Efficient consignment auctions*

Bing Liu[†] Simon Loertscher[‡] Leslie M. Marx[§]

June 19, 2023

Abstract

Consignment auctions are two-stage mechanisms to (re)-allocate emission permits. Firms are first endowed with permits and then allowed to trade them. We determine theoretically endowments that enable efficient allocation, subject to incentive compatibility, individual rationality, and no deficit. All firms prefer efficient consignment auctions to efficient standard auctions, making them politically palatable. Firms' investment incentives align with the first-best in efficient consignment auctions. Grandfathering based on efficient long-run allocations induces efficiency-permitting endowments. A simple calibration to data from Southern California's RECLAIM program validates our no-deficit assumption and shows that grandfathering provides the best theoretical match for the empirically observed endowments.

[*] We are grateful to three anonymous referees and the editor, Kareen Rozen, for comments and feedback that have helped us to improve the paper. We also thank David Byrne, Peter Cramton, Toan Le, Andy Wu and seminar audiences at the Stanford Theory Lunch, the Toulouse School of Economics and at the University of Melbourne for insightful discussions and suggestions. Financial support from the Samuel and June Hordern Endowment, a University of Melbourne Faculty of Business & Economics Eminent Research Scholar Grant, and Australian Research Council Discovery Project Grant DP200103574 is gratefully acknowledged.

[†] Department of Economics, Stanford University, Stanford, CA 94305. Email: bingliu@stanford.edu.

[‡] Department of Economics, Level 4, FBE Building, 111 Barry Street, University of Melbourne, Victoria 3010, Australia. Email: simonl@unimelb.edu.au.

[§] Duke University, 100 Fuqua Drive, Durham, NC 27708, USA. Email: marx@duke.edu.

Keywords: cap-and-trade, transaction costs, market mechanisms, optimal endowments, grandfathering, benchmarking, Coase Theorem, RECLAIM permit market

JEL Classification: C72, D44, L13

1 Introduction

Markets for emission permits often involve a central authority that first endows firms with permits and then allows these firms to trade those permits in a secondary market into which the authority may sell additional permits. Customarily, these markets are called *consignment auctions* or cap-and-trade systems. From a theoretical perspective, consignment auctions raise the question of why such two-stage procedures are employed rather than simply auctioning off the permits in the first place, thereby avoiding problems associated with the impossibility of efficient secondary markets as emphasized by Vickrey (1961) and Myerson and Satterthwaite (1983).[1] Given that a consignment auction is to be used, the question of how initial endowments should be chosen is of considerable practical relevance. The appropriateness of determining endowments based on historical allocations, often referred to as *grandfathering*, is being debated, and questions arise regarding how permit markets affect firms' incentives to invest in abatement technologies.[2]

In this paper, we analyze an independent private values model that allows us to respond to these questions within a framework in which transaction costs, like in the work of Vickrey and Myerson and Satterthwaite, arise from agents' private information. The model assumes that all firms have privately known marginal values for permits. These values are drawn independently from continuous distributions and are constant up to a maximum demand. Distributions and maximum demands are common knowledge among the firms and the cen-

[1] Of course, reallocation mechanisms such as the "incentive auction" (see e.g., Milgrom, 2017) are sensible in the presence of unforeseen and unforeseeable technological change, in which case resources cannot be allocated efficiently in the first place; see also Loertscher et al. (2015). However, this is not the case here because the reallocation mechanism could be dispensed with if an efficient primary market mechanism were used.

[2] See, for example, Cramton and Kerr (2002), Burtraw and McCormack (2017), and Hahn and Stavins (2011).

tral authority, as is the fact that marginal values are constant.

We first show that for any allocation problem of this form, there always exists a vector of endowments such that first-best (ex post efficient) reallocation can be achieved while respecting the firms' incentive compatibility and interim individual rationality constraints without running a deficit.[3] We also show that the total revenue extracted from the firms decreases in the share of permits with which they are endowed, which provides an explanation for why these schemes are used in the first place—they are politically more palatable than pure auctions because they tax the firms less.[4] Moreover, if the allocation mechanism always achieves the first-best, then each firm's incentive to invest in abatement technologies aligns with those of a designer that is interested in social surplus maximization. Hence, within this framework, any argument about suboptimal investment incentives must involve an argument as to why the secondary market is not efficient.

Our analysis allows us to relate grandfathering schemes to endowments that permit first-best reallocation via the secondary market. In a stationary environment and abstracting away from incentives to boost consumption to increase endowments in the future, endowing each firm with its long-run average consumption allows the secondary market to achieve the first-best without running a deficit on average if the past allocations were efficient. The

[3] Respecting individual rationality constraints is a natural requirement when trade and participation are voluntary as is the case, for example, in the secondary market in Southern California's RECLAIM scheme (see Section 4). But even when, in principle, the agents could be coerced to participate, such as in public good problems, respecting individual rationality is sensible because it guarantees political acceptance of the scheme by safeguarding the agents; see Hellwig (2003) for an insightful discussion.

[4] The issue of the need for political buy-in is raised by Hahn and Shapiro (2011), who mention the possibility of "leaving it up to the legislature to construct a constituency in support of the program by allocating the allowances to various interests" (Hahn and Shapiro, 2011, p. S267).

model thus provides a theoretical rationale for grandfathering.[5] We also analyze endowments when firms can be overendowed, that is, in the first stage some firms may obtain permits in excess of their maximum demands.

A simple calibration of our model to data from the RECLAIM nitrogen oxides (NOx) and sulfur oxides (SOx) permit market in Southern California, which were collected and analyzed by Fowlie and Perloff (2013), validates our assumption that the reallocation mechanism does not run a deficit. Specifically, given the estimated distributions and empirical endowments, the computed revenue is always positive. The calibration also shows that of the theoretical endowments we derive, those under grandfathering are the best match to the empirical ones, both when measured by the revenue generated and by the dissimilarities among the distributions of endowments. Further, our empirical analysis provides some evidence of overendowments and thereby support for the empirical relevance of this possibility.

There is a rich body of literature on emission trading. Xiong et al. (2017) provide an overview of the carbon allowance allocation mechanisms used in the EU and California and in China's pilot programs. Grandfathering is used in the early stages of these programs, for reasons such as political palatability and lack of data.[6] The analysis in our paper gives an efficiency justification to using grandfathering: it achieves efficient reallocation of the carbon permits even when the authority lacks data on the individual emission abatement costs. Busch et al. (2018) raise concerns that free allocation using grandfathering can lead to thin secondary markets and poor carbon price discovery, but recognizes that using a centralized consignment auction addresses the thin market problem. Our paper furthers the analysis by showing that grandfathering is not the cause of a thin secondary market.

[5] To our knowledge, Segal and Whinston (2011) were the first to observe that what amounts to grandfathering in consignment auctions can achieve the first-best.

[6] Grandfathering can be viewed as driven by a lack of data insofar as the designer does not need to know the distributions from which the agents draw their types. It suffices to observe average consumption from historical data.

Besides grandfathering, benchmarking—allocating permits to firms according to observable characteristics such as their industrial sector and production technology—is used as another method to determine the initial allocation of carbon allowances (Xiong et al., 2017; Goulder and Morgenstern, 2018; Pizer and Zhang, 2018). We determine theoretically the optimal endowments when the firms belong to strength-ordered groups, where all firms within a given group have the same maximum demand and draw their values from the same distribution, with firms in stronger groups having larger maximum demands and first-order stochastically dominant distributions. This resonates with the benchmarking idea because agents within each group are treated equally, with the benchmark used for each group stemming from the distribution of the abatement costs of that group. Our calibration exercise is also based on a specification with heterogeneous groups, with firms within each group having the same distribution and the same maximum demand.

Related to consignment auctions, some prior literature dismisses the relevance of initial endowments by referencing Coase theorem arguments (see, e.g., Montgomery, 1972). But, of course, transaction costs remain a feature of the real world,[7] as well as a feature of theoretical models, including ours, in which transaction costs arise from agents' private information. Recent theoretical analyses of consignment auctions include Khezr and MacKenzie (2018) and Liu and Tan (2021), who study consignment auctions with a uniform price assuming, respectively, constant common values and decreasing private values. Our paper departs from this approach by focusing on efficient, incentive compatible mechanisms, which by Holmström (1979) means focusing on Groves' mechanisms. Making the interim individual rationality constraint binding so as not to "leave money on the table,"[8] this further narrows down the

[7] Hahn and Shapiro (2011) note the violation of conditions for the Coase theorem in the context of cap-and-trade systems, saying that they may be violated "when there are transaction costs, when firms have market power, and when firms are subject to differential regulatory treatment" (Hahn and Shapiro, 2011, p. S267).

[8] That is the interim individual rationality constraint of each agent is binding when eval-

mechanism to the Vickrey-Clarke-Groves (VCG) mechanism (see Vickrey, 1961; Clarke, 1971; Groves, 1973). While our approach is more abstract, it allows us to connect consignment auctions to the mechanism design literature on partnership dissolution and asset markets, initiated by Cramton et al. (1987) with subsequent contributions by Lu and Robert (2001), Che (2006), Figueroa and Skreta (2012), Loertscher and Wasser (2019), and Loertscher and Marx (2020). This permits us to relate the efficient operation of the secondary market to the initial endowments, make statements about the optimal endowments and grandfathering, and discuss investment incentives.

Our paper is also related to Loertscher and Marx (2020) and Delacrétaz et al. (2022), which study, respectively, trade sacrifice and efficient mechanisms for asset markets and focus on dominant strategies and ex post individual rationality. Delacrétaz et al. (2022) show that there is never a budget surplus under efficiency. Thus, our possibility results derive from the fact that we only require interim individual rationality. Finally, the simple status quo that ensures voluntary participation in efficient bargaining derived by Segal and Whinston (2011) is equivalent to grandfathering. From a theoretical perspective, our paper adds to this conditions—for example, two agents and identical distributions—under which grandfathering yields the revenue-maximizing endowments.

The remainder of this paper is organized as follows. The setup is introduced in Section 2. Section 3 presents the main results. In Section 4, we apply the model to data from Southern California's RECLAIM pollution permit trading program. Section 5 extends the model to allow for investment, and Section 6 concludes the paper. Further extensions and results are contained in the Online Appendix.

uated at this agent's worst-off type.

2 Setup

We assume that there is set of n agents denoted by $\mathcal{N}$ and a homogeneous good. In our application, the agents are emitting firms and the good is emission permits. Our assumptions allow the possibility that each emitting firm i is known to have a minimum permit requirement that is essential to its business and also a privately known value for a range of additional permits. Consistent with this, we assume that the willingness to pay for each agent i is fixed at $\overline{\theta} > 0$ for the first $d_i \geq 0$ units and equal to θ_i for an additional $k_i > 0$ units, where θ_i is drawn independently from the continuous distribution F_i with identical support $[\underline{\theta}, \overline{\theta}]$ and density f_i that is positive on the interior of the support. While the distributions F_i are assumed to be commonly known, the realization of θ_i is agent i's private information. Further, agent i's willingness to pay for units beyond the $d_i + k_i$-th unit is assumed to be zero. Thus, we refer to d_i as agent i's *minimal demand*, to k_i as agent i's *variable demand*, and to $d_i + k_i$ as agent i's *total demand*. Agents in this model exhibit decreasing marginal values for the good, consistent with the data in our application in Section 4.

The total supply of the good is assumed to be sufficient to cover agents' minimal demands, but not sufficient to cover agents' total demand. Given this, it is without loss of generality to focus on the allocation of emission permits above and beyond agents' minimal demands because any efficient mechanism will first satisfy all agents' minimal demands and then allocate the remaining permits according to the agents' private types for additional units.

We consider a consignment auction, described in detail below, in which the agents are first endowed with units of the good and then, following the realization of their private information, participate in an auction that facilitates trading of the good among the agents. We normalize the total supply in excess of agents' minimum demands, denoted by R, to be equal to 1, and we denote by $\mathbf{r} = (r_1, \ldots, r_n)$ the initial endowment of permits to the agents above and beyond their minimal demands, where $\sum_{i \in \mathcal{N}} r_i = R$.[9] Given our assumptions, we

[9] In Section 3.3, we generalize the model by allowing the designer to withhold part of the

have:[10]

$$R = 1 < \sum_{i \in \mathcal{N}} k_i.$$

The agents' total endowments are assumed to be sufficient to at least cover their minimal demands, which implies that $r_i \geq 0$, and, except where noted (e.g., Section 3.4), we assume that no agent's total endowment exceeds its total demand, implying that $r_i \leq k_i$. That is, no agent is endowed with less than what it must have to operate its business, and no agent is endowed with more than it could possibly use.

We refer to the game consisting of both the endowment of agents with units of the good and the subsequent auction-based trading as the *consignment auction*. Recognizing that downstream interactions may play an important role depending on the application, but are typically difficult to model, we focus on auction-based trading and refer to this as the *second-stage auction* or simply the *auction*. The timing is as follows: First, prior to the realization of agents' types, the designer chooses the endowment vector $\mathbf{r}$, which is observed by all. Second, after agents' private types are realized, the agents choose whether to participate in the auction.[11] We restrict attention to auctions that are ex post efficient and satisfy dominant strategy incentive compatibility and interim individual rationality constraints. Because agent

supply until the auction is run. That is, we allow for $\sum_{i \in \mathcal{N}} r_i \leq R$ there.

[10] More formally, if $\hat{R}$ is the total supply of the good, then $R \equiv \hat{R} - \sum_{i \in \mathcal{N}} d_i$, and if agents' total initial endowments are $\hat{\mathbf{r}} = (\hat{r}_1, \ldots, \hat{r}_n)$, then $r_i = \hat{r}_i - d_i$.

[11] As noted, agent i's endowment r_i is assigned before its type θ_i are realized. This assumption mimics an aspect of real-world emission permit markets in that tradeable permits tend to be issued at sufficiently long intervals that allowing trading of those permits in the interim makes economic sense. In light of this, it seems useful to view some details of a firm's value for permits as being unknown at the time that endowments are assigned, with actual values being realized at a later date. In this sense, one can view a firm's type θ_i as reflecting relatively short-term characteristics or information, whereas the assignment of endowments must rely on more permanent characteristics of firms, which we capture by having firms'

i obtains value $\min\{r_i, k_i\}\theta_i$ if it does not participate in the auction, the auction must offer agent i an expected payoff of at least this outside option in order for agent i's interim individual rationality constraint to be satisfied. As we show below, such an auction exists. We assume that the auction runs if and only if its expected revenue (conditional on the endowments $\mathbf{r}$ and taking expectations over agents' types) is nonnegative.[12]

Standard mechanism design results imply that the VCG mechanism maximizes ex ante expected revenue subject to ex post efficiency, (dominant-strategy) incentive compatibility, and interim individual rationality (see the Online Appendix, Section 1.1).[13] Thus, we assume that the designer uses the VCG mechanism for the auction. We denote the VCG mechanism by $\langle \mathbf{Q}, \mathbf{T_r} \rangle$, where $\mathbf{Q} : [\underline{\theta}, \overline{\theta}]^n \to \mathbb{R}^n_+$ is the ex post efficient allocation rule and $\mathbf{T_r} : [\underline{\theta}, \overline{\theta}]^n \to \mathbb{R}^n$ is the VCG payment rule (where payments are from agents to the market designer). For a given vector of reported types $\boldsymbol{\theta}$, maximal social surplus, denoted $W(\boldsymbol{\theta})$, is

$$W(\boldsymbol{\theta}) = \max_{\hat{\mathbf{Q}} \in \Delta, \hat{\mathbf{Q}} \leq \mathbf{k}} \sum_{i \in \mathcal{N}} \theta_i \hat{Q}_i(\boldsymbol{\theta}) = \sum_{i \in \mathcal{N}} \theta_i Q_i(\boldsymbol{\theta}),$$

where Δ is the $(n-1)$-dimensional simplex and the second equality holds because $\mathbf{Q}$ is the

distributions and maximum demands be common knowledge, including for the designer.

[12] As shown by, e.g., Börgers and Norman (2009), the mechanism can always be made to balance the budget with equality by returning any expected surplus to the agents through fixed payments.

[13] The notion of incentive compatibility is not material. The VCG mechanism being dominant-strategy incentive compatible implies that it is also Bayesian incentive compatible. With independent private values, for any efficient mechanism that is Bayesian incentive compatible, there exists an equivalent mechanism that is dominant strategy incentive compatible; see, e.g., Gershkov et al. (2013).

ex post efficient allocation rule. Given $\hat{\theta}_i \in [\underline{\theta}, \overline{\theta}]$, the VCG transfer from agent i is

$$T_{\mathbf{r},i}(\boldsymbol{\theta}) = W(\hat{\theta}_i, \boldsymbol{\theta}_{-i}) - (W(\boldsymbol{\theta}) - \theta_i Q_i(\boldsymbol{\theta})) - \hat{\theta}_i \min\{r_i, k_i\}.$$

Accordingly, the designer's revenue ex post, denoted $\Pi(\boldsymbol{\theta}, \mathbf{r})$, is

$$\Pi(\boldsymbol{\theta}, \mathbf{r}) = \sum_{i \in \mathcal{N}} T_{\mathbf{r},i}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{N}} W(\hat{\theta}_i, \boldsymbol{\theta}_{-i}) - (n-1)W(\boldsymbol{\theta}) - \sum_{i \in \mathcal{N}} \hat{\theta}_i \min\left\{r_i, k_i\right\}, \tag{1}$$

where the second equality follows from substituting $T_{\mathbf{r},i}(\boldsymbol{\theta})$. The transfer $T_{\mathbf{r},i}(\boldsymbol{\theta})$ depends both directly on r_i, as is evident from the definition, and indirectly because, as shown after (2) below, $\hat{\theta}_i$ is also a function of r_i.

By the payoff equivalence theorem (see, e.g., Börgers, 2015), any ex post efficient mechanism that is incentive compatible and that satisfies the interim individual rationality constraints with equality for the worst-off types induces the same interim payoffs, and thus ex ante expected payoffs, as the VCG mechanism. Consequently, focusing on the VCG mechanism is without loss of generality within the context of ex post efficient mechanisms. This thus leaves only the question of why one should focus on ex post efficient mechanisms since, notwithstanding the normative appeal of efficiency, there are so many ways in which markets can and arguably do depart from that benchmark. A priori, we see two reasons why efficiency is a good assumption. First, efficiency provides a well-defined benchmark whereas the multitude of reasons for which markets may operate inefficiently begs the question of which one analysts should zero in on. Lacking any direct evidence as to the source(s) of inefficiency, any alternative modeling assumption would, arguably, be stronger and seem more arbitrary than assuming efficiency. Second, efficiency is a distribution-free concept that permits tractable modeling. This contrasts with, say, second-best mechanisms à la Myerson and Satterthwaite (1983), in which the allocation rule is intertwined with the agents' distributions. Moreover, our application in Section 4 shows that the data are largely consistent with the assumption that the market reallocates efficiently.

3 Consignment auctions

In this section, we show that the market's expected and ex post revenue is nonnegative with appropriately chosen endowments. Under additional assumptions, we relate the choice of endowments to benchmarking and grandfathering. We also analyze hybrid consignment auctions, in which only a fraction of the total supply is allocated as endowments ex ante, with the remainder being auctioned off once private information has been realized.

3.1 Ex post and expected revenue

Denoting the interim expected efficient allocation for agent i with type θ_i by $q_i(\theta_i)$, we have

$$q_i(\theta_i) = \mathbb{E}_{\boldsymbol{\theta}_{-i}}[Q_i(\theta_i, \boldsymbol{\theta}_{-i})],$$

where, given the assumed properties of the distributions, $q_i(\theta_i)$ is continuous and increasing in θ_i. Given the assumptions of a maximum demand for agent i of k_i and marketwide scarcity, it follows that $q_i(\theta_i)$ is bounded above by k_i. Analogously, agent i's interim expected payment to the mechanism, $t_{\mathbf{r},i}(\theta_i)$, is

$$t_{\mathbf{r},i}(\theta_i) = \mathbb{E}_{\boldsymbol{\theta}_{-i}}[T_{\mathbf{r},i}(\theta_i, \boldsymbol{\theta}_{-i})].$$

Combining these, the interim expected payoff from the consignment auction of agent i with type θ_i, net of the agent's outside option, can thus be expressed as

$$q_i(\theta_i)\theta_i - t_{\mathbf{r},i}(\theta_i) - \min\{r_i, k_i\}\theta_i. \tag{2}$$

For each agent i, there exists a worst-off type, denoted by $\hat{\theta}_i$, which is the type that minimizes (2). As we show in the proof of Lemma 1, $\hat{\theta}_i$ is such that $q_i(\hat{\theta}_i) = \min\{r_i, k_i\}$. (This is a generalization of the result of Cramton et al. (1987) to heterogeneous distributions and maximum demands k_i less than 1.) Intuitively, an agent benefits from trade, both as a buyer or seller, so an agent's worst-off type is the one whose interim expected allocation in the second-stage auction is equal to its endowment or, if its endowment is greater than its maximum demand, then equal to its maximum demand. Because q_i is continuous and increasing in agent i's type, $\hat{\theta}_i$ is unique.

Lemma 1. *There exists a unique $\hat{\theta} \in (\underline{\theta}, \overline{\theta})$ and endowments $\mathbf{r}^* \equiv (q_1(\hat{\theta}), \ldots, q_n(\hat{\theta}))$ with $r_i^* \in [0, k_i]$ for all $i \in \mathcal{N}$ and $r_j^* \in (0, k_j)$ for some $j \in \mathcal{N}$ such that when the endowments are $\mathbf{r}^*$, all agents have the same worst-off type $\hat{\theta}$.*

Proof. See Appendix A.1.

By Lemma 1, endowments $\mathbf{r}^*$ exist that induce all agents to have the same worst-off type, even though the agents themselves are not necessarily symmetric. The following proposition shows that if no agent can have an endowment in excess of its maximum demand, then $\mathbf{r}^*$ maximizes the expected revenue from the second-stage auction, $\Pi_{\mathbf{r}} \equiv \mathbb{E}_{\boldsymbol{\theta}}[\Pi(\boldsymbol{\theta}, \mathbf{r})]$.

Proposition 1. *If $r_i \leq k_i$ has to hold for all $i \in \mathcal{N}$, then endowments $\mathbf{r}^*$ maximize the expected revenue from the second-stage auction $\Pi_{\mathbf{r}}$.*

Proof. See Appendix A.2.

The intuition for why equalizing the worst-off types maximizes revenue is based on an envelope-type argument. The only effect that changes in r_i have on expected revenue under ex post efficiency is the direct effect on the worst-off types. Because each agent i's worst-off type increases in r_i and only depends on r_i, expected revenue $\Pi_{\mathbf{r}}$ is strictly concave in $\mathbf{r}$. Hence, it is maximized when $\mathbf{r}$ is such that all worst-off types are the same. Che (2006) and Figueroa and Skreta (2012) use the property that equal worst-off types maximize the designer's revenue in a partnership dissolution model (that is $k_i = 1$ for all $i \in \mathcal{N}$), anticipating that an efficient dissolution mechanism is used.[14] Loertscher and Wasser (2019)

[14] Of course, in the original partnership model of Cramton et al. (1987), the worst-off

show that, within a partnership model, this property generalizes beyond the use of an efficient dissolution mechanism. For example, anticipating that, at the dissolution stage, the designer maximizes its revenue, the revenue-maximizing partnership shares are such that all agents with positive shares have the same worst-off type.[15] That said, none of these papers studies a more general settings than ours because they all impose $k_i = 1$ for all i, which also means that $r_i \leq k_i$ is automatically satisfied in these earlier works. Because $\mathbf{r}^*$ maximizes $\Pi_{\mathbf{r}}$, subject to $r_i \leq k_i$ for all $i \in \mathcal{N}$, we refer to $\mathbf{r}^*$ as the optimal endowments.[16]

Theorem 1. *For all* $\boldsymbol{\theta}$, $\Pi(\boldsymbol{\theta}, \mathbf{r}^*) \geq 0$*; moreover,* $\Pi_{\mathbf{r}^*} > 0$.

Proof. See Appendix A.3.

The proof of Theorem 1 generalizes the revealed-preference argument used by Che (2006) to prove the possibility of efficient dissolution in a partnership model with heterogeneous type distributions, that is, for a model with $k_i = 1$ for all $i \in \mathcal{N}$ that permits $F_i \neq F_j$. Theorem 1 generalizes Cramton et al. (1987) by allowing for $k_i \neq k_j$ and $F_i \neq F_j$.[17]

While Myerson and Satterthwaite (1983) show that with one buyer and one seller, i.e., $\mathbf{r} = (0, 1)$, efficient trade is not possible without running a deficit, Theorem 1 shows that

types are also the same when all agents have the same ownership share, given that their distributions are all the same.

[15] Lu and Robert (2001) allow for $k_i \neq k_j$ and rent extraction by the designer, but impose identical distributions and do not optimize over endowments.

[16] The terminology of "optimal endowments" is in line with the mechanism design literature, which refers to mechanisms that maximize revenue as optimal (see e.g. Myerson, 1981). In our setting, however, the second-stage mechanism itself is not optimal in this sense because it maximizes social surplus.

[17] Other papers that provide conditions for the first-best to be possible include Makowski and Mezzetti (1994), Williams (1999), Neeman (1999), Krishna and Perry (2000), Schweizer (2006), and Segal and Whinston (2011).

with appropriately chosen endowments, ex post efficient trade *is possible* without running a deficit.[18] This is achieved even though trade involves transactions costs in the form of information rents that must be covered. The fact that whether ex post efficient trade is possible or not hinges on the initial endowments highlights the importance of having endowments that are well chosen. In particular, Theorem 1 shows that the optimal endowments $\mathbf{r}^*$ generate positive ex ante revenue and so permit ex post efficient trade.

As mentioned, our paper is related to Loertscher and Marx (2020) and Delacrétaz et al. (2022). In the context of Theorem 1, the identification by Loertscher and Marx (2020) of asset market environments in which the VCG mechanism has a balanced budget subject to ex post individual rationality, together with the result in Delacrétaz et al. (2022) that subject to ex post individual rationality the VCG mechanism never runs a budget surplus, highlights the role of the weaker notion of interim individual rationality that we impose.[19]

[18] To see why the designer may run a deficit in the consignment auction, consider a parameterization such that for some $\mathcal{S} \subset \mathcal{N}$, we have $\sum_{i \in \mathcal{S}} k_i = 1$, implying that total demand by agents in $\mathcal{S}$ is equal to the total supply. If $r_i = k_i$ for all $i \in \mathcal{S}$, then we have a two-sided allocation problem, with the agents in $\mathcal{S}$ acting as sellers and the remaining agents acting as buyers. The results in Loertscher and Mezzetti (2019) then imply that under the VCG mechanism, the deficit on each unit traded is at least as large as the difference between the highest and the lowest market-clearing Walrasian price. Hence, $\Pi(\boldsymbol{\theta}, \mathbf{r}) \leq 0$ for all $\boldsymbol{\theta}$ and $\Pi(\boldsymbol{\theta}, \mathbf{r}) < 0$ for a positive measure set of $\boldsymbol{\theta}$, implying an expected deficit from the second-stage auction. Moreover, under the VCG mechanism, the interim individual-rationality constraints are binding for the highest possible types of the sellers (agents in $\mathcal{S}$) and for the lowest possible types of the buyers (agents in $\mathcal{N} \setminus \mathcal{S}$). Thus, there are no fixed payments that increase revenue without violating the interim individual-rationality constraints.

[19] The gist of the environments identified by Loertscher and Marx (2020) is as follows. Suppose there is an odd number of traders, each of whom has an endowment of one and a maximum demand to use the good of two units. For any given realization of agents'

A natural question that may arise is why one would or should care about the no-deficit constraint in the first place. Indeed, given the environment with quasilinear utility, one could simply use transfers, for example, from the general public, to balance the budget. Even if one accepts the no-deficit constraint, one may wonder why or to what extent $\mathbf{r}^*$ would be preferable to any $\mathbf{r}$ satisfying $\Pi_{\mathbf{r}} \geq 0$.[20] To address these questions, it is useful to recall that the assumption of quasilinear utility is imposed not because it is believed to hold universally, but because it makes the model tractable. Without quaslinear utility, transfers can be costly, and there are numerous reasons for why taxes that fill the public coffers are socially costly, making them unsuitable to finance particular markets or projects.[21] Put differently, the seminal impossibility results of Vickrey (1961) and Myerson and Satterthwaite (1983) and the entire literature on incentive compatible provision of public goods implicitly or explicitly rest on arguments along these lines. As Vickrey (1961, p. 8) notes, using public funds for these purposes would "prove to be inordinately expensive in terms of their demands on the fiscal resources of the state relative to the net benefits to be realized." While this is not explicitly modeled, it seems plausible and intuitive that the larger is the budget surplus

types, the median type is the Walrasian price, which is a singleton, and the agent with that type consumes its endowment under efficiency. Under these conditions, the VCG prices for all units that are traded are equal to the Walrasian price, and so the budget of the VCG mechanism is balanced.

[20] In the Online Appendix, we provide an AGV mechanism (d'Aspremont and Gérard-Varet, 1979; Arrow, 1979) with zero revenue ex post—as this shows, our requirement of nonnegative revenue in expectation can be strengthened to require nonnegative revenue ex post because, as shown by Börgers and Norman (2009), given any Bayesian incentive compatible mechanism that generates nonnegative revenue in expectation, one can construct another Bayesian incentive compatible mechanism with the same allocation rule and the same interim expected payments that generates nonnegative revenue for all type realizations.

[21] See also Hellwig (2003) for a related discussion.

under ex post efficiency, subject to incentive compatibility and individual rationality, the larger is the leeway to obtain efficient outcomes.[22]

Furthermore, in settings like those of an emission permits market, there are multiple practical reasons why a designer might care about the revenue generated from a consignment auction. For example, running the auction is not costless and revenue generated from the auction can be used to cover administrative costs of the consignment auction or for other environmental causes that generate social benefits (see, e.g., ICAP, 2021). In addition, the presence of an expected budget surplus may increase the motivation for a government or third party to initiate an emission market.

3.2 Benchmarking and grandfathering

As mentioned in the introduction, benchmarking and grandfathering are widely used means to determine endowments $\mathbf{r}$. We now show that in a setting with what we call strength-ordered groups, endowment vector $\mathbf{r}^*$ amounts to benchmarking in the sense that agents within each group obtain the same endowments. Then we provide conditions under which grandfathering—that is, determining endowments on the basis of historical consumption levels—permits efficient consignment auctions. Because historical averages are simple statis-

[22] To illustrate this point, consider a simple bilateral trade setup in the tradition of Myerson and Satterthwaite (1983). That is, there is one buyer and one seller and a single good to be traded. Letting $[\underline{v}, \overline{v}]$ and $[\underline{c}, \overline{c}]$ be the supports of the buyer's and the seller's type distributions satisfying $\overline{v} > \underline{c}$, ex post efficiency is possible if and only if $\underline{v} \geq \overline{c}$. If ex post efficiency is possible, it can be achieved by setting a posted price $p \in [\overline{c}, \underline{v}]$. Thus, the larger is the gap $\underline{v} - \overline{c}$, the more leeway there is to choose a p that achieves ex post efficiency. Because $p_B = \underline{v}$ is the largest price that the buyer can be charged and $p_S = \overline{c}$ is the smallest that the seller can be offered without violating ex post efficiency, this is equivalent to saying that the larger is the designer's revenue $p_B - p_S$ under ex post efficiency, the more leeway there is to achieve ex post efficiency.

tics, the data requirements for grandfathering are small.

Benchmarking and strength-ordered groups

We begin by defining strength-ordered agents and then extend that to strength-ordered groups. We say that we have *strength-ordered agents* if lower-indexed agents are stronger in the sense of having first-order stochastically dominating distributions and larger maximum demands, i.e., for all $i \in \{1, \dots, n-1\}$ and $\theta \in (\underline{\theta}, \overline{\theta})$,

$$F_i(\theta) < F_{i+1}(\theta) \quad \text{and} \quad k_i > k_{i+1}. \tag{3}$$

We then have the following result:

Proposition 2. *With strength-ordered agents, stronger agents have greater optimal endowments; that is, $i < j$ implies $r_i^* > r_j^*$.*

Proof. See Appendix A.4.

Proposition 2 provides support for a benchmarking approach that gives larger endowments to firms in industries that have greater value for the permits, i.e., relatively more permits are given to firms in industries with relatively more emissions. For additional results on how optimal endowments change with maximum demands, see the Online Appendix. There we show that with two strength-ordered agents, the optimal endowment of the stronger agent increases with proportional increases of the maximum demands and with the maximum demand of the stronger agent, and decreases with the maximal demand of the weaker agent. This provides support for a benchmarking approach that shifts permits to heavier emitters following a demand shift that affects all industries proportionally or following demand shifts that favor the heavy emitting industries or disfavor other industries.

Under benchmarking, a facility's endowment is based on a fixed benchmark for an appropriate level of the emissions per unit of output, multiplied by the facility's expected output. For example, the benchmark might be set somewhere between the industry average and the best performing facility (Pizer and Zhang, 2018). In such a regime, benchmarks can be expected to differ across industries, with, for example, different benchmarks for power, cement, and aluminum.

To allow for such differences, we now amend the above setup by allowing agents to belong to different groups ex ante, where agents within a group have the same type distribution and same maximum demand to use the good. For the setup with groups, we show an "equal treatment of the equals" result: at $\mathbf{r}^*$, agents in the same group have the same optimal endowment. When agents can be grouped into a strong group and a weak group, which we define below, the optimal endowment weakly increases for a weak agent that transitions from the weak to the strong group.

Denoting the set of groups by $\mathcal{G} \equiv \{G_1, \ldots, G_m\}$,[23] we obtain the result of equal treatment of the equals stated in Corollary 1 by noting that the interim efficient allocation q_i is the same for all agents in the same group, so $r_i^* = q_i(\hat{\theta})$ is the same for agents in the same group:

Corollary 1. *Ex ante identical agents have the same optimal endowment; that is, for all $G_\ell \in \mathcal{G}$ and all $i, j \in G_\ell$, $r_i^* = r_j^*$.*

Corollary 1 provides support for the use of a common benchmark across all firms in the same group.

We say that a setup has *strength-ordered groups* if for all $\theta \in (\underline{\theta}, \overline{\theta})$,

$$F_{G_1}(\theta) < \cdots < F_{G_m}(\theta) \quad \text{and} \quad k_{G_1} > \cdots > k_{G_m},$$

where F_{G_i} and k_{G_i} are the distribution and maximum demand, respectively, for every agent in G_i. A lower-numbered group is *stronger* in that it has a better distribution in the sense of first-order stochastic dominance and a larger maximum demand to use the good. Proposition 2 and Corollary 1 then imply that agents in stronger groups have greater optimal

[23] Formally, we assume $\cap_{G \in \mathcal{G}} G = \emptyset$ and $\cup_{G \in \mathcal{G}} G = \mathcal{N}$. The model studied thus far is a special case of this with $|\mathcal{G}| = n$ and, for each group G, $|G| = 1$.

endowments:

Proposition 3. *In the setup with strength-ordered groups, all agents in a stronger group have a greater optimal endowment than agents in a weaker group; that is, $\ell_1 < \ell_2$ implies that for all $i \in G_{\ell_1}$ and $j \in G_{\ell_2}$, we have $r_i^* > r_j^*$.*

Proposition 3 implies that endowments will be greater for industries that have better distributions in the sense of first-order stochastic dominance and/or larger maximum demands.[24] We provide additional results for strength-ordered groups in the Online Appendix.

Grandfathering and efficient consignment auctions

A widely used scheme to determine endowments is so-called *grandfathering*, whereby firms are endowed with permits according to their long-run average consumption. Grandfathering is often criticized on various grounds, including incentives to inefficiently boost consumption to increase future endowments and concerns for equity because larger polluters receive larger endowments. Without disputing these possible concerns, we now show that endowments equal to each firm's expected consumption under efficiency allow the designer to break even in expectation and so allows an efficient second-stage auction to operate. This means that "grandfathering gets it right" if the the historical consumption reflects efficient allocations. Of course, as discussed at the end of Section 2, efficiency is a strong assumption, but without direct evidence as to the nature of inefficiency, any departure from that assumption seems even stronger.

Formally, letting $\overline{r}_i = \mathbb{E}_{\boldsymbol{\theta}}[Q_i(\boldsymbol{\theta})]$ and $\overline{\mathbf{r}} = (\overline{r}_i)_{i \in \mathcal{N}}$ be the endowments under grandfathering and $T_{\overline{\mathbf{r}},i}(\boldsymbol{\theta})$ be the associated VCG transfers for all $i \in \mathcal{N}$, we have the following result:

Proposition 4. *Consider grandfathering. Then, each agent's interim individual rationality constraint is satisfied. Moreover, the ex ante expected transfer of every agent is non-negative,*

[24] Proposition 3 generalizes a result obtained by Che (2006) for the case with single-agent groups, $k_i = 1$ for all $i \in \mathcal{N}$, and distributions ranked by first-order stochastic dominance.

that is, for all $i \in \mathcal{N}$, $\mathbb{E}_{\boldsymbol{\theta}}[T_{\bar{\mathbf{r}},i}(\boldsymbol{\theta})] \geq 0$. Consequently, the designer's expected revenue is nonnegative, that is, $\Pi_{\bar{\mathbf{r}}} \geq 0$.

Proof. See Appendix A.5.

To develop intuition for Proposition 4, observe that if it so happened that $r_i = Q_i(\boldsymbol{\theta})$ for each i, there would be no trade in the second stage of the consignment auction because every agent's endowment equals its efficient consumption level. Consequently, the market would not run a deficit. Proposition 4 shows that this no-deficit property extends to the case where each agent i is endowed with its expected consumption under ex post efficiency, that is, $\bar{r}_i = \mathbb{E}_{\boldsymbol{\theta}}[Q_i(\boldsymbol{\theta})]$. Moreover, this intuition also makes clear that, perhaps counterintuitively, low trading volumes in the second-stage do not imply that emission markets are inefficient.

Segal and Whinston (2011) refer to the vector $\bar{\mathbf{r}}$ as a simple status quo. They show that it ensures voluntary participation and efficient bargaining with private as well as interdependent values (see their Corollaries 1 and 2). To keep our paper self-contained, and because the proof is short, we provide an independent proof of Proposition 4 in Appendix A.5. For the rest of the paper, we refer the endowments $\bar{\mathbf{r}}$ as the Segal-Whinston (S-W) endowments.

While we consider grandfathering in the context of a consignment auction, Cramton and Kerr (2002) contrast grandfathering, where the authority gives permits away to specific groups, with a direct auction of permits by the authority, arguing that auctioning is preferable because it "provides greater incentives for innovation, provides more flexibility in distribution of costs, and reduces the need for politically contentious arguments over the allocation of rents" (Cramton and Kerr, 2002, p. 339). Proposition 4 shows that grandfathering in a consignment auction based on the expected consumption of permits has the advantage of inducing endowments that permit efficient reallocation in a second-stage auction without running a deficit in expectation. As Proposition 5 below shows, firms prefer efficient consignment auctions to permits being sold via an efficient auction, which makes consignment auctions politically more appealing. Further, as Proposition 7 in Section 5 shows, efficient consignment auctions induce efficient investment.[25]

In general $\mathbf{r}^*$ and $\overline{\mathbf{r}}$ differ, even with only two agents.[26] However, the case with identical distributions and two agents implies that $\mathbf{r}^* = \overline{\mathbf{r}}$. To see this, notice first that with identical distributions and two agents, the worst-off type $\hat{\theta}$ is such that $F(\hat{\theta}) = 1/2$.[27] This implies that $r_1^* = (k_1 + 1 - k_2)/2$. Next, observe that agent 1's ex ante expected allocation is

$$\overline{r}_1 = \mathbb{E}_{\boldsymbol{\theta}}[Q_1(\boldsymbol{\theta})] = \int_0^1 [k_1 F(\theta) + (1-k_2)(1-F(\theta))] f(\theta) d\theta = (k_1+k_2-1)\int_0^1 F(\theta) f(\theta) d\theta + 1 - k_2.$$

But this is the same as r_1^* if $\int_0^1 F(\theta) dF(\theta) = 1/2$, which is indeed the case,[28] establishing that $\mathbf{r}^* = \overline{\mathbf{r}}$. In addition, if all firms are ex ante identical, that is, if $k_i = k$ and $F_i = F$ for all $i \in \mathcal{N}$, then each firm's ex ante expected allocation is $1/n$, which is also r_i^*. Hence, $\mathbf{r}^* = \overline{\mathbf{r}}$. Of course, for cases in which $\mathbf{r}^* = \overline{\mathbf{r}}$, we have, using Theorem 1, for all $\boldsymbol{\theta}$, $\Pi(\boldsymbol{\theta}, \mathbf{r}^*) = \Pi(\boldsymbol{\theta}, \overline{\mathbf{r}}) \geq 0$ and $\Pi_{\mathbf{r}^*} = \Pi_{\overline{\mathbf{r}}} > 0$, which gives us the following result:

Corollary 2. *Assume that for all $i \in \mathcal{N}$, $F_i = F$ and either (i) $n = 2$ or (ii) for all $i \in \mathcal{N}$, $k_i = k$. Then under grandfathering, the designer never runs a deficit ex post and obtains a positive budget surplus in expectation.*

[25] That said, efficient investments would also be a Nash equilibrium outcome if an efficient standard auction were used to allocate permits.

[26] For example, with two agents and distributions $F_1(\theta) = \theta^2$ and $F_2(\theta) = \theta$, we have $\hat{\theta} = (\sqrt{5} - 1)/2$ and $r_1^* = (k_1(\sqrt{5} - 1) + (1 - k_2)(3 - \sqrt{5}))/2$ and $\overline{r}_1 = (1 + 2k_1 - k_2)/3$.

[27] To see this, observe first that $q_i(\theta) = k_i F(\theta) + \max\{1 - k_j, 0\}(1 - F(\theta))$. Next, distinguishing between the three cases $k_j \leq k_i < 1$; $k_j < k_i = 1$; and $k_j = k_i = 1$, one obtains $\sum_h q_h(\theta) = 2(k_1 + k_2 - 1)F(\theta) + 2 - (k_1 + k_2)$; $\sum_h q_h(\theta) = 2k_j F(\theta) + 1 - k_j$; and $\sum_h q_h(\theta) = 2F(\theta)$. Simple algebra shows that $\hat{\theta}$ such that $\sum_h q_h(\theta) = 1$ implies $F(\hat{\theta}) = 1/2$ for each of these three cases.

[28] This follows because quantiles are uniformly distributed, so the expected quantile is 1/2. Formally, using the change of variables $y = F(\theta)$, we have $\int_0^1 F(\theta) dF(\theta) = \int_0^1 y dy = 1/2$.

3.3 Hybrid consignment auctions

Emission abatement imposes costs on producers, which may pose political challenges for the imposition of regimes that restrict emissions. For a given total emission level, allocating the corresponding number of permits to emitting entities via an efficient auction may not be politically palatable because these entities will aim to resist this additional "tax."[29] The designer can reduce these political obstacles by running what may be called a hybrid consignment auction. To formalize this notion, we focus on the "variable" part of demand parameterized by k_i for each i and the supply $R = 1$, which is net of the minimum demands $\sum_{i \in \mathcal{N}} d_i$.

To capture the possibility mentioned in the Introduction that the authority may withhold permits in the first stage and sell them in the second stage, we now define an *α-hybrid consignment auction* as a mechanism in which the agents are endowed with a fraction α of the total supply R in the first stage and the remaining fraction $1 - \alpha$ is auctioned off in the second-stage auction. Let $\hat{\theta}(\alpha)$ be such that $\sum_i q_i(\hat{\theta}(\alpha)) = \alpha$. For all $i \in \mathcal{N}$, let $r_i^*(\alpha) = q_i(\hat{\theta}(\alpha))$ and $\Pi_H(\boldsymbol{\theta}, \mathbf{r}^*(\alpha))$ be the associated ex post revenue in the second-stage auction from selling the α proportion of the permits and from the agents' trading their endowments and denote by $\Pi_H(\alpha) = \mathbb{E}[\Pi_H(\boldsymbol{\theta}, \mathbf{r}^*(\alpha))]$ the ex ante expected revenue. Then the α-hybrid consignment auction generates nonnegative expected revenue. Moreover, $\Pi_H(\alpha)$ weakly decreases in α while each agent's payoff weakly increases in α.

Proposition 5. *In an α-hybrid consignment auction, maximized expected revenue $\Pi_H(\alpha)$ is nonnegative and weakly decreases in α, and each agent's expected payoff weakly increases in α.*

Proof. See Appendix A.6.

Proposition 5 implies that the designer obtains nonnegative expected revenue from a

[29] See e.g., Kreibich and Hermwille (2021) for the problems faced by the firms in coping with climate policies.

hybrid consignment auction with endowments $\mathbf{r}^*(\alpha)$ *and* that the agents themselves prefer that to having the designer auction off all the permits in the second stage. Moreover, the designer can choose the trade-off between revenue and political palatablity by varying α. The Online Appendix provides a more detailed analysis of the properties of $\mathbf{r}^*(\alpha)$ for cases with two agents.

3.4 Overendowment of agents

We now drop the restriction that $r_i \leq k_i$ for all i. By the assumption that the agents have zero value for any units that are beyond their maximum demand, endowments greater than k_i function similarly to the $(1-\alpha)$ proportion of supply that is auctioned off by the designer in the α-hybrid auction. In this way, overendowments can improve revenue as we now show.

Let agent i^* be the agent with the minimal type-integrated distance between its maximum demand and its interim expected allocation under ex post efficiency, that is,

$$i^* \in \arg\min_{i \in \mathcal{N}} \int_{\underline{\theta}}^{\overline{\theta}} (k_i - q_i(x))dx,$$

which, for ease of exposition, we assume to be unique. Proposition 6 shows that agent i^* receives the full supply if overendowments are optimal.

Proposition 6. *If overendowments* $\mathbf{r}^o$ *maximize the designer's expected revenue, then* $\mathbf{r}^o$ *has the form that, for all* $j \in \mathcal{N} \backslash \{i^*\}$,

$$r_j^o \in [0, \max\{0, 1 - \sum_{\ell \in \mathcal{N} \backslash \{j\}} k_\ell\}], \quad \text{and} \quad r_{i^*}^o = 1 - \sum_{j \in \mathcal{N} \backslash \{i^*\}} r_j^o.$$

Proof. See the Online Appendix.

4 Application

Our theoretical analysis assumes that the revenue of the VCG mechanism in the second-stage of the consignment auction is non-negative. It also provides three different endowment vectors: $\mathbf{r}^*$ and $\bar{\mathbf{r}}$, which generate non-negative revenue in the second-stage, and $\mathbf{r^o}$, which may maximize second-stage revenue when overendowments are allowed. In this section, we test whether, given empirically observed endowments and estimated distributions, the VCG revenue is nonnegative and which of the theoretical endowments best matches the empirical ones.

Our application uses data from Southern California's RECLAIM pollution permit trading program collected and analyzed by Fowlie and Perloff (2013). The RECLAIM setting involves an initial assignment of permits followed by unstructured trading. Thus, the second stage is not literally an auction as modeled above. Nevertheless, it is useful to take an as-if approach to modeling this market and to assume that the reallocation stage is efficient, and hence equivalent to a VCG mechanism being run.[30] In line with this, we change the terminology for the second stage of the consignment auction from second-stage "auction" to second-stage "reallocation."

4.1 Data and procedure

The data that we use for our analysis consist of the initial assignment of emission permits and the actual emissions for ten periods for the 56 facilities in the data of Fowlie and Perloff (2013) that are designated as "cycle 1" facilities and that have observations for each of the final 10 semi-annual periods in the data, spanning 2001-02 to 2006-01.[31] Denoting the set

[30] See Larsen and Zhang (2021) for an empirical application of this as-if approach and Loertscher and Marx (2022) for additional motivation and a theoretical generalization.

[31] Facilities are divided into two staggered compliance cycles designated as "cycle 1" and "cycle 2."

of periods by T, for each period $t \in T$ and each facility $i \in \{1, \ldots, 56\}$, the data contain the initial assignment of emission permits to facility i, $\widehat{alloc}_{i,t}$, and facility i's emissions for period t, $\widehat{emission}_{i,t}$.

It is our understanding that permit holdings following trading must be equal to emissions to avoid penalty, so we use the facilities' emissions as a proxy for their permits following trading.[32] However, in the data, total assigned permits in each assignment period are not equal to total emissions in that assignment period as a result of excluded facilities and the trading of permits across periods.[33] Thus, to balance the total emissions and total allocations within each period, we create an artificial facility, facility 0, whose initial assignment and emissions balance the totals within each period.[34] Consistent with our model, the data reflect that whether a facility is a buyer or a seller in a particular consignment auction is endogenous. On average, in a given year, 35% of facilities are net buyers and 65% are net sellers. An individual facility may be a buyer in some years and a supplier in others. Only 8.7% of facilities are net buyers in every year, and 32% of facilities are net sellers in every year. The remainder—59% of facilities—are net buyers in some years and net sellers in other years. Figure 1 illustrates the initial assignment and emissions of one facility, Mission Clay Product. As shown in the figure, in some years, the facility has higher emissions than its endowment and thus is a net buyer, but in other years, it has a higher endowment than its emissions and is thus a net seller.

To match the theoretical model, we assume that in each period t, each facility i draws

[32] We presume that an agent's emissions cannot be more than its permit holdings. Further, if emissions were lower, then one would expect the agent to sell its excess permits. This supports the assumption of permit holdings after trading being equal to emissions.

[33] In addition, Fowlie and Perloff (2013) mention that the government reserved some permits, which it then sold in the secondary market.

[34] For the artificial facility, we have $\widehat{alloc}_{0,t} = \max\{0, \sum_{i=1}^{56} \widehat{emission}_{i,t} - \sum_{i=1}^{56} \widehat{alloc}_{i,t}\}$ and $\widehat{emission}_{0,t} = \max\{0, \sum_{i=1}^{56} \widehat{alloc}_{i,t} - \sum_{i=1}^{56} \widehat{emission}_{i,t}\}$.

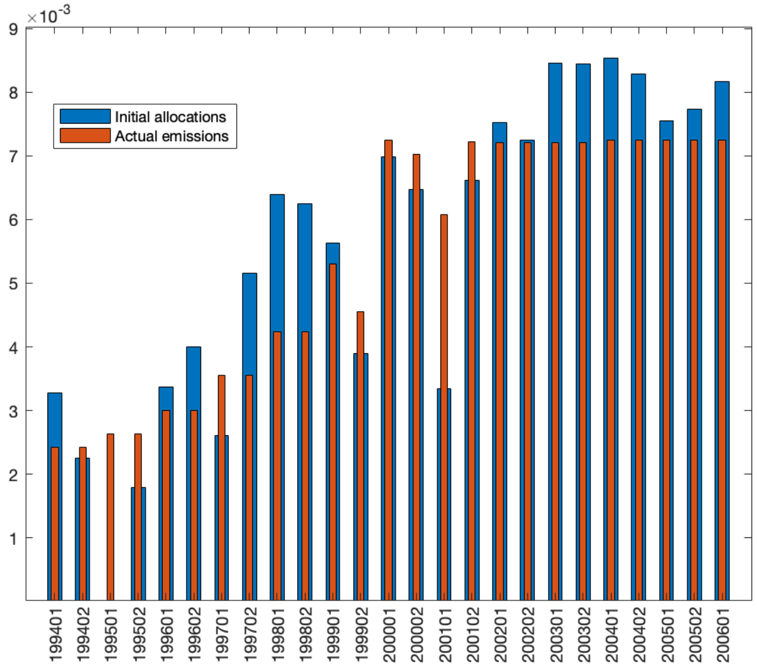

Figure 1: Data for Mission Clay Product on initial assignments and actual emissions by year. Source: Fowlie and Perloff (2013).

a type $\theta_{i,t}$ independently from a known distribution F_i that captures facility i's value of a unit *share* of the permits, rather than the absolute quantity of NOx emissions. We further assume that the type draws $(\theta_{i,t})_{t \in T}$ are independent across periods.

For each period t, let $supply_t \equiv \sum_{i=0}^{56} \widehat{alloc}_{i,t} = \sum_{i=0}^{56} \widehat{emission}_{i,t}$ be the total supply in that period. As the RECLAIM program is designed to reduce NOx emissions overtime. $supply_t$ is reduced overtime. We first normalize the supply for all periods to $\hat{R}$. Hence, for each facility i and each period t, we have its emission shares, denoted by $emission_{i,t}$, and its allocation shares, denoted by $alloc_{i,t}$, where

$$emission_{i,t} \equiv \frac{\widehat{emission}_{i,t}}{supply_t} \hat{R} \quad \text{and} \quad alloc_{i,t} \equiv \frac{\widehat{alloc}_{i,t}}{supply_t} \hat{R}.$$

We further assume that the facilities can be divided into groups where the facilities in the same group draw their types independently from the same distribution and have the same maximum and minimum demand. We divide the data into 12 groups based on each facility i's average emission overtime, $Average\ Emission_i \equiv \frac{\sum_{t\in T} emission_{i,t}}{T}$, and we denote the set of groups by $\mathcal{G} = \{1, \ldots, 12\}$. If facilities i and j are in the same group $g \in \mathcal{G}$, we have $F_i = F_j = F_g$, $k_i = k_j = k_g$, and $d_i = d_j = d_g$, but facilities i and j are still distinct facilities that participate in the reallocation stage independently. We parameterize the distributions by assuming that $F_g(\theta) = \theta^{a_g}$.[35] We denote the vector of distributional parameters by $\mathbf{a} = (\mathbf{a_g})_{\mathbf{g}\in\mathcal{G}}$ and by $\mathbf{F}(\mathbf{a})$ the vector of distributions parameterized by $\mathbf{a}$.

For each group $g \in \mathcal{G}$, we approximate d_g and k_g using its emission data, $(emission_{i,t})_{i\in g,t\in T}$. More specifically, we use $\tilde{d}_g = \min_{i\in g,t\in T}$ as the empirical minimum demand for group g and $\tilde{k}_g + \tilde{d}_g = (\tilde{F}_g((emission_{i,t})_{i\in g,t\in T}))^{-1}(0.9)$, that is the 90th percentile, as the empirical maximum demand for group g, where $\tilde{F}_g((emission_{i,t})_{i\in g,t\in T})$ is the empirical distribution extrapolated from the sample data $(emission_{i,t})_{i\in g,t\in T}$. Given $\tilde{\mathbf{d}}$ and $\hat{R}$, the empirical emission of each facility $i \in g$ in period t is $\tilde{Q}_{i,t} \equiv emission_{i,t} - d_g$. We use $\tilde{\mathbf{Q}}$ to denote the empirical emission.

In line with our theoretical analysis, we assume that the reallocation stage operates efficiently. For i in group g, let $\mathbb{E}_{\mathbf{F}(\mathbf{a})}[Q_i(\boldsymbol{\theta})]$ be the ex ante expected allocation and let $\tilde{Q}_g = \frac{\sum_{i\in g,t\in T} \tilde{Q}_{i,t}}{|g||T|}$ be the empirical average emission for group g facility. The estimated distributional parameters $\tilde{\mathbf{a}}$ then minimize the sum of squared errors between the empirical average emission and the theoretical ex ante expected allocation, $SSE(\mathbf{a})$, where

$$SSE(\mathbf{a}) \equiv \sum_{g\in\mathcal{G}} \left(\frac{\sum_{i\in g} \mathbb{E}_{\mathbf{F}(\mathbf{a})}[Q_i(\boldsymbol{\theta})]}{|g|} - \tilde{Q}_g \right)^2 \tag{4}$$

[35] We perform the same calibration exercise for two alternative families of distributions: cumulative distribution functions $F_g(\theta) = 1 - (1-\theta)^{a_g}$, and the Beta distributions with two equal parameters, $Beta(a_g, a_g)$, that is, $f_g(\theta) = \frac{\Gamma(2a_g)}{\Gamma^2(a_g)}\theta^{a_g-1}(1-\theta)^{a_g-1}$, where $\Gamma(\cdot)$ is the Gamma function. We show that the results are robust to the distribution assumptions (see Table 2.2 and the measures of distributional similarity that follow that table).

Table 2.1 in Appendix 2.1 provides summary statistics for each facility's allocations and emissions across periods. Table 2.2 in Appendix 2.1 summarizes each facility's maximum demand, fitted distributional parameter $\tilde{\mathbf{a}}$, and optimal normalized endowment $\mathbf{r}^*$.

4.2 Testing for nonnegative revenue

Observe that $\tilde{\mathbf{a}}$ does not make use of information about the empirical endowments nor does our calibration use the assumption that the VCG mechanism does not run a deficit in the second-stage. Given $\tilde{\mathbf{a}}$ and the empirical endowments, it is therefore possible that the VCG mechanism would run a deficit. Thus, computing revenue of the VCG mechanism given $\tilde{\mathbf{a}}$ and the observed endowments provides a test for our identifying assumption: if the computed revenue is nonnegative, the identifying assumption is validated, and otherwise it is rejected. Thus, our identifying assumption is validated by the positive expected revenues shown in Table 1, which reports the expected VCG revenues from the empirical endowments for the sample periods using the distributions $\mathbf{F}(\tilde{\mathbf{a}})$.

Periods	2001-02	2002-01	2002-02	2003-01	2003-02
Revenues	0.0451	0.0451	0.0295	0.0451	0.0400
Periods	2004-01	2004-02	2005-01	2005-02	2006-01
Revenues	0.0446	0.0415	0.0448	0.0446	0.0446

Table 1: Expected VCG revenues from empirical endowments

Given that the empirical endowments permit efficient reallocations, if there were significant transaction costs in the reallocation stage, one would expect to observe a lower trading volume than under efficiency. We explore this by comparing the empirical trading volumes with trading volumes simulated under the assumption of ex post efficient reallocation, given the empirical endowments.[36] The simulations are based on 1000 independent draws of the

[36] We define the trading volume in a period to be one-half of the sum of the absolute value of the differences between each facility's endowment and emission.

type profile $\boldsymbol{\theta}$ for each period from the fitted distributions $\mathbf{F}(\tilde{\mathbf{a}})$. Figure 2 shows the results, focusing on the mean, 5-th, and 95-th percentiles. The observed trading volumes are, with one exception, within 90% of the simulated trading volumes, thus providing additional evidence of consistency of the theoretical assumptions with the RECLAIM data.

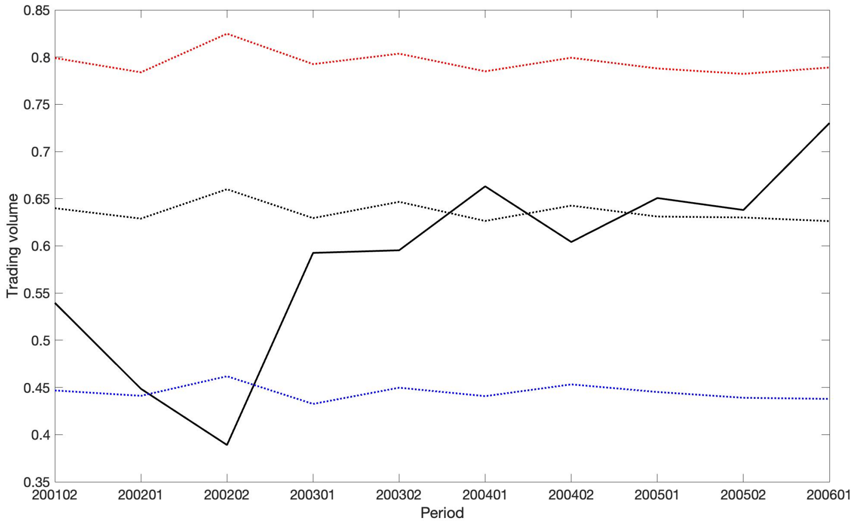

Figure 2: Observed trading volumes (black solid line) and simulated trading volumes under ex post efficient trade relative to the empirical endowments (dotted lines), at the 95th percentile (red dotted lines), mean (black dotted lines), and the 5th percentile (blue dotted lines).

4.3 Theoretical and empirical endowments

The theoretical analysis derives three endowments, $\mathbf{r}^*$, $\bar{\mathbf{r}}$, and $\mathbf{r}^o$. We now use the data to see which of these is the best match to the empirically observed endowments. We use two measures: the comparison with the VCG revenue that these endowments imply versus that implied by the empirical endowments, and the Kullback–Leibler distance measure that is commonly used to formalize how good one distribution matches another one.

We start with the comparison of VCG revenues. Figure 3 displays the revenue implied by the empirical endowments by diamond-shaped markers, and the dotted black, blue and red lines correspond to the revenues implied by $\mathbf{r}^*$, $\bar{\mathbf{r}}$, and $\mathbf{r}^o$, respectively.

Figure 3 shows that each of the three theoretical endowments generates more revenue than

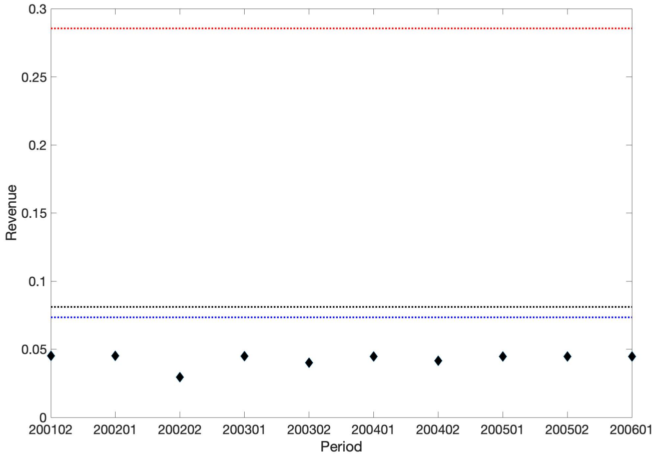

Figure 3: Comparing the revenues implied by the empirical endowments (diamond shaped markers) with the revenues implied by $\mathbf{r}^*$ (dotted black line), $\bar{\mathbf{r}}$ (dotted blue line), and $\mathbf{r}^o$ (dotted red line).

the empirical endowment and that, measured by the implied revenue, $\bar{\mathbf{r}}$ is the best match to the empirical endowment, suggesting that grandfathering may be a good approximation to the process that generates these data. Further, because as mentioned grandfathering results in relatively low trading volumes, this observation contrasts with the common interpretation that high (low) empirical trading volumes indicate successful (unsuccessful) emission markets (see, e.g., Zhang et al., 2020). Figure 3 also shows that $\mathbf{r}^{\mathbf{o}}$ is associated with the largest revenue. This happens because there are many firms with small maximum demands, which means that overendowing those firms results in a reallocation mechanism with little private information by the owners. Thus, the reallocation stage resembles a one-sided auction.

Next, we directly compare the average empirical endowments with the theoretical ones. Because the overendowment $\mathbf{r}^{\mathbf{o}}$ is not a good match empirically, we confine attention to $\mathbf{r}^*$ and $\bar{\mathbf{r}}$. Figure 4 displays the average empirical endowments by the wide white bars, $\mathbf{r}^*$ by thin black bars, and $\bar{\mathbf{r}}$ by blue bars. As the figure shows, the distributions of the endowments of all three types across facilities share a similar shape.

To more formally compare theoretical and empirical endowments, we now treat each facility as a state in a probability space and its endowment as the probability measure assigned to that state. This then allows us to compare the resulting distributions of endowments using Kullback–Leibler (KL) distance, which is commonly used as a measure of dissimilarity between two probability distributions. The KL distance is equal to zero for two identical probability distributions, and its absolute value increases as the dissimilarity increases. Using the average empirical endowment as the reference distribution, the KL distance between the average empirical endowment with $\mathbf{r}^*$ is 0.2193, while the KL distance between the average empirical endowment with $\bar{\mathbf{r}}$ is 0.2074. As a point of reference, the KL distance between the empirical endowment and a completely uniform endowment—in which each facility receives a share of 1/57—is 1.3705. This shows that the average empirical endowment is close to $\mathbf{r}^*$ and $\bar{\mathbf{r}}$, consistent with our observation from Figure 4 and the notion that grandfathering is a good approximation to the data generating process.

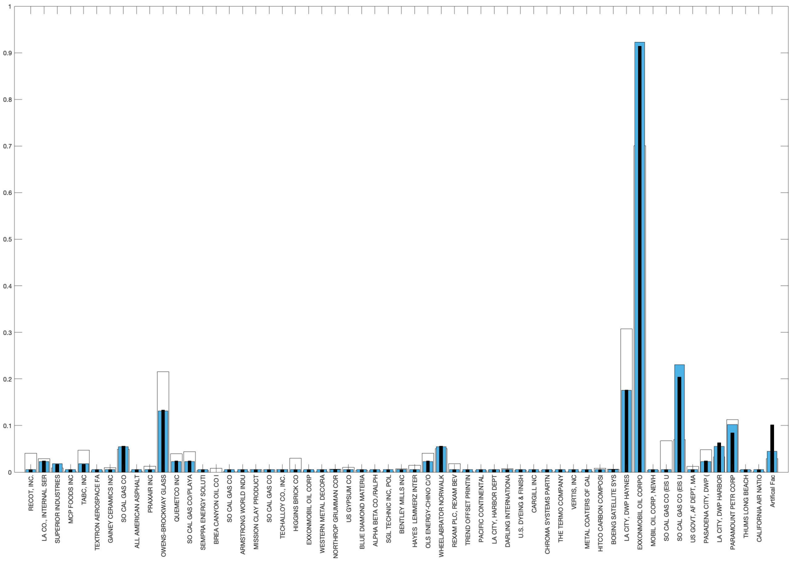

Figure 4: Comparing the empirical endowments (average across time, wide white bars) with $\mathbf{r}^*$ (thin black bars) and $\bar{\mathbf{r}}$ (blue bars).

5 Extension to allow investment

An important policy question is whether emission permits and associated emission markets diminish incentives to invest in emission abatement. With this in mind, we now amend the setup to allow investment and to consider the effects of endowments on investment incentives. We show here that an efficient consignment auction—or any efficient mechanism as foreshadowed by footnote 25—gives agents exactly the right incentives to invest; so any argument about suboptimal incentives to invest due to emission markets must be an argument that these markets are not efficient.

Without denying political economics arguments against the use of auctions, Cramton and Kerr (2002) raise concerns regarding incentives for innovation when permits are given away to specific groups because under a direct auction of permits by the authority, "innovators do not receive scarcity rents, so they unambiguously benefit from the innovation-induced fall in permit prices" because these rents are collected as revenue by the authority. In contrast, "in a grandfathered system the scarcity rents belong to the industry, so there is no aggregate gain to industry from reducing them" (Cramton and Kerr, 2002, p. 340). We now show that an efficient consignment auction provides incentives for efficient investment.

To incorporate investment, we assume that the agents have investment options to reduce their abatement costs. We model this by assuming that each agent's investment (or effort) affects its own distribution. We denote by $C_i(e)$ agent i's cost associated with investment $e \geq 0$ and let $F_i(\cdot; e)$ be its type distribution with an investment e, whose density we denote by $f_i(\cdot; e)$. We allow $f_i(\cdot; e)$ to vary with e but assume that the support $[\underline{\theta}, \overline{\theta}]$ is independent of investments. For each $i \in \mathcal{N}$, we assume that $f_i(\theta; e) > 0$ for any $e \geq 0$ and any $\theta \in (\underline{\theta}, \overline{\theta})$, which is analogous to the assumptions in the model without investment.

We assume that investments are not observable and that the market mechanism can condition on equilibrium investments. Because investments are not observable, the market mechanism does not vary with off-equilibrium investments.

We first consider the designer's problem of maximizing ex ante expected welfare under the assumption that upon investments $\mathbf{e} = (e_1, \dots, e_n)$, efficient reallocation is possible. Recalling that $W(\boldsymbol{\theta})$ denotes social welfare at type profile $\boldsymbol{\theta}$, the designer's optimal investments maximize

$$W^I(\mathbf{e}) = \int_{[\underline{\theta},\overline{\theta}]^n} W(\boldsymbol{\theta}) f(\boldsymbol{\theta};\mathbf{e}) d\boldsymbol{\theta} - \sum_{i \in \mathcal{N}} C_i(e_i),$$

where $f(\boldsymbol{\theta};\mathbf{e}) \equiv \times_{i\in\mathcal{N}} f_i(\theta_i;e_i)$ denotes the joint density. We denote by $\overline{\mathbf{e}}$ a solution to the designer's problem $\max_{\mathbf{e}} W^I(\mathbf{e})$. We impose the minimal assumption that $\overline{\mathbf{e}}$ exists. In particular, we require neither that $\overline{\mathbf{e}}$ is given by first-order conditions nor that it is unique.

Next, we consider the *investment game*, in which for a given endowment $\mathbf{r}$, the agents simultaneously choose investments $\mathbf{e}$. After investments are chosen, agents' types are realized, and they participate in an ex post efficient market. Given equilibrium investments $\mathbf{e}$, we denote by $q_i(\theta;\mathbf{e}_{-i})$ agent i's interim expected allocation under ex post efficiency, which is independent of i's investment. Adapting the insight underlying Lemma 1 that $q_i(\hat{\theta}_i) = \min\{r_i, k_i\}$ to the model augmented by investments, it follows that for a given r_i, agent i's worst-off type $\hat{\theta}_i(\mathbf{e}_{-i})$ is such that

$$q_i(\hat{\theta}_i(\mathbf{e}_{-i}), \mathbf{e}_{-i}) = \min\{r_i, k_i\}.$$

Note that $\hat{\theta}_i(\mathbf{e}_{-i})$ does not depend on e_i because $q_i(\theta;\mathbf{e}_{-i})$ is independent of e_i.

Relying on the payoff equivalence theorem, we can focus without loss of generality on the VCG mechanism. The VCG transfer for agent i at the type profile $\boldsymbol{\theta}$ when its worst-off type is $\hat{\theta}_i(\mathbf{e}_{-i})$ is $T_{\mathbf{r},i}(\boldsymbol{\theta}) = Q_i(\boldsymbol{\theta})\theta_i - W(\boldsymbol{\theta}) + W(\hat{\theta}_i(\mathbf{e}_{-i}), \boldsymbol{\theta}_{-i}) - \min\{r_i, k_i\}\hat{\theta}_i(\mathbf{e}_{-i})$ (see the proof of Theorem 1). Accordingly, agent i's payoff at type profile $\boldsymbol{\theta}$ is

$$W(\boldsymbol{\theta}) - \left(W(\hat{\theta}_i(\mathbf{e}_{-i}), \boldsymbol{\theta}_{-i}) - \min\{r_i, k_i\}\hat{\theta}_i(\mathbf{e}_{-i})\right).$$

Consequently, i's investment problem given $\mathbf{e}_{-i}$ is to maximize

$$\begin{aligned} U_i^I(\mathbf{e}) = &\int_{[\underline{\theta},\overline{\theta}]^n} W(\boldsymbol{\theta}) f(\boldsymbol{\theta};\mathbf{e}) d\boldsymbol{\theta} - C_i(e_i) \\ &- \int_{[\underline{\theta},\overline{\theta}]^{n-1}} \Big(W(\hat{\theta}_i(\mathbf{e}_{-i}), \boldsymbol{\theta}_{-i}) - \min\{r_i, k_i\} \hat{\theta}_i(\mathbf{e}_{-i}) \Big) f_{-i}(\boldsymbol{\theta}_{-i}; \mathbf{e}_{-i}) d\boldsymbol{\theta}_{-i}, \end{aligned}$$

where $f_{-i}(\boldsymbol{\theta}_{-i};\mathbf{e}_{-i}) \equiv \times_{j\neq i} f_j(\theta_j; e_j)$ is the joint density of $\boldsymbol{\theta}_{-i}$. Thus, $\mathbf{e}^*$ is a Nash equilibrium outcome of the investment game if and only if for all $i \in \mathcal{N}$, $e_i^* \in \arg\max_{e_i} U_i^I(e_i, \mathbf{e}_{-i}^*)$. The key observation is that

$$U_i^I(\mathbf{e}) = W^I(\mathbf{e}) + K(\mathbf{e}_{-i}),$$

where $K(\mathbf{e}_{-i})$ is constant with respect to e_i; that is, $K(\mathbf{e}_{-i})$ is independent of e_i. Notice that $\overline{\mathbf{e}}$ being a Nash equilibrium of the investment game means that $U_i^I(\overline{\mathbf{e}}) \geq U_i^I(e_i, \overline{\mathbf{e}}_{-i})$ for all $e_i \geq 0$. Using the definition of $U_i^I(\mathbf{e})$, this is equivalent to $W^I(\overline{\mathbf{e}}) \geq W^I(e_i, \overline{\mathbf{e}}_{-i})$ for all $e_i \geq 0$, which is the case by the definition of $\overline{\mathbf{e}}$.[37]

Let $\hat{\theta}(\overline{\mathbf{e}})$ be the unique number between $\underline{\theta}$ and $\overline{\theta}$ such that $\sum_{i\in\mathcal{N}} q_i(\hat{\theta}(\overline{\mathbf{e}}); \overline{\mathbf{e}}_{-i}) = 1$, and denote by

$$\mathbf{r}_{\overline{\mathbf{e}}}^* = (q_1(\hat{\theta}(\overline{\mathbf{e}}); \overline{\mathbf{e}}_{-1}), \ldots, q_n(\hat{\theta}(\overline{\mathbf{e}}); \overline{\mathbf{e}}_{-n}))$$

the associated optimal endowments. It follows that given endowments $\mathbf{r}_{\overline{\mathbf{e}}}^*$ and investments $\overline{\mathbf{e}}$, the market does not run a deficit ex post and a positive budget surplus in expectation, where the expectation is taken with respect to $f(\boldsymbol{\theta}; \overline{\mathbf{e}})$. Summarizing, we have shown:

Proposition 7. *The investment game has a Nash equilibrium in which each agent i chooses $e_i^* = \overline{e}_i$. Moreover, given endowments $\mathbf{r}_{\overline{\mathbf{e}}}^*$, in this equilibrium, for all $\boldsymbol{\theta}$, $\Pi(\boldsymbol{\theta}, \mathbf{r}_{\overline{\mathbf{e}}}^*) \geq 0$, and*

[37] The intuition is that Nash equilibrium imposes less stringent conditions than the designer's optimum. For example, if the designer's solution is given by first-order conditions, the designer needs to account for cross-partials whereas in a Nash equilibrium players do not account for the effects of their actions on other players' payoffs, implying that only second own partials need to be taken into account.

$\Pi_{\mathbf{r}^*_{\bar{\mathbf{e}}}} > 0$.

It follows from Proposition 7 and our assumptions about investments that any argument that markets distort investment incentives away from the first-best must be an argument that these markets are *not* efficient. In addition, intuition gleaned from complete-information models suggesting that there is a tradeoff between the efficiency of markets and investments is misleading: efficient markets with incomplete information imply efficient investments, and so that tradeoff does not arise.

6 Conclusion

Consignment auctions are widely used but raise questions from a theoretical perspective given that authorities could alternatively allocate efficiently via an efficient, standard auction. Taking a mechanism design perspective, we show that with appropriately determined endowments, consignment auctions are politically more palatable than standard auctions because they leave the firms better off without sacrificing efficiency. This approach also shows that efficient consignment auctions imply that the firms' incentives to invest in abatement technologies are aligned with the first-best. A simple calibration to data from Southern California's RECLAIM market shows that our assumption that the firms' endowments are such that the VCG reallocation mechanism does not run a deficit is upheld empirically. It further shows that of the three theoretical endowment structures that we considered, the one corresponding to grandfathering matches the empirical endowments best.

Interesting avenues for future research include relaxing the assumption of constant marginal values, which underlies our analysis, and allowing for a misalignment between maximizing participating firms' profits and social surplus. If such a misalignment can be fixed through the use of set-asides or bid credits, this raises in turn the question of which endowments, if any, permit efficient reallocation in the second stage without running a deficit.

References

Arrow, K. (1979): "The Property Rights Doctrine and Demand Revelation under Incomplete Information," in *Economics and Human Welfare*, ed. by M. Boskin, New York: Academic Press, 23–39.

Börgers, T. (2015): *An Introduction to the Theory of Mechanism Design*, Oxford University Press.

Börgers, T. and P. Norman (2009): "A Note on Budget Balance Under Interim Participation Constraints: The Case of Independent Types," *Economic Theory*, 39, 477–489.

Burtraw, D. and K. McCormack (2017): "Consignment Auctions of Free Emissions Allowances," *Energy Policy*, 107, 337–344.

Busch, C., H. Harvey, H. Min, and L. Shuang (2018): "Consignment Auctioning of Carbon Allowances in Cap-and-Trade Program Design," Energy Innovation: Policy and Technology.

Che, Y.-K. (2006): "Beyond the Coasian Irrelevance: Asymmetric Information," Unpublished Lecture Notes, Columbia University.

Clarke, E. (1971): "Multipart Pricing of Public Goods," *Public Choice*, 11, 17–33.

Cramton, P., R. Gibbons, and P. Klemperer (1987): "Dissolving a Partnership Efficiently," *Econometrica*, 55, 615–632.

Cramton, P. and S. Kerr (2002): "Tradeable Carbon Permit Auctions: How and Why to Auction Not Grandfather," *Energy Policy*, 30, 333–345.

d'Aspremont, C. and L.-A. Gérard-Varet (1979): "Incentives and Incomplete Information," *Journal of Public Economics*, 11, 25–45.

Delacrétaz, D., S. Loertscher, and C. Mezzetti (2022): "When Walras Meets Vickrey," *Theoretical Economics*, 17, 1803–1845.

Figueroa, N. and V. Skreta (2012): "Asymmetric Partnerships," *Economics Letters*, 115, 268–271.

Fowlie, M. and J. M. Perloff (2013): "Distributing Pollution Rights in Cap-and-Trade Programs: Are Outcomes Independent of Allocation?" *Review of Economics and Statistics*, 95, 1640–1652.

Gershkov, A., J. K. Goeree, A. Kushnir, B. Moldovanu, and X. Shi (2013): "On the Equivalence of Bayesian and Dominant Strategy Implementation." *Econometrica*, 81, 197–220.

Goulder, L. H. and R. D. Morgenstern (2018): "China's Rate-Based Approach to Reducing CO 2 Emissions: Attractions, Limitations, and Alternatives," in *AEA Papers and Proceedings*, vol. 108, 458–62.

Groves, T. (1973): "Incentives in Teams," *Econometrica*, 41, 617–631.

Hahn, R. W. and J. S. Shapiro (2011): "The Effect of Allowance Allocations on Cap-and-Trade System Performance," *Journal of Law & Economics*, 54, S267–S295.

Hahn, R. W. and R. N. Stavins (2011): "The Effect of Allowance Allocations on Cap-and-trade System Performance," *The Journal of Law and Economics*, 54, S267–S294.

Hellwig, M. F. (2003): "Public-Good Provision with Many Participants," *Review of Economic Studies*, 70, 589–614.

Holmström, B. (1979): "Groves' Scheme on Restricted Domains," *Econometrica*, 47, 1137–1144.

ICAP (2021): "Emissions Trading Worldwide: Status Report 2021," International Carbon Action Partnership Berlin, Germany.

Khezr, P. and I. A. MacKenzie (2018): "Consignment Auctions," *Journal of Environmental Economics and Management*, 87, 42–51.

Kreibich, N. and L. Hermwille (2021): "Caught in Between: Credibility and Feasibility of the Voluntary Carbon Market Post-2020," *Climate Policy*, 21, 939–957.

Krishna, V. and M. Perry (2000): "Efficient Mechanism Design," Working Paper, Penn State University.

Larsen, B. J. and A. Zhang (2021): "Quantifying Bargaining Power Under Incomplete Information: A Supply-Side Analysis of the Used-Car Industry," Working Paper, Stanford University.

Liu, Y. and B. Tan (2021): "Consignment Auctions Revisited," *Economics Letters*, 203, 1–4.

Loertscher, S. and L. M. Marx (2020): "A Dominant-Strategy Asset Market Mechanism," *Games and Economic Behavior*, 120, 1–15.

——— (2022): "Incomplete Information Bargaining with Applications to Mergers, Investment, and Vertical Integration," *American Economic Review*, 112, 616–649.

Loertscher, S., L. M. Marx, and T. Wilkening (2015): "A Long Way Coming: Designing Centralized Markets with Privately Informed Buyers and Sellers," *Journal of Economic Literature*, 53, 857–897.

Loertscher, S. and C. Mezzetti (2019): "The Deficit on Each Trade in a Vickrey Double Auction Is at Least as Large as the Walrasian Price Gap," *Journal of Mathematical Economics*, 84, 101–106.

Loertscher, S. and C. Wasser (2019): "Optimal Structure and Dissolution of Partnerships," *Theoretical Economics*, 14, 1063–1114.

Lu, H. and J. Robert (2001): "Optimal Trading Mechanisms with Ex Ante Unidentified Traders," *Journal of Economic Theory*, 97, 50–80.

Makowski, L. and C. Mezzetti (1994): "Bayesian and Weakly Robust First Best Mechanisms: Characterizations," *Journal of Economic Theory*, 64, 500–519.

Milgrom, P. (2017): *Discovering Prices*, New York: Columbia University Press.

Montgomery, W. D. (1972): "Markets in Licenses and Efftcient Pollution Control Programs," *Journal of Economic Theory*, 5, 395–418.

Myerson, R. and M. Satterthwaite (1983): "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory*, 29, 265–281.

Myerson, R. B. (1981): "Optimal Auction Design," *Mathematics of Operations Research*, 6, 58–73.

Neeman, Z. (1999): "Property Rights and Efficiency of Voluntary Bargaining under Asymmetric Information." *Review of Economic Studies*, 66, 679–691.

Pizer, W. A. and X. Zhang (2018): "China's New National Carbon Market," in *AEA Papers and Proceedings*, vol. 108, 463–67.

Schweizer, U. (2006): "Universal Possibility and Impossibility Results," *Games and Economic Behavior*, 57, 73–85.

Segal, I. and M. D. Whinston (2011): "A Simple Status Quo that Ensures Participation (With Application to Efficient Bargaining)," *Theoretical Economics*, 6, 109–125.

Vickrey, W. (1961): "Counterspeculation, Auction, and Competitive Sealed Tenders," *Journal of Finance*, 16, 8–37.

Williams, S. R. (1999): "A Characterization of Efficient, Bayesian Incentive Compatible Mechanisms," *Economic Theory*, 14, 155–180.

Xiong, L., B. Shen, S. Qi, L. Price, and B. Ye (2017): "The Allowance Mechanism of China's Carbon Trading Pilots: A Comparative Analysis with Schemes in EU and California," *Applied Energy*, 185, 1849–1859.

Zhang, S., K. Jiang, L. Wang, G. Bongers, G. Hu, and J. Li (2020): "Do the Performance and Efficiency of China's Carbon Emission Trading Market Change Over Time?" *Environmental Science and Pollution Research*, 27, 33140–33160.

A Appendix: Proofs

A.1 Proof of Lemma 1

Standard arguments imply that given r_i, i's worst-off satisfies $q_i(\hat{\theta}_i) = \min\{r_i, k_i\}$; Section 1.1 in the Online Appendix provides the derivation.

Because $\sum_{i\in\mathcal{N}} q_i(\underline{\theta}) < 1 < \sum_{i\in\mathcal{N}} q_i(\overline{\theta})$,[38] the continuity and monotonicity of the $q_i(\cdot)$ imply that there exists a unique $\hat{\theta} \in (\underline{\theta}, \overline{\theta})$ such that $\sum_{i\in\mathcal{N}} q_i(\hat{\theta}) = 1$. Because $q_i(\theta) \in [0, k_i]$ for all θ, it then follows that for all $i \in \mathcal{N}$, $q_i(\hat{\theta}) \in [0, k_i]$. Further, it cannot be that $q_i(\hat{\theta}) \in \{0, k_i\}$ for all $i \in \mathcal{N}$: if all are equal to k_i, then by our assumption of excess demand, $\sum_{i\in\mathcal{N}} q_i(\hat{\theta}) > 1$, which is a contradiction; and if $q_j(\hat{\theta}) = 0$, then $\hat{\theta} = \underline{\theta}$, and so $q_i(\hat{\theta}) = 0$ for all $i \in \mathcal{N}$, implying that $\sum_{i\in\mathcal{N}} q_i(\hat{\theta}) = 0$, which is also a contradiction. Thus, letting $r_i^* = q_i(\hat{\theta})$ for all $i \in \mathcal{N}$ and using $q_i(\hat{\theta}_i) = \min\{r_i, k_i\}$, we have proven the statement in the lemma. ■

A.2 Proof of Proposition 1

In the first part of the proof, we formalize the constrained optimization problem. Let $u_i(\theta_i, \mathbf{r})$ denote agent i's interim expected payoff from the consignment auction, not including agent i's outside option $\theta_i r_i$: $u_i(\theta_i, \mathbf{r}) \equiv q_i(\theta_i)\theta_i - t_{i,\mathbf{r}}(\theta_i)$. By the payoff equivalence theorem and the definition of $\hat{\theta}_i$ and letting individual rationality bind for type $\hat{\theta}_i$, we have

$$t_{\mathbf{r},i}(\theta_i) = q_i(\theta_i)\theta_i - \int_{\hat{\theta}_i(r_i)}^{\theta_i} q_i(x)dx - \hat{\theta}_i(r_i)r_i, \tag{A.1}$$

[38] To see that $\sum_{j\in\mathcal{N}} q_j(\underline{\theta}) < 1$, first note that $q_i(\underline{\theta}) = \max\{0, 1 - \sum_{j\neq i} k_j\}$. If for all i, $1 - \sum_{j\neq i} k_j \leq 0$, then we are done because $\sum_{j\in\mathcal{N}} q_j(\underline{\theta}) = 0 < 1$. If not, then let i be such that $1 - \sum_{j\neq i} k_j > 0$. Then $q_i(\underline{\theta}) = 1 - \sum_{j\neq i} k_j$. Also note that $q_i(\underline{\theta}) < k_i$ by the assumption that $\sum_{j\in\mathcal{N}} k_j > 1$. If i is the only agent with $1 - \sum_{j\neq i} k_j > 0$, then we are also done because $\sum_{j\in\mathcal{N}} q_j(\underline{\theta}) = q_i(\underline{\theta}) < k_i < 1$. If i is not the only one, let ℓ be the only other one. Then $q_\ell(\underline{\theta}) < k_\ell$ and $\sum_{j\in\mathcal{N}} q_j(\underline{\theta}) = q_i(\underline{\theta}) + q_\ell(\underline{\theta}) < q_i(\underline{\theta}) + k_\ell < 1$. And so $q_i(\underline{\theta}) = 1 - \sum_{j\neq i} k_j$. By induction on the number of agents with $1 - \sum_{j\neq i} k_j > 0$, we have $\sum_{j\in\mathcal{N}} q_j(\underline{\theta}) < 1$.

which implies that the ex ante expected budget surplus generated in the consignment auction is

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i\in\mathcal{N}}\left(q_i(\theta_i)\theta_i-\int_{\underline{\theta}}^{\theta_i}q_i(x)dx\right)\right]+\sum_{i\in\mathcal{N}}\left(\int_{\underline{\theta}}^{\hat{\theta}_i(r_i)}q_i(x)dx-\hat{\theta}_i(r_i)r_i\right). \tag{A.2}$$

The first component of (A.2) is determined by exogenous parameters, that is, F_i and k_i, and the interim expected efficient allocation $q_i(\theta_i)$. Maximizing (A.2) thus requires maximizing the second component of (A.2). Letting

$$\Delta(\mathbf{k})\equiv\left\{\mathbf{r}\in\mathbb{R}^n\mid\sum_{i\in\mathcal{N}}r_i=1\text{ and }\forall i\in\mathcal{N},\ \max\{0,1-\sum_{j\in\mathcal{N}\setminus\{i\}}k_j\}\le r_i\le k_i\right\},$$

the problem of maximizing the expected budget surplus generated in the consignment auction is

$$\max_{\mathbf{r}\in\Delta(\mathbf{k})}\sum_{i\in\mathcal{N}}\left[\int_{\underline{\theta}}^{\hat{\theta}_i(r_i)}q_i(x)dx-\hat{\theta}_i(r_i)r_i\right].$$

The Lagrangian for this problem is

$$\begin{aligned}L(\mathbf{r},\mu,\boldsymbol{\lambda}^o,\boldsymbol{\lambda}^N)=&\sum_{i\in\mathcal{N}}\left[\int_{\underline{\theta}}^{\hat{\theta}_i(r_i)}q_i(x)dx-\hat{\theta}_i(r_i)r_i\right]\\&+\mu(\sum_{i\in\mathcal{N}}r_i-1)+\sum_{i\in\mathcal{N}}\lambda_i^o(r_i-\max\{0,1-\sum_{j\in\mathcal{N}\setminus\{i\}}k_j\})-\sum_{i\in\mathcal{N}}\lambda_i^N(r_i-k_i).\end{aligned}$$

The Karush–Kuhn–Tucker conditions are as follows:

(a) (stationarity) $\forall i\in\mathcal{N}$, $\frac{\partial L}{\partial r_i}=0$, i.e., for all $i\in\mathcal{N}$, $\hat{\theta}_i(r_i)=\mu+\lambda_i^o-\lambda_i^N$;

(b) (complementary slackness) for all $i\in\mathcal{N}$, $\lambda_i^o(r_i-\max\{0,1-\sum_{j\in\mathcal{N}\setminus\{i\}}k_j\})=0$ and $\lambda_i^N(r_i-k_i)=0$;

(c) (primal feasibility) $\mathbf{r}\in\Delta(\mathbf{k})$;

(d) (dual feasibility) for all $i\in\mathcal{N}$, $\lambda_i^o,\lambda_i^N\ge 0$.

To characterize the local maxima, we examine three exhaustive cases depending on the signs of λ_i^o and λ_i^N:

Case 1. For all $i \in \mathcal{N}$, $\lambda_i^o = \lambda_i^N = 0$. Lemma 1 shows that this is a feasible solution. The concavity of the objective function in r_i ensures that $\mathbf{r}^*$ characterizes a local maximum.

Case 2. For some $i \in \mathcal{N}$, $\lambda_i^o > 0$ and for all $i \in \mathcal{N}$, $\lambda_i^N = 0$. Let $\mathcal{N}_o \equiv \{i \in \mathcal{N} \mid \lambda_i^o > 0\}$. Then stationarity implies that for all $i \in \mathcal{N}_o$, $\hat{\theta}_i(r_i) = \mu + \lambda_i^o$, and for all $i \in \mathcal{N}_o^c$, $\hat{\theta}_i(r_i) = \mu$. By the definition of $\hat{\theta}_i$ and complementary slackness, we then have for all $i \in \mathcal{N}_o$,

$$q_i(\mu + \lambda_i^o) = \max\{0, 1 - \sum_{j \in \mathcal{N} \setminus \{i\}} k_j\},$$

and for all $i \in \mathcal{N}_o^c$, $q_i(\mu) = r_i$. Then primary feasibility implies that

$$\sum_{i \in \mathcal{N}_o} q_i(\mu + \lambda_i^o) + \sum_{i \in \mathcal{N}_o^c} q_i(\mu) = 1.$$

If such μ exists, then for all $i \in \mathcal{N}_o$ and for any type $x \leq \mu + \lambda_i^o$, we have $q_i(x) = q_i(\mu + \lambda_i^o) = \max\{0, 1 - \sum_{j \in \mathcal{N} \setminus \{i\}} k_j\}$. Thus, for all $i \in \mathcal{N}_o$, we have $q_i(\mu) = q_i(\mu + \lambda_i^o)$, and hence, $\sum_{i \in \mathcal{N}_o} q_i(\mu) + \sum_{i \in \mathcal{N}_o^c} q_i(\mu) = 1$. The corresponding local maximum $r_i = q_i(\mu)$ for all $i \in \mathcal{N}$ is the interior solution $\mathbf{r}^*$ with $\hat{\theta} = \mu$.

Case 3. For some $i \in \mathcal{N}$, $\lambda_i^o > 0$, and for some $i \in \mathcal{N}$, $\lambda_i^N > 0$. Define $\mathcal{N}_o$ as in Case 2 and define $\mathcal{N}_N \equiv \{i \in \mathcal{N} \mid \lambda_i^N > 0\}$. Then stationarity implies that for all $i \in \mathcal{N}_N$, $\hat{\theta}_i(r_i) = \mu - \lambda_i^N$, for all $i \in \mathcal{N}_o$, $\hat{\theta}_i(r_i) = \mu + \lambda_i^o$, and for all $i \notin \mathcal{N}_o \cup \mathcal{N}_N$, $\hat{\theta}_i(r_i) = \mu$. By the definition of $\hat{\theta}_i$ and complementary slackness, we then have $\forall i \in \mathcal{N}_N$,

$$q_i(\mu - \lambda_i^N) = k_i, \tag{A.3}$$

for all $i \in \mathcal{N}_o$ and $j \notin \mathcal{N}_o \cup \mathcal{N}_N$,

$$q_i(\mu + \lambda_i^o) = \max\{0, 1 - \sum_{j \in \mathcal{N} \setminus \{i\}} k_j\}, \tag{A.4}$$

and

$$r_i = q_i(\mu). \tag{A.5}$$

Because $f_i(\theta) > 0$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$, (A.3) implies that $\mu > \overline{\theta}$. Then equations (A.4) and (A.5) only hold if $\mathcal{N}_N = \mathcal{N}$. However, this violates the primary feasibility that $\sum_{i \in \mathcal{N}} r_i = 1$. Thus, a local maximum with $\lambda_i^o > 0$ for some $i \in \mathcal{N}$ and $\lambda_j^N > 0$ for some $j \in \mathcal{N}$ does not exist. Therefore, the interior solution $\mathbf{r}^*$ maximizes the expected budget surplus in the consignment auction. ■

A.3 Proof of Theorem 1

Take $\boldsymbol{\theta}$ and $\mathbf{r}$ as given, assume that for all $i \in \mathcal{N}$, $\hat{\theta}_i = \hat{\theta}$ and fix an agent $i \in \mathcal{N}$. When the type profile is $(\hat{\theta}, \boldsymbol{\theta}_{-i})$, social welfare under the efficient allocation for that type profile, $\mathbf{Q}(\hat{\theta}, \boldsymbol{\theta}_{-i})$, is at least as large as the social welfare under the allocation $\mathbf{Q}(\boldsymbol{\theta})$, because by construction $\mathbf{Q}(\hat{\theta}, \boldsymbol{\theta}_{-i})$ maximizes social welfare given the type profile $(\hat{\theta}, \boldsymbol{\theta}_{-i})$. That is, $W(\hat{\theta}, \boldsymbol{\theta}_{-i}) \geq W(\boldsymbol{\theta}) + (\hat{\theta} - \theta_i) Q_i(\boldsymbol{\theta})$, which is equivalent to

$$W(\boldsymbol{\theta}) - W(\hat{\theta}, \boldsymbol{\theta}_{-i}) \leq (\theta_i - \hat{\theta}) Q_i(\boldsymbol{\theta}). \tag{A.6}$$

Because $\sum_{i \in \mathcal{N}} Q_i(\boldsymbol{\theta}) = 1$, summing up yields

$$\sum_{i \in \mathcal{N}} \Big(W(\boldsymbol{\theta}) - W(\hat{\theta}, \boldsymbol{\theta}_{-i}) \Big) \leq W(\boldsymbol{\theta}) - \hat{\theta}. \tag{A.7}$$

Substituting $\hat{\theta}_i = \hat{\theta}$ for all $i \in \mathcal{N}$ into (1) and using $\min\{r_i, k_i\} = r_i$, which follows from Lemma 1, yields $\Pi(\boldsymbol{\theta}, \mathbf{r}) = W(\boldsymbol{\theta}) - \hat{\theta} - \sum_{i \in \mathcal{N}} \Big(W(\boldsymbol{\theta}) - W(\hat{\theta}, \boldsymbol{\theta}_{-i}) \Big) \geq 0$, where the inequality follows from (A.7). Because our initial choice of $\boldsymbol{\theta}$ was arbitrary, $\Pi(\boldsymbol{\theta}, \mathbf{r}^*) \geq 0$ follows because by Lemma 1 $\mathbf{r}^*$ induces $\hat{\theta}_i = \hat{\theta}$ for all $i \in \mathcal{N}$. Because the inequality in (A.6) is strict for a positive measure of types (To see this, note that (A.6) holds with equality only if $Q_i(\boldsymbol{\theta}) = Q_i(\hat{\theta}, \boldsymbol{\theta}_{-i})$, which only holds for a restricted set of $\boldsymbol{\theta}$ given that $\underline{\theta} < \overline{\theta}$ and $f_i > 0$ for all i on $(\underline{\theta}, \overline{\theta})$.), $\Pi_{\mathbf{r}^*} > 0$ follows. ■

A.4 Proof of Proposition 2

We make use of the following lemma:

Lemma A.1. *If $k_1 \geq \cdots \geq k_n$ and $F_1 \leq \cdots \leq F_n$, then $q_i(\theta) \geq q_{i+1}(\theta)$ for any $\theta \in [\underline{\theta}, \overline{\theta}]$.*

Proof. Denote by $Q_i(x, y, \boldsymbol{\theta}_{-\{i,i+1\}})$ the ex post allocation to i when its type is x and $i+1$'s type is y and all other agents' types are $\boldsymbol{\theta}_{-\{i,i+1\}}$; accordingly, in this instance, $i+1$'s ex post allocation is denoted $Q_{i+1}(y, x, \boldsymbol{\theta}_{-\{i,i+1\}})$, and then note that

$$\begin{aligned} q_{i+1}(\theta) &= \mathbb{E}_{\theta_i}[\mathbb{E}_{\boldsymbol{\theta}_{-\{i,i+1\}}}[Q_{i+1}(\theta, \theta_i, \boldsymbol{\theta}_{-\{i,i+1\}})]] \leq \mathbb{E}_{\theta_i}[\mathbb{E}_{\boldsymbol{\theta}_{-\{i,i+1\}}}[Q_i(\theta, \theta_i, \boldsymbol{\theta}_{-\{i,i+1\}})]] \\ &\leq \mathbb{E}_{\theta_{i+1}}[\mathbb{E}_{\boldsymbol{\theta}_{-\{i,i+1\}}}[Q_i(\theta, \theta_{i+1}, \boldsymbol{\theta}_{-\{i,i+1\}})]] = q_i(\theta), \end{aligned}$$

where $Q_i(\theta, \theta_i, \boldsymbol{\theta}_{-\{i,i+1\}})$ in the second line denotes the allocation of firm i when firm i's type is θ, firm $i+1$'s type is θ_i, and the other types are $\boldsymbol{\theta}_{-\{i,i+1\}}$. The first inequality follows because $k_{i+1} \leq k_i$, implies $Q_{i+1}(x, y, \boldsymbol{\theta}_{-\{i,i+1\}}) \leq Q_i(y, x, \boldsymbol{\theta}_{-\{i,i+1\}})$ for all $\boldsymbol{\theta} \in [\underline{\theta}, \overline{\theta}]$. The second inequality follows from first-order stochastic dominance and $\mathbb{E}_{\boldsymbol{\theta}_{-\{i,i+1\}}}[Q_i(\theta, \theta_{i+1}, \boldsymbol{\theta}_{-\{i,i+1\}})]$ being nonincreasing in θ_{i+1}. □

Proposition 2 then follows as a direct result of Lemma A.1 and that the endowments $\mathbf{r}^*$ imply same worst-off type for all agents. With strength-ordered agents, we have $q_i(\theta) > q_j(\theta)$ for any $\theta \in (\underline{\theta}, \overline{\theta})$ and in particular for $\theta = \hat{\theta}$. Thus, $r_i^* = q_i(\hat{\theta}) > r_j^*(\hat{\theta})$ follows. ■

A.5 Proof of Proposition 4

The last part of proposition is an implication of the second, so we only need to show the first two parts. For notational ease, let $\overline{\theta}_i$ be i's worst-off type given $\overline{r}_i$. From the definition of the VCG transfers, we have $T_{\overline{\mathbf{r}},i}(\boldsymbol{\theta}) = W(\overline{\theta}_i, \boldsymbol{\theta}_{-i}) - (W(\boldsymbol{\theta}) - \theta_i Q_i(\boldsymbol{\theta})) - \overline{\theta}_i \overline{r}_i$.

We first establish interim individual rationality (IR). Because $\overline{\theta}_i = \arg\min_{\theta_i \in [\underline{\theta}, \overline{\theta}]} \mathbb{E}_{\boldsymbol{\theta}_{-i}}[\theta_i Q_i(\boldsymbol{\theta}) - T_{\overline{\mathbf{r}},i}(\boldsymbol{\theta})]$, it suffices to show that $\mathbb{E}_{\boldsymbol{\theta}_{-i}}[\overline{\theta}_i Q_i(\overline{\theta}_i, \boldsymbol{\theta}_{-i}) - T_{\overline{\mathbf{r}},i}(\overline{\theta}_i, \boldsymbol{\theta}_{-i})] \geq \overline{\theta}_i \overline{r}_i$. Observe then that, for any $\boldsymbol{\theta}_{-i}$,

$$\overline{\theta}_i Q_i(\overline{\theta}_i, \boldsymbol{\theta}_{-i}) - T_{\overline{\mathbf{r}},i}(\overline{\theta}_i, \boldsymbol{\theta}_{-i}) = \overline{\theta}_i \overline{r}_i$$

holds. Thus, interim IR is satisfied. (Notice that this actually shows that for the interim worst-off type $\overline{\theta}_i$, both the interim and the ex post IR constraints are satisfied. But this does not imply that the ex post IR constraints are satisfied for all types since there is no reason why $\overline{\theta}_i$ should be the worst-off type ex post for all $\boldsymbol{\theta}_{-i}$.)

Consider then the ex ante expected transfer from i. We have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[T_{\overline{\mathbf{r}},i}(\boldsymbol{\theta})] &= \mathbb{E}_{\boldsymbol{\theta}}[W(\overline{\theta}_i, \boldsymbol{\theta}_{-i})] - \mathbb{E}_{\boldsymbol{\theta}}[W(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}}[\theta_i Q_i(\boldsymbol{\theta})] - \overline{\theta}_i \overline{r}_i \\
&\geq \mathbb{E}_{\boldsymbol{\theta}}[(\overline{\theta}_i - \theta_i) Q_i(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}}[\theta_i Q_i(\boldsymbol{\theta})] - \overline{\theta}_i \overline{r}_i \\
&= \overline{\theta}_i \mathbb{E}_{\boldsymbol{\theta}}[Q_i(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[\theta_i Q_i(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}}[\theta_i Q_i(\boldsymbol{\theta})] - \overline{\theta}_i \overline{r}_i = 0,
\end{aligned}$$

where the inequality follows because $W(\overline{\theta}_i, \boldsymbol{\theta}_{-i}) - W(\boldsymbol{\theta}) \geq (\overline{\theta}_i - \theta_i) Q_i(\boldsymbol{\theta})$ holds (for the same reasons as (A.6) holds) and the last equality holds because $\mathbb{E}_{\boldsymbol{\theta}}[Q_i(\boldsymbol{\theta})] = \overline{r}_i$. ■

A.6 Proof of Proposition 5

We complete the proof in three steps. First, we show that ex ante revenue under the VCG mechanism is nonnegative if agents share the common worst-off type $\hat{\theta} \geq 0$ and $\sum_{i \in \mathcal{N}} r_i = \alpha$.

This revenue, denoted, $\Pi(\boldsymbol{\theta}, \mathbf{r}, \hat{\theta}, \alpha)$, is

$$\begin{aligned}\Pi(\boldsymbol{\theta}, \mathbf{r}, \hat{\theta}, \alpha) &= W(\boldsymbol{\theta}) - \sum_{i \in \mathcal{N}} \min\{r_i, k_i\}\hat{\theta} - \sum_{i \in \mathcal{N}} \Big(W(\boldsymbol{\theta}) - W(\hat{\theta}, \boldsymbol{\theta}_{-i}) \Big) \\ &\geq W(\boldsymbol{\theta}) - \alpha\hat{\theta} - \underbrace{\sum_{i \in \mathcal{N}} \Big(W(\boldsymbol{\theta}) - W(\hat{\theta}, \boldsymbol{\theta}_{-i}) \Big)}_{\leq W(\boldsymbol{\theta}) - \hat{\theta}} \geq (1-\alpha)\hat{\theta} \geq 0.\end{aligned}$$

The first inequality follows because $\sum_{i \in \mathcal{N}} \min\{r_i, k_i\} \leq \alpha$, which holds because $\sum_{i \in \mathcal{N}} r_i = \alpha$ and so $\sum_{i \in \mathcal{N}} \min\{r_i, k_i\} \leq \sum_i r_i = \alpha$. The inequality in the underbrace expression is established in (A.7) in the proof of Theorem 1 and implies the second-to-last inequality. The last one follows from $\hat{\theta} \geq 0$ and $\alpha \leq 1$.

Second, we characterize the revenue-maximizing endowments $\mathbf{r}^*(\alpha)$. The problem is the same as that considered in Proposition 1 if one changes the feasibility constraint to $\sum_{i \in \mathcal{N}} r_i = \alpha$. We continue to have the result that there exists a common worst-off type x satisfying $\sum_{i \in \mathcal{N}} q_i(x) = \alpha$. We have previously shown that there is a unique solution to $\sum_{i \in \mathcal{N}} q_i(\hat{\theta}) = 1$. Using $\alpha \in [0, 1]$, the result that for all $i \in \mathcal{N}$, $q_i(x)$ increases in x, and continuity, there is unique solution x to $\sum_{i \in \mathcal{N}} q_i(x) = \alpha$, which defines the common worst-off type, which we denote by $\hat{\theta}(\alpha)$. For all $i \in \mathcal{N}$, $q_i(x)$ is nondecreasing in x. Hence, $\hat{\theta}(\alpha)$ weakly increases in α, and strictly so if $\hat{\theta}(\alpha) > \underline{\theta}$.

Third, we show that the optimal ex ante revenue decreases in the endowment fraction α. The designer's ex ante expected profit is (see equation (A.2)):

$$\Pi_H(\alpha) = \sum_{i \in \mathcal{N}} \mathbb{E}_{\boldsymbol{\theta}} \left[q_i(\boldsymbol{\theta})\theta_i - \int_{\underline{\theta}}^{\theta_i} q_i(x)dx \right] + \sum_{i \in \mathcal{N}} \int_{\underline{\theta}}^{\hat{\theta}(\alpha)} q_i(x)dx - \alpha\hat{\theta}(\alpha).$$

Using the envelope theorem, for all $\hat{\theta}(\alpha) > 0$, $\Pi'(\alpha) = -\hat{\theta}(\alpha) < 0$, which establishes that the ex ante expected revenue weakly decreases with α, and strictly so as long as $\hat{\theta}(\alpha) > 0$. The result that the firm's payoff (weakly) increases in α follows from the facts that social welfare remains the same and the value of each firm's outside option, $q_i(\hat{\theta}(\alpha))\hat{\theta}(\alpha)$, increases. ■