)$ and $\hat{t}(0, \overline{\phi}, m) = (0 \succ 2 \succ 1)$, namely by choosing R^{da}. Further consider the case where reports are strategy proof. Pareto efficiency implies that $\hat{t}(1, \overline{\phi}, m) = 1 \succ 2 \succ 0$ and $\hat{t}(0, \overline{\phi}, m) = 0 \succ 2 \succ 1$ with probability 1. As a result, $\hat{t} = 1 \succ 2 \succ 0 \implies t = 1$, and $\hat{t} = 0 \succ 2 \succ 1 \implies t = 0$, as the agent's posterior is 1 in each case. After allocating these clients, the agent has a fraction $1 - \frac{2h}{3}$ clients for which

$$\tau^a = \begin{cases} 2 & \text{with probability } \frac{1-\frac{2h}{3}}{3} \\ 1 & \text{with probability } \frac{1-\frac{2h}{3}}{3} \\ 0 & \text{with probability } \frac{1-\frac{2h}{3}}{3}. \end{cases} \tag{69}$$

It allocates these to $\frac{1}{3}$ slots for $t = 2$, $\frac{1-h}{3}$ slots for $t = 1$ and $\frac{1-h}{3}$ slots for $t = 0$. To maximize client welfare from the perspective of the agent, all clients with $\tau^a = i$ are

allocated $t=1$. There are $\frac{2h}{9}(1-h)$ remaining $t=2$ slots, and $\frac{h}{9}$ clients with $\tau^a=1$ and $\tau^a=0$. These are matched. The expected treatment surplus per client is then

$$\mu+\left(\frac{1-\frac{2(1-h)}{3}}{3}(\lambda q+\frac{1-\lambda}{3})+\frac{2(1-h)}{9}\frac{1-q}{2}\right)\hat{\phi}+\frac{2h}{3}\underline{\phi}+2(\frac{1-\frac{2(1-h)}{3}}{3})q\underline{\phi}. \quad (70)$$

The increase in expected treatment surplus over Agent Assignment is $\left(\frac{2(1-h)}{9}\frac{1-q}{2}\right)\hat{\phi}+\frac{2h}{3}\overline{\phi}(1-(\lambda q+\frac{1-\lambda}{3}))+2(\frac{1-\frac{2(1-h)}{3}}{3})(q-(\lambda q+\frac{1-\lambda}{3}))\underline{\phi}>0$. However, now the reduction in wages is $\frac{1-\frac{2(1-h)}{3}}{3}\alpha(1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}$, which is lower than in Proposition 3 by $(1-\frac{1-\frac{2(1-h)}{3}}{3})\alpha(1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}$.

Despite client welfare being higher, there are two costs. First, revenue is r^{da} rather than r^a. Second, the agent must be paid a higher wage than in Proposition 3. This is because he makes less of a difference. Pareto efficient assignment reduces the marginal value of the agent: in effect, a fraction $\frac{2h}{3}$ clients determine their own service and the agent is less pivotal. The wage must increase by $(1-\frac{1-\frac{2(1-h)}{3}}{3})\alpha(1-\lambda)(1-\frac{\lambda}{3}-(1-\lambda)\rho)\hat{\phi}$, and the number of clients covered falls accordingly. The firm prefers non-Pareto efficient Agent Assignment for low enough funding.[64] □

Proof of Proposition 4: Let $\hat{t}$ now refer to the good chosen by the client. Unrestricted Choice arises when the client chooses $t_i=\tau_i$ at a price r_i for all i. As $r_0\geq 0$, this implies that the Fairness Constraint can only be satisfied if $\frac{V'(\hat{y}+\overline{\eta})}{V'(\hat{y}+\underline{\eta})}\equiv\frac{\overline{m}}{\underline{m}}\geq\frac{2\mu-\underline{\phi}}{2\mu+\underline{\phi}}$. 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. This implies that there exists a critical average income $\hat{y}^{uc}$ such that $\frac{V'(\hat{y}+\overline{\eta})}{V'(\hat{y}+\underline{\eta})}\geq\frac{2\mu-\underline{\phi}}{2\mu+\underline{\phi}},\forall\hat{y}\geq\hat{y}^{uc}$. Lemma 1 generates the optimal contract, given by $R^*=\{r_0^*,\frac{\mu+\underline{\phi}}{\underline{m}},\frac{2\mu+\underline{\phi}}{\underline{m}}\}$, for all $\hat{y}\geq\hat{y}^{uc}$. Straightforward calculations show that $r_0^*=\frac{\underline{\phi}}{\underline{m}}$ above a critical average income $\tilde{y}^{uc}$ where $\frac{V'(\tilde{y}^{uc}+\overline{\eta})}{V'(\tilde{y}^{uc}+\underline{\eta})}=1-\frac{\underline{\phi}}{2\mu}$. Then revenue per client from the mechanism is $r^{uc}=\frac{3\mu+r_0^*+5\underline{\phi}}{3V'(y^{uc}-\underline{\eta})}$ for $y^{uc}\leq\hat{y}\leq\tilde{y}^{uc}$ and $r^{uc}=\frac{\mu+\underline{\phi}}{V'(y^{uc}-\underline{\eta})}$ for $y^{uc}>\tilde{y}^{uc}$. As $T^{uc}=T^*=\mu+\underline{\phi}$ and $I^{uc}=\frac{D}{\overline{U}-r^{uc}}$,

$$ES^{uc}=\frac{D}{\overline{U}-r^{uc}}\left(v+T^*-\frac{\delta D}{2(\overline{U}-r^{uc})}\right)-\frac{\delta\sigma_\epsilon^2}{2(\overline{U}-r^{uc})^2}.\square \quad (71)$$

[64] We have not considered an intermediate mechanism between DA and Agent Allocation, where the firm implements a DA mechanism using the agent's signal instead of its own. This is because there is little conceptually novel about it (and we know of no empirical examples). Instead, it offers an intermediate outcome where it allows Pareto efficient outcomes, but at the cost of increasing wages compared to Agent Assignment. This is shown in the Appendix. Agent allocation remains preferred when funding is low.

The Restricted Choice Mechanism: The firm knows I^{rc} when allocating services to all clients. The Restricted Choice mechanism is:

1. The firm chooses quantities κ_1^*, κ_2^* and unit prices $R^{rc} = (r_0^{rc}, r_1^{rc}, r_2^{rc})$.
2. All clients $i \leq I^{rc}$ report $\hat{t}$ and receive their first choice.
3. The firm uses a Deferred Acceptance mechanism to allocate the remaining $t = 1$ and $t = 2$ using prices $R^{da} = (r_1^{da}, r_2^{da})$.

The outcome is given in Proposition 5.

Proof of Proposition 5: *The firm chooses quantities of each* $t_i = (1, \kappa_1^*, \kappa_2^*)$ *at unit prices* $(r_0^{rc}, r_1^{rc}, r_2^{rc})$. 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. There are three cases:

- $\hat{t}(0, \underline{\phi}, m) = 0 \succ \hat{t}(0, \underline{\phi}, m) = 1$ if $\underline{\phi} - mr_0 \geq \kappa_1(\mu - mr_1), \forall m$, for $\frac{\underline{\phi}}{\underline{m}} \geq r_0 \geq 0$ and $r_1 \leq \frac{\mu + \underline{\phi}}{\underline{m}}$.
- $\hat{t}(0, \underline{\phi}, m) = 0 \succ \hat{t}(0, \underline{\phi}, m) = 2$ if $\underline{\phi} - mr_0 \geq \kappa_2(2\mu - mr_2)$ for $r_0 \geq 0$ and $r_2 \leq \frac{2\mu + \underline{\phi}}{\underline{m}}$.
- $\hat{t}(0, \underline{\phi}, m) = 1 \succ \hat{t}(0, \underline{\phi}, m) = 2$ if $\kappa_1(\underline{\phi} + \mu - mr_1) \geq \kappa_2(2\mu - mr_2), \forall m$, for $r_1 \leq \frac{\mu + \underline{\phi}}{\underline{m}}$ and $r_2 \leq \frac{2\mu + \underline{\phi}}{\underline{m}}$.

If the Fairness Constraint can be overcome, where all three conditions hold, then for the same reasons as in Lemma 1, $r_2^* = \frac{2\mu + \underline{\phi}}{\underline{m}}$, and $r_1^* = \frac{\mu + \underline{\phi}}{\underline{m}}$ as this maximizes revenue and the first two constraints. For $\hat{y} < \tilde{y}^{uc}$, the constraints on r_0 bind and

$$\underline{\phi} + \overline{m} r_0^* = \kappa_1^*(\mu + \overline{m}\frac{\mu + \underline{\phi}}{\underline{m}}), \tag{72}$$

and

$$\underline{\phi} + \overline{m} r_0^* = \kappa_2^*(2\mu + \overline{m}\frac{2\mu + \underline{\phi}}{\underline{m}}), \tag{73}$$

for $r_0^* \geq 0$. The benefits and costs of reducing r_0^* from $\underline{\phi}$ to 0 are linear, as the cost of diminishing returns is also linear, so there are two relevant corner solutions as optima, where $r_0^* = \frac{\underline{\phi}}{\underline{m}}$ and where $r_0^* = 0$. We consider each in turn.

First, consider κ_i at prices R^{uc}. Then (48) with $k_0 = 1$ implies that $(\hat{t}|\tau = 0) = 0 \succ (\hat{t}|\tau = 0) = 2$ iff

$$(1 - \frac{m}{\underline{m}})\underline{\phi} \geq \kappa_2(2\mu - \frac{m}{\underline{m}}(2\mu + \underline{\phi})). \tag{74}$$

for all m. Then (48) implies that maximum κ_2^* is

$$\kappa_2^* = \frac{1}{\frac{2\mu}{\underline{\phi}} - \frac{\overline{m}}{\underline{m}-\overline{m}}} = \frac{\underline{\phi}}{2\mu - \frac{\overline{m}\underline{\phi}}{\underline{m}-\overline{m}}}. \tag{75}$$

There are two necessary conditions for κ_1. (48) with $k_0 = 1$ implies that $(\hat{t}|\tau = 0) = 0 \succ (\hat{t}|\tau = 0) = 1$ iff

$$(1 - \frac{m}{\underline{m}})\underline{\phi} \geq \kappa_1(\mu - \frac{m}{\underline{m}}(\mu + \underline{\phi})) \tag{76}$$

for all m. This requires that

$$\kappa_1^* \leq \frac{1}{\frac{\mu}{\underline{\phi}} - \frac{\overline{m}}{\underline{m}-\overline{m}}}. \tag{77}$$

Finally, (48) with $k_0 = 1$ implies that $(\hat{t}|\tau = 1) = 1 \succ (\hat{t}|\tau = 1) = 2$ iff

$$\kappa_1(\mu + \underline{\phi} - \frac{m}{\underline{m}}(\mu + \underline{\phi})) \geq \kappa_2(2\mu - \frac{m}{\underline{m}}(2\mu + \underline{\phi})). \tag{78}$$

for all m. This implies that

$$\kappa_1^* \geq \frac{\kappa_2(2\mu - \frac{m}{\underline{m}}(2\mu + \underline{\phi}))}{(\mu + \underline{\phi} - \frac{m}{\underline{m}}(\mu + \underline{\phi}))}. \tag{79}$$

The two requirements for κ_1 hold for $\kappa_1 \in [\frac{\kappa_2(2\mu - \frac{m}{\underline{m}}(2\mu+\underline{\phi}))}{(\mu+\underline{\phi}-\frac{m}{\underline{m}}(\mu+\underline{\phi}))}, \frac{1}{\frac{\mu}{\underline{\phi}} - \frac{\overline{m}}{\underline{m}-\overline{m}}}]$ if $\frac{1}{\frac{\mu}{\underline{\phi}} - \frac{\overline{m}}{\underline{m}-\overline{m}}} > \frac{\kappa_2(2\mu - \frac{m}{\underline{m}}(2\mu+\underline{\phi}))}{(\mu+\underline{\phi}-\frac{m}{\underline{m}}(\mu+\underline{\phi}))}$. This always holds as $\kappa_1(\mu + \underline{\phi} - \frac{m}{\underline{m}}(\mu + \underline{\phi})) \geq \kappa_2(2\mu - \frac{m}{\underline{m}}(2\mu + \underline{\phi})$ when evaluated at κ_2^* and κ_1^* in (77). Hence κ_i^* are given by (75) and (77). As v_i is decreasing in i, the firm has (75) and (77) hold with equality.

Now consider the Deferred Acceptance component. The mechanism begins with clients $i = I^{rc} + 1, I^{rc} + 2, ...I^{rc}(1 + \frac{1-\kappa_1^*}{3} + \frac{1-\kappa_2^*}{3})$ reporting $\hat{t}$, where they only rank $t = 1$ and $t = 2$. The firm chooses prices $\tilde{R}^{da} = \{\infty, \frac{\mu}{\underline{m}}, \frac{2\mu}{\underline{m}}\}$. At prices $\tilde{R}^{da}$ (45) and (46) hold so $\hat{t}(\tau, \underline{\phi}, m) = (2 \succ 1)$, $\hat{t}(2, \overline{\phi}, m) = (2 \succ 1)$, $\hat{t}(1, \overline{\phi}, m) = (1 \succ 2)$, and $\hat{t}(0, \overline{\phi}, m) = (2 \succ 1)$ for all m. Note that

$$\hat{t} = \begin{cases} 2 \succ 1 & \text{with probability } 1 - \frac{h}{3} \\ 1 \succ 2 & \text{with probability } \frac{h}{3} \end{cases} \tag{80}$$

There are $I^{rc}(\frac{1-\kappa_2^*}{3})$ available $t = 2$ slots and $I^{rc}(1-\frac{h}{3})(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3})$ clients for whom $\hat{t} = 2 \succ 1$. Of those who have $\hat{t} = 2 \succ 1$, there are $I^{rc}(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3})(q+(2-h)\frac{1-q}{2})$ for whom $\tau^a = 2$. There are two cases. First, if $(q+(2-h)\frac{1-q}{2})(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3}) \geq (\frac{1-\kappa_2^*}{3})$, then the firm allocates $t = 2$ to $I^{rc}\frac{1-\kappa_2^*}{3}$ clients for which $\tau^a = 2$ and $\hat{t} = (2 \succ 1)$

and $t=1$ randomly to all other clients. Let $\hat{p}_1 = \frac{\frac{1-\kappa_2^*}{3}}{(q+(2-h)\frac{1-q}{2})(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3})}$. This has expected per capita treatment surplus of

$$\tilde{T}^{da} = \mu + (\frac{1-\kappa_2^*}{2-\kappa_1^*-\kappa_2^*})(\frac{q}{q+(2-h)\frac{1-q}{2}})\hat{\phi} + (\frac{1-\kappa_1^*}{2-\kappa_1^*-\kappa_2^*})(\hat{p}_1(\frac{1-q}{2}(1-h)\underline{\phi}) + (1-\hat{p}_1)(h\overline{\phi}+(1-h)q\underline{\phi})) \quad (81)$$

Second, if $(q+(2-h)\frac{1-q}{2})(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3}) < (\frac{1-\kappa_2^*}{3})$, then the firm allocates $t=2$ to all clients for which $\tau^a = 2$. It offers $t=2$ randomly to $I^{rc}(\frac{1-\kappa_2^*}{3} - (q+(2-h)\frac{1-q}{2})(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3}))$ clients for whom $\hat{t}=2 \succ 1$ and $\tau^a = 1$. The probability that a such a client receives an offer of $t=2$ is $\hat{p}_2 = \frac{\frac{1-\kappa_2^*}{3} - (q+(2-h)\frac{1-q}{2})(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3})}{(1-h)q+\frac{1-q}{2}(2-h)} < 1$. This has expected per capita treatment surplus of

$$\tilde{T}^{da} = \mu + (\frac{1-\kappa_2^*}{2-\kappa_1^*-\kappa_2^*})[\hat{p}_2(\frac{q}{q+(2-h)\frac{1-q}{2}})\hat{\phi} + (1-\hat{p}_2)\frac{1-q}{2}\hat{\phi}] + (\frac{1-\kappa_1^*}{2-\kappa_1^*-\kappa_2^*})(h\overline{\phi}+(1-h)q\underline{\phi}). \quad (82)$$

This has expected per capita treatment surplus of

$$\tilde{T}^{da} = \mu + (\frac{1-\kappa_2^*}{2-\kappa_1^*-\kappa_2^*})(\frac{q}{q+(2-h)\frac{1-q}{2}})\hat{\phi} + (\frac{1-\kappa_1^*}{2-\kappa_1^*-\kappa_2^*})(h\overline{\phi}+(1-h)q\underline{\phi}) + (1-h)\frac{1-q}{2}\underline{\phi}\left(1 - \frac{\frac{1-\kappa_2^*}{3}}{(q+(2-h)\frac{1-q}{2})(1+\frac{1-\kappa_1^*}{3}+\frac{1-\kappa_2^*}{3})}\right). \quad (83)$$

The Artificial Currency Game: The firm offers supply of $t_0 = \frac{I^{rc}}{3}$, $t_1 = \frac{\kappa_1^* I^{rc}}{3}$ and $t_2 = \frac{\kappa_2^* I^{rc}}{3}$ at unit fiat money prices of $R = R^*$. Each client $i = 1, 2, ...I^{rc}$ is given 1 unit of artificial currency to spend only on services. Each client reports a minimum quantities λ_i of each t_i that they are willing to offer for their unit of currency at a unit fiat money price of r_i^{1st}. 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 (t_0^*, t_1^*, t_2^*). Then $(t_0^*, t_1^*, t_2^*) = (1, \frac{1}{\kappa_1^*}, \frac{1}{\kappa_2^*})$ is a competitive equilibrium outcome. Consider a client i of type τ_{ik}, $k = 0, 1, 2$. Then consider the following equilibrium strategy: $\lambda_k = 0$ and $\lambda_{-k} = 1$. With these strategies, there is demand for each good of $\frac{I^{rc}}{3}$ at a minimum allocation of 0, so with equal rationing the equilibrium quantities of each good are $(t_0^*, t_1^*, t_2^*) = (1, \frac{1}{\kappa_1^*}, \frac{1}{\kappa_2^*})$. 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 r_i^{1st}, where $\kappa_0^* = 1$ to any κ_{-i}^*. Hence this is a competitive equilibrium.

Proof of Proposition 6: We show this in two parts. First, we show existence of an 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. Second, we show that there cannot be an equilibrium where a type other than $\tau = 1$ clients choose $t = 1$.

To show the first part, we consider incentive compatibility in this proposed equilibrium at the contract R^b. First, type $\tau = 2$ does not choose $t = 1$ if

$$\mu(1 - \frac{\overline{m}}{\underline{m}}) \leq (q + \frac{1-q}{4})(2\mu(1 - \frac{\overline{m}}{\underline{m}}) + \phi. \tag{84}$$

which always holds. Second, type $\tau = 1$ chooses $t = 1$ if

$$\mu(1 - \frac{\overline{m}}{\underline{m}}) + \phi \geq (\frac{q}{2} + \frac{1-q}{2})(2\mu(1 - \frac{\overline{m}}{\underline{m}}), \tag{85}$$

which always holds. Finally, a client with preferences $\underline{\phi}$ and $\tau = 0$ does not choose $t = 1$ only if

$$\mu(1 - \frac{\overline{m}}{\underline{m}}) \leq \frac{3(1-q)}{4})(2\mu(1 - \frac{\overline{m}}{\underline{m}}) + (q + \frac{1-q}{2})\underline{\phi}, \tag{86}$$

which holds only if

$$\frac{\overline{m}}{\underline{m}} \geq 1 - \frac{(1+3q)\underline{\phi}}{2\mu(3q-1)}. \tag{87}$$

Note that $1 - \frac{(1+3q)\underline{\phi}}{2\mu(3q-1)} < 1 - \frac{\underline{\phi}}{\overline{\mu}}$, which defined $\tilde{y}^{uc}$. If these conditions hold, then an equilibrium exists a prices R^b. But if all are slack, the firm can increase r_1 above $r_1 = \frac{\mu}{\underline{m}}$, thereby showing $r^{1st} \geq r^b$.

Partial Choice entails that one $\underline{\phi}$ type chooses their allocation $t = \tau_i$ and the other two are assigned by Deferred Acceptance. We first show that the only Partial Choice equilibrium is where type $\tau = 1$ choose $t = 1$ by ruling out the other two types.

1. Type $\tau = 0$. In order for type $\tau = 0$ to choose $t = 0$ it must be that

$$\underline{\phi} - mr_0 \geq min\{\mu - mr_1, 2\mu - mr_2\} \tag{88}$$

for all m. As $t = 1$ and $t = 2$ are assigned by DA, $r_1 \leq \frac{\mu}{\underline{m}}$ and $r_2 \leq \frac{2\mu}{\underline{m}}$ as services are misallocated with positive probability. Similarly, $r_1 \geq 0$. This implies that

$$\frac{\overline{m}}{\underline{m}} \geq min\{\frac{2\mu - \underline{\phi}}{2\mu}, \frac{\mu - \underline{\phi}}{\mu}\}. \tag{89}$$

Note however that the firm can use Unrestricted Choice, which has higher payoff, if $\frac{\overline{m}}{\underline{m}} \geq \frac{2\mu - \underline{\phi}}{2\mu + \underline{\phi}}$. But $\frac{2\mu - \underline{\phi}}{2\mu + \underline{\phi}} < min\{\frac{2\mu - \underline{\phi}}{2\mu}, \frac{\mu - \underline{\phi}}{\mu}\}$ 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. □

Proof of Proposition 7: For any assignment mechanism A,

$$S^* = \int_{\underline{\epsilon}}^{\overline{\epsilon}} \frac{(\mathbb{D}(b) - \pi^A(x) + x)}{w^A - r^A}(v + T^A - \frac{\delta(\mathbb{D}(b) - \pi^A(x) + x)}{2(w^A - r^A)})f(x)dx, \tag{90}$$

where $\pi(x) = \pi(\tilde{I} - E\tilde{I})$, as $E\pi = 0$. The firm chooses $\pi^*(\tilde{I})$ holding donations fixed, though in equilibrium $\hat{\pi}^A = \int_{\underline{\epsilon}}^{\overline{\epsilon}} \pi^{*\prime A}(\tilde{I}(x))f(x)dx$. As $E\pi = 0$ this simplifies to

$$S^* = \frac{\mathbb{D}(b)}{w^A - r^A}(v + T^A - \frac{\delta\mathbb{D}(b)}{2(w^A - r^A)}) - \int_{\underline{\epsilon}}^{\overline{\epsilon}} \frac{\delta(-\pi(x) + x)^2}{w^A - r^A} f(x)dx. \tag{91}$$

This is maximized at $\pi^*(\epsilon) = \epsilon$ or $\pi^{*\prime} = 1$ with equilibrium surplus

$$S^* = \frac{\mathbb{D}(b)}{w^A - r^A}(v + T^A - \frac{\delta\mathbb{D}(b)}{2(w^A - r^A)}). \tag{92}$$

As $\hat{\pi}^A = \int_{\underline{\epsilon}}^{\overline{\epsilon}} \pi^{*\prime A}(\tilde{I}(x))f(x)dx = 1$, this implies that $d = 0$. □

Proposition 10 *The link between profit status and assignment mechanisms is as follows:*

- *If $\tilde{m}^{da} < V'(y^{uc} + \underline{\eta})$, no non-profit uses Unrestricted Choice. Then if*
  - *If $\tilde{m}^{uc} < V'(y^{uc} + \underline{\eta})$, some non-profit firms and all for-profit firms use Unrestricted Choice,*
  - *If $\tilde{m}^{uc} \geq V'(y^{uc} + \underline{\eta})$, all firms using Unrestricted Choice are for-profit.*
- *If $\tilde{m}^{da} \geq V'(y^{uc} + \underline{\eta})$, for profit firms can use any assignment mechanism.*

Proof of Proposition 10:

- *If $\tilde{m}^{da} < V'(y^{uc} + \underline{\eta})$, no non-profit uses Unrestricted Choice.* The firm always uses Unrestricted Choice if $\underline{m} \geq V'(y^{uc} + \underline{\eta})$. As a result, if $\tilde{m}^{da} < V'(y^{uc} + \underline{\eta}), \nexists \underline{m} \in [V'(y^{uc} + \underline{\eta}), \tilde{m}^{da}]$ where a for-profit firm uses DA.
  Then if
  - *If $\tilde{m}^{uc} < V'(y^{uc} + \underline{\eta})$, some non-profit firms and all for-profit firms use Unrestricted Choice,* The firm always uses Unrestricted Choice if $\underline{m} \geq V'(y^{uc} + \underline{\eta})$. If $\tilde{m}^{uc} < V'(y^{uc} + \underline{\eta}), \exists \underline{m} \in [V'(y^{uc} + \underline{\eta}), \tilde{m}^{uc}]$ where a non-profit firm uses Unrestricted Choice.

– *If $\tilde{m}^{uc} \geq V'(y^{uc} + \underline{\eta})$, all firms using Unrestricted Choice are for-profit.* The firm always uses Unrestricted Choice if $\underline{m} \geq V'(y^{uc} + \underline{\eta})$. If $\tilde{m}^{uc} \geq V'(y^{uc} + \underline{\eta}), \nexists \underline{m} \in [V'(y^{uc} + \underline{\eta}), \tilde{m}^{uc}]$ where a non-profit firm uses Unrestricted Choice.

- If $\tilde{m}^{da} \geq V'(y^{uc} + \underline{\eta})$, for profit firms can use any assignment mechanism. The firm always uses Unrestricted Choice if $\underline{m} \geq V'(y^{uc} + \underline{\eta})$. If $\tilde{m}^{da} \geq V'(y^{uc} + \underline{\eta}), \exists \underline{m} \in [\tilde{m}^{da}, V'(y^{uc} + \underline{\eta})]$ where a for-profit firm uses DA or First Choice.

Agency Contracting when all revenue can be paid out: Here the firm pays itself all revenue as a Bonus if I high enough: $\pi(\overline{I}) = d$ for enough high I.

$$\pi = \begin{cases} \mathbb{D}(b) - \overline{\pi} & \text{if } I \geq \hat{I} \\ -\overline{\pi} & \text{if } I < \hat{I} \end{cases} \tag{93}$$

where $\overline{\pi} = \int_{I^*}^{\overline{I}} d(\underline{m}) f(x) dx$ to ensure that $E\pi = 0$. Then $\hat{I}$ is set such that

$$f_e(\hat{I})\mathbb{D}(b) = \\ \mathbb{D}(b)(v + T^A) \left(\frac{w^A + C(e^*) - r^A e^* - e^* C'(e^*)}{(w^A + C(e^*) - r^A e^*)^2} - \frac{w^A + C(\tilde{e}) - r^A \tilde{e} - \tilde{e} C'(\tilde{e})}{(w^A + C(\tilde{e}) - r^A \tilde{e})^2} \right). \tag{94}$$

Here $\mathbb{D}(b) = \mathbb{D}(b) F(\hat{I})$. □

When a Fixed Fee can be charged to clients: Assume that the firm chooses prices $\hat{R} = \{\hat{r}, \hat{r}_0, \hat{r}_1, \hat{r}_2\}$. Let $U^A(0, \underline{m})$ is the expected utility of a client with type $\tau = 0$ and income $\underline{m}$ under assignment mechanism A. Then for any Assigned Consumption allocation other than Deferred Acceptance,

$$\hat{R} = \{\hat{r}, 0, \frac{\mu}{\underline{m}}, \frac{2\mu}{\underline{m}}\}, \tag{95}$$

where $\hat{r} = \frac{U^a(0, \underline{m}) - E(r_i | \tau = 0)}{\underline{m}}$.

Proposition 11 *Assume that a fixed fee can be charged. The expected surplus from Deferred Acceptance is*

$$ES^{da} = \frac{D}{\overline{U} - r^{da}} \left(v + T^{da} - \frac{\delta D}{2(\overline{U} - r^{da})} \right) - \frac{\delta \sigma_\epsilon^2}{2(\overline{U} - r^{da})^2}. \tag{96}$$

where $r^{da} = \frac{\mu - \underline{\phi} + \underline{\phi}(1-q)(1+p_2)}{\overline{m}}$.

Proposition 12 *Assume a fixed fee can be charged. The expected surplus from Agent Allocation is*

$$ES^{da} = \frac{D}{w^* - r^a}\left(v + T^a - \frac{\delta D}{2(w^* - r^a)}\right) - \frac{\delta\sigma_\epsilon^2}{2(w^* - r^a)^2}. \tag{97}$$

where $r^a = \frac{\mu+\phi(1-(q\rho+\frac{(1-q)(1-\rho)}{2}))}{\underline{m}}$.

The qualitative outcomes of each are unchanged.