

Housing Problems with Non-Rationalizable Behavior

SOPHIE BADE^{*†}

September 17, 2008

Abstract

The assumption that agents' behavior can be derived from well-defined preferences is standardly made in the theory of housing problems and housing markets. This assumption seems a strong one for some of the real life allocation problems for which economists have designed allocation mechanisms. The present paper drops the assumption of rationalizable behavior from the definition of housing problems and housing markets. Based on a suitable definition of Pareto optimality it is shown that serial dictatorship and the top trading cycles mechanism yield Pareto optimal outcomes. However, Abdulkadiroglu and Sonmez' [1] strong equivalence between these two mechanisms fails when agents' behavior cannot necessarily be rationalized. Furthermore, it turns out that without the assumption of rationalizable behavior, consistency is a strong condition on a mechanism; no consistent Pareto optimal mechanism exist when agents behavior is not rationalizable. Conversely, serial dictatorship is a consistent and Pareto optimal mechanism when agents' behavior is rationalizable. It is shown that very modest deviations from the assumption of rationalizability suffice to generate all negative results in the paper.

KEYWORDS: Housing Problems, Housing Markets, Sequential Rationalizability, Multiple Rationales.

JEL Classification Numbers: C78, D01, D60.

^{*}Department of Economics, Penn State University, 518 Kern Graduate Building, University Park, PA 16802-3306, USA. Phone: (814) 865 8871, Fax: (814) 863 - 4775

[†]I would like to thank David Bielen for thorough research assistance.

1 Introduction

Many strong results have been obtained in the literature on housing problems and housing markets. Allocation mechanisms that are Pareto optimal, neutral with respect to the names of the houses, consistent, conversely consistent and strategy-proof have been found and characterized. Relations between housing markets and housing problems have been established (some of these results can be found in Hylland and Zeckhauser [9], Shapely and Scarf [17], Roth and Postlewaite [15], Roth [16], Ma [11], Abdulkadiroglu and Sonmez [1], Ergin [6]).

While the assumption that agents' behavior can be derived from preference maximization is shared by all the named studies, it seems that this assumption is a strong one in some environments in which mechanisms have been implemented. Take kidney allocation mechanisms as an example. One difficulty with the implementation of these mechanisms is that doctors are reluctant to state complete preference rankings over kidneys. However, the same doctors do not seem to have any problem choosing the "best" kidney for a particular patient from a given set.¹ The following story might explain the contrast between the hesitance to state preferences and the readiness to choose. Consider the task to find the "best" kidney for patient x out of a set of 10 kidneys. Financial constraints might force doctors to use some preliminary quick and cheap tests, to limit the set of kidneys on which they run some more detailed and expensive tests. Call the kidney chosen according to this procedure kidney a .

Does this mean that a should be ranked above any of the other kidney's in the set? Maybe not. Consider the case in which only a and some other kidney b are available, and assume that b was eliminated following the preliminary tests in the case of the choice problem with 10 kidneys. Given that there are only 2 kidneys in the new choice problem the doctors might now be able to run the detailed and expensive tests on both of them and discover that kidney b is actually better than a for patient x . Doctors might be aware that their decision procedures can lead to such non-rationalizable choices. They might know that the decisions based on some quick and cheap tests face a risk of being overturned by more thorough research. It can therefore be quite reasonable to refuse to state complete rankings over kidney's if decisions are derived from procedures other than preference maximization. Similar stories could be told (and will be told in Section 7) to argue that the assumption of rationalizability can be a strong one in some of the typical mechanism design problems.

Observe that the deviation from rationalizability described in the above example is not an extreme one, in the sense that the decision procedure described above cannot be used to derive

¹These statements reflect a private conversation with Utku Unver, who was involved in the design and practical implementation of several kidney exchange mechanisms.

any possible choice function. Choice functions that are not rationalizable but do follow some structure have recently garnered a lot of interest in decision theory. There is a large set of studies devoted to the analysis of choice functions that can be rationalized by multiple instead of one rationale (these many rationales might be applied to different sets as in Kalai, Rubinstein and Spiegler [10], or sequentially as in Tadenuma [18], Manzini and Mariotti [12], Houy [8], Apestegua and Ballester [2] and Tadenuma and Houy [19]). Other studies concern choice functions arising from such procedures as choosing the second best, or the median according to some predetermined preferences (see Baigent and Gaertner [3] and Gaertner and Xu [7]). Choice correspondences that can be “rationalized” by incomplete preferences have been characterized by Eliaz and Ok [5]; choice functions that can be derived as subgame perfect equilibrium outcomes of extensive form games have been characterized by Xu and Zhou [21]. In this study I will first provide some results on housing problems that allow for any kind of choice functions, I will then go on to see whether these results remain valid when restricting attention to choice functions that are only “minimally irrational” according to some of the decision theoretic articles mentioned above.

The paper is organized as follows: In Section 2 I define housing problems and housing markets and develop the necessary vocabulary to describe the choice functions of agents. Using this vocabulary I characterize the Pareto sets and cores of housing problems and housing markets in Sections 3 and 4. In the latter section I show that Roth and Postlewaite’s [15] result that every housing market has a unique core has some limited validity for the case of preferences that cannot be rationalized. In Section 5 I show that the outcomes of the top trading cycles mechanism with given initial allocations differ markedly from the outcomes of serial dictatorships with given hierarchies. This result stands in sharp contrast with Abdulkadiroglu and Sonmez’ [1] strong equivalence result between these two mechanisms. In Section 6 I show that consistency is a very strong requirement on mechanisms when the agents’ choices are not necessarily rationalizable. Finally, in Section 7 I show that most of the results of the prior sections do not require extreme deviations from rationalizability. In subsection 7.2 I give detailed consideration to choice functions that can be derived from procedures like the one described in the above example on kidney exchanges.

2 The Environment

A house allocation problem (or simply **problem**) is a triplet $\mathcal{E} = (N, H, (c_i)_{i \in N})$ where $0 \neq |N| = |H| < \infty$. The set of all agents is denoted by $N = \{1, \dots, |N|\}$ and the set of all houses is denoted by H . Any non-empty subset $K \subset N$ is called a **coalition**. The set of all subsets of a set X is denoted by $\mathcal{P}(X)$. For each $i \in N$, $c_i : \mathcal{P}(H) \rightarrow H$ with $c_i(S) \in S$ is a choice function representing the choice that agent i would make when given the opportunity to choose from a set S . If a choice function c_i is rationalizable I write \succsim_i for the preferences that rationalize it. If in some housing problem $(N, H, (c_i)_{i \in N})$ all agents' choice functions are rationalizable, the housing problem can also be written as $(N, H, (\succsim_i)_{i \in N})$. Note that the preferences \succsim_i yield single-valued choice correspondences if and only if \succsim_i is a **linear order** meaning that it is transitive, asymmetric and complete. Consequently, the present definition of a house allocation problem is standard except for the assumption of choice functions instead of linear orders to describe the behavior of agents. Throughout the text, I refer to problems $(N, H, (\succsim_i)_{i \in N})$ with all \succsim_i linear orders as **standard** problems.

An **allocation** for a particular problem $\mathcal{E} = (N, H, (c_i)_{i \in N})$ as a bijection $\mu : N \rightarrow H$, where $\mu(i)$ denotes the house assigned to agent i . For a given allocation μ , agent i is **assigned** house $\mu(i)$, this house is also called agent i 's **assignment**. At times allocations will be denoted by vectors, with the i 'th entry denoting agent i 's assignment. Most examples in the text concern the case of just three agents and three houses, in this case the houses are always called x, y and z . If $|N| = |H| = 4$ the houses are called x, y, z and w . Generic houses are denoted by the Greek letters α, β and γ .

A housing market or simply **market** is defined as a quadruple $(N, H, (c_i)_{i \in N}, \mu)$, with the interpretations of N, H and c_i as above and $\mu : N \rightarrow H$ an allocation describing the **initial endowment** of the agents. The only difference between the classical definition of markets given by Shapley and Scarf [17] and the present definition is that the former uses preferences instead of choice functions to describe the behavior of agents. If for some market $(N, H, (c_i)_{i \in N}, \mu)$ all agents choice functions c_i are rationalizable by some linear orders \succsim_i , such a market can also be denoted by $(N, H, (\succsim_i)_{i \in N}, \mu)$ and is called a **standard** market.

2.1 Characterizing the Choice function

Throughout the text I will use two notions of “preference.” All of the following definitions presume a fixed problem $\mathcal{E} = (N, H, (c_i)_{i \in N})$. According to the first and weaker definition I say that agent i ***R*-prefers** house α over house β , formally $\alpha R_i \beta$, if there exists a set of houses

$S \subset H$ such that $\alpha, \beta \in S$ and $\alpha = c_i(S)$. On the other hand I say that agent i *P-prefers* house α over house β , formally $\alpha P \beta$, if $\beta \neq c_i(S)$ for all $S \subset H$ with $\alpha, \beta \in S$.² Observe that $\alpha P \beta$ holds if there exists no set such that β is chosen when α is available; conversely, there needs to exist only one set containing α and β such that α is chosen for $\alpha R \beta$ to hold. A linear order B^* is considered a *linear extension* of a binary relation B if $\alpha B \beta$ implies $\alpha B^* \beta$ for all $\alpha, \beta \in H$. Conversely, a linear order B^* is considered a *linear selection* from a binary relation B if $\alpha B^* \beta$ implies $\alpha B \beta$ for all $\alpha, \beta \in H$. When speaking about the preference of a particular agent i , R and P will be indexed by the subscript i . The following Lemma compares the notions of P -preference and R -preference to each other and establishes the point that P -preference is in a sense “deeper,” “stronger” or “more reliable” than R -preference.

Lemma 1

- P is asymmetric and acyclic, but need not be transitive or complete.
- R is complete and reflexive, but need not be transitive.
- The statement $\alpha P \beta$ holds if and only if the statement $\beta R \alpha$ does not hold. If $\alpha P \beta$ holds then $\alpha R \beta$ holds.
- The choice function c is rationalizable if and only if $P = R$.
- P has a linear extension. There exists a linear selection from R .³

Proof

- Suppose P was cyclical, that is suppose there existed $\alpha_1, \alpha_2, \dots, \alpha_n \in H$ such that $\alpha_j P \alpha_{j+1}$ for $j = 1, \dots, n - 1$ and $\alpha_n P \alpha_1$. Then we must have that $c(\{\alpha_1, \alpha_2, \dots, \alpha_n\}) = \emptyset$ as $\alpha_{j+1} \neq c(S)$ for any S with $\alpha_j, \alpha_{j+1} \in S$ for $j = 1, \dots, n - 1$ and $c(\{\alpha_1, \alpha_2, \dots, \alpha_n\}) \neq \alpha_1$ as $\alpha_n P \alpha_1$. This would imply a contradiction with the assumption that c is a function. Acyclicity implies asymmetry. To see that P need not be transitive or complete, take the following choice function on the set $H = \{x, y, z\} : c(\{x, y, z\}) = x, c(\{x, y\}) = x, c(\{y, z\}) = y, c(\{x, z\}) = z$. Observe that P consists of the two statements $x P y$ and $y P z$, however x and z remain unranked by P .

²Bernheim and Rangel [4] define a similar relation P^* for the case of choice correspondences that might not be defined on all subsets of a grand set and might depend on some ancillary condition. For the simpler case covered here (the case of a choice *function* mapping *all* subsets of H without any ancillary condition) P and P^* coincide.

³Some of the statements and proofs are very close to Bernheim and Rangel [4]. However, the results presented here differ due to the slight difference in the definition of the relations P in the present paper and P^* in their work.

- Take any $\alpha, \beta \in H$ we either have $\alpha = c(\{\alpha, \beta\})$ and therefore $\alpha R \beta$ or $\beta = c(\{\alpha, \beta\})$ and therefore $\beta R \alpha$. The completeness of R also implies its reflexivity. To see that R need not be transitive observe that zRx and xRy hold for the above example, whereas zRy does not.
- The statement $\beta R \alpha$ does not hold if and only if there exists no S with $\alpha, \beta \in S$ and $\beta \in c(S)$, which is exactly the definition of $\alpha P \beta$. If $\alpha P \beta$ holds, then $\alpha R \beta$ must hold as R is complete and $\beta R \alpha$ cannot hold by the above argument.
- The equality $P = R$ does not hold if there exist some α, β such that $\alpha R \beta$ and $\beta R \alpha$. In this case there must exist sets $S, T \subset H$ with $\alpha, \beta \in T \cap S$ such that $c(S) = \alpha$ and $c(T) = \beta$. Clearly, such a choice function cannot be rationalized. On the other hand, if c is not rationalizable then the α -axiom is violated, so there exist sets $S \subset T$ and some $\alpha \in S$ such that $\alpha \in c(T)$ but $\alpha \notin c(S)$. Let $\beta = c(S)$. Then we have that $\alpha R \beta$ as $\alpha = c(S)$ with $\alpha, \beta \in S$ but not $\alpha P \beta$ as the set T with $\alpha \in T$ has $\beta = c(T)$.
- To see that a P has a linear extension, first construct a binary relation P' such that $\alpha P' \beta$ if there exist $\alpha_1, \dots, \alpha_n \in H$ such that $\alpha P \alpha_1 P \alpha_2 \dots P \alpha_n P \beta$. Now let $\alpha P' \beta$ and $\beta P' \gamma$. This implies that there exist $\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_m \in H$ such that $\alpha P \alpha_1 P \alpha_2 \dots P \alpha_n P \beta$ and $\beta P \beta_1 P \beta_2 \dots P \beta_m P \gamma$. Combining these latter two statements yields $\alpha P' \gamma$. So P' is transitive. Observe that P' is also asymmetric. Szpilrajin's theorem guarantees the existence of a linear extension P^* of P' .⁴ Since $\alpha P \beta$ implies that $\alpha P' \beta$, P^* is also a linear extension of P . Next observe that any linear extension P^* of P is a linear selection of R . To see this observe that $\alpha P^* \beta$ implies that either $\alpha P \beta$ or that α and β are not ranked by P . In the first case $\alpha R \beta$ holds as $\alpha P \beta$ implies that $\alpha R \beta$, in the second case $\alpha R \beta$ holds as α and β are unranked by P if and only if both $\alpha R \beta$ and $\beta R \alpha$ hold.

□

⁴Given that we are only concerned with finite sets here, Szpilrajin's theorem is not needed for the conclusion; instead an elementary proof could easily be constructed. Szpilrajin's Theorem can be found in Ok [13].

3 Pareto Optimality

Mechanisms can be judged on the basis of whether they lead to Pareto optimal outcomes. This criterion cannot directly be applied here as the notion of Pareto optimality builds on the assumption of agents' preferences as primitives in the definition of housing problems and markets. In the present study preferences are neither assumed as primitives nor implied by an assumption of rationalizability. The two notions preference defined above allow me to define two kinds of Pareto rankings that are appropriate for the present context. The following Lemma relates the two notions of Pareto-optimality to each other.

Definition 1 *An allocation μ' represents a P - (R-)improvement over μ if there exists a coalition $K \subset N$ such that $\mu'(i)P_i\mu(i)$ ($\mu'(i)R_i\mu(i)$ and $\mu'(i) \neq \mu(i)$) for all $i \in K$ and $\mu'(i) = \mu(i)$ for all $i \notin K$. An allocation μ in $\mathcal{E} = (N, H, (c_i)_{i \in N})$ is called P - (R-) **Pareto optimal** if there does not exist another allocation μ' that represents a P - (R-)improvement over μ . I write $P - PO(\mathcal{E})$ ($R - PO(\mathcal{E})$) for the sets of P - (R-)Pareto optimal allocations for problem \mathcal{E} .*

Lemma 2 *Fix a housing problem $(N, H, (c_i)_{i \in N})$. $R - PO(\mathcal{E}) \subset P - PO(\mathcal{E})$. If c_i rationalizable for all i , then $R - PO(\mathcal{E}) = P - PO(\mathcal{E})$; the inverse does not hold. The set $P - PO(\mathcal{E})$ is always non-empty. The set $R - PO(\mathcal{E})$ might be empty.*

Proof Suppose $\mu \notin P - PO(\mathcal{E})$. This means there exists another allocation μ' and a coalition $K \subset N$ such that $\mu'(i)P_i\mu(i)$ for all $i \in K$ and $\mu(i) = \mu'(i)$ for $i \notin K$. Part 1 of Lemma 1 implies that $\mu'(i)R_i\mu(i)$ for all $i \in K$, $\mu'(i) \neq \mu(i)$ for all $i \in K$ and $\mu(i) = \mu'(i)$ for $i \notin K$ which in turn implies that $\mu \notin R - PO(\mathcal{E})$. The equality of $R - PO(\mathcal{E})$ and $P - PO(\mathcal{E})$ for the case of rationalizable choice functions is implied by the fact that P and R coincide in that case (Lemma 1).

To see that the inverse conclusion does not hold take the following 3 agent example:

Example 1 Let the choice functions of agents 1 and 2 be rationalizable by $x \succ_1 y \succ_1 z$ and $z \succ_2 y \succ_2 x$ respectively. Assume that the choice function of the third agent is given as follows:

$$c_3(\{x, y, z\}) = x, \quad c_3(\{x, y\}) = x, \quad c_3(\{y, z\}) = y, \quad c_3(\{x, z\}) = z.$$

Observe that (y, x, z) , (z, x, y) and (z, y, x) are not contained in either $R - PO(\mathcal{E})$ or $P - PO(\mathcal{E})$ as for each of these allocations agents 1 and 2 would be better off by exchanging their houses. Next observe that $(x, z, y) \in R - PO(\mathcal{E}) \cap P - PO(\mathcal{E})$ as agents 1 and 2 each obtain their

most preferred house under this allocation. Any alternative allocation would make at least one of them worse off (according to P_i or R_i , which coincide for the two agents as they are rational). Next observe that $(x, y, z) \notin P - PO(\mathcal{E})$ (and therefore by the first part of the current Lemma $(x, y, z) \notin R - PO(\mathcal{E})$) as agents 2 and 3 are better off according to P_2 and P_3 when exchanging their respective houses. Finally, $(y, z, x) \in R - PO(\mathcal{E}) \subset P - PO(\mathcal{E})$ as agent 2 already owns her most preferred house while agent 3 P -prefers x to y .

No separate proof of $P - PO(\mathcal{E}) \neq \emptyset$ is provided here as this is a consequence of Theorem 1, which is shown below. For an example that $R - PO(\mathcal{E})$ might be empty, take a three agent example with the following choice functions:

Example 2

$$\begin{aligned} c_1(\{x, y, z\}) &= x, & c_1(\{x, y\}) &= x, & c_1(\{y, z\}) &= y, & c_1(\{x, z\}) &= z \\ c_2(\{x, y, z\}) &= y, & c_2(\{x, y\}) &= y, & c_2(\{y, z\}) &= z, & c_2(\{x, z\}) &= x \\ c_3(\{x, y, z\}) &= z, & c_3(\{x, y\}) &= y, & c_3(\{y, z\}) &= z, & c_3(\{x, z\}) &= x. \end{aligned}$$

Observe that, on the one hand, $(x, y, z) \notin R - PO(\mathcal{E})$ as (z, y, x) R -Pareto dominates (x, y, z) . On the other hand any allocation $\mu \neq (x, y, z)$ is R -Pareto dominated by (x, y, z) .

□

Lemma 2 has the consequence that we cannot hope for mechanisms that are Pareto optimal in the sense that any problem \mathcal{E} is mapped to an allocation in $R - PO(\mathcal{E})$. Example 2 shows that for some problems \mathcal{E} the set $R - PO(\mathcal{E})$ is empty. At the same time one might develop the suspicion that it could be difficult to find mechanisms to map problems to any of the allocations in the larger of P -Pareto optimal allocations. Just as as the set $R - PO(\mathcal{E})$ can be very small the set $P - PO(\mathcal{E})$ can be very large. The allocations reached by “good” mechanism will probably be nested between the two Pareto-sets. In the next section some of the current results on R - and P -Pareto optimality are paralleled by some results on the suitably defined P - and R -cores of markets.

4 The Core

Roth and Postlewaite [15] show that any standard housing market has a non-empty and unique core (defined by weak domination). In the present section I show that this result has

some limited validity when agents' choice functions are not necessarily rationalizable. The core for the R -ranking contains at most one element and the core for the P -ranking is always non-empty. The two cores are nested; therefore Roth and Postlewaite's result directly translates to the case of non-rationalizable choice functions if and only if the cores according to both notions of preferences coincide. Before I formally state and prove all this, note that two important intuitions can be gleaned from the preceding section for the main result of this section. First of all, just like the standard definition of the core by weak domination, the present notion of the core has the feature that any of its elements are Pareto optimal. Since we know from Lemma 2 that the R -Pareto optimal set of some problems can be empty, we can already conclude that the core of some markets must be empty. Secondly, the proof that the cores according to the two notions of preference are nested follows along the same lines as the proof that the R - and P -Pareto set of a problem are nested.

Definition 2 *A matching η is in the P - (R -) core of the housing market $(N, H, (c_i)_{i \in N}, \mu)$ if there exists no coalition $K \subset N$ and matching ν such that, $\nu(K) = \mu(K)$ and $\nu(i)P_i\eta(i)$ for some $i \in K$ and $\nu(i) = \eta(i)$ for all other $i \in K$ ($\nu(i)R_i\eta(i)$ for all $i \in K$ and $\nu(i) \neq \eta(i)$ for some $i \in K$). The P - and R -cores of a market $(N, H, (c_i)_{i \in N}, \mu)$ are denoted by $P\text{-core}(N, H, (c_i)_{i \in N}, \mu)$ and $R\text{-core}(N, H, (c_i)_{i \in N}, \mu)$ respectively.*

If we consider only rationalizable choice functions, the definitions of the P - and the R -core coincide with the standard definition of the core by weak domination. In that case the core is denoted by $\text{core}(N, H, (\succsim_i)_{i \in N}, \mu)$. An analogue to the core by strict domination can be obtained by requiring that no allocation ν exists such that $\nu(i)P_i\eta(i)$ (or in the case of the R -ranking $\nu(i)R_i\eta(i)$ together with $\nu(i) \neq \eta(i)$) holds for all $i \in K$ in addition to $\nu(K) = \mu(K)$. Observe that any allocation in the P -core is P -Pareto optimal and that $\mu \in P\text{-core}(\mathcal{E}, \mu)$ for any μ that is P -Pareto optimal. The same holds for the R -core.

Theorem 1 *Fix a housing market $(N, H, (c_i)_{i \in N}, \mu)$. The P -core of this market is non-empty, and it might contain more than one allocation. The R -core of this market might be empty, and it contains at most one allocation. Finally, $R\text{-core}(N, H, (c_i)_{i \in N}, \mu) \subset P\text{-core}(N, H, (c_i)_{i \in N}, \mu)$.*

Proof Let P_i^* be a linear extension of P_i for all i . We know from Roth and Postlewaite [15] that the $\text{core}(N, H, (P_i^*)_{i \in N}, \mu)$ is non-empty; say it contains some η . There exists no coalition $K \subset N$ and matching ν such that, $\nu(K) = \mu(K)$ and $\nu(i)P_i^*\eta(i)$ for some $i \in K$ and $\nu(i) = \eta(i)$ for all other $i \in K$. This in turn implies that there exists no coalition $K \subset N$ and matching ν such that, $\nu(K) = \mu(K)$ and $\nu(i)P_i\eta(i)$ for some $i \in K$ and $\nu(i) = \eta(i)$ for all other $i \in K$.

Consequently η is in the P -core of the market $(N, H, (c_i)_{i \in N}, \mu)$.⁵

To see that P -cores need not be unique, reconsider Example 2 given in the proof of Lemma 2. Consider the housing market $(N, H, (c_i)_{i \in N}, \mu)$ with $\mu = (x, y, z)$. The allocation μ is an element of $P - core(N, H, (c_i)_{i \in N}, \mu)$, as it is itself P -Pareto optimal. To see that $\eta = (x, z, y)$ is also an element of $P - core(N, H, (c_i)_{i \in N}, \mu)$, suppose an allocation ν and a coalition K existed such that $\nu(K) = \mu(K)$ and $\nu(i) P_i \eta(i)$ for some $i \in K$ and $\nu(i) = \eta(i)$ for all other $i \in K$. Since agent 1 P -prefers neither y nor z to x , $\nu(1) = x$ must hold. For ν to differ from η we must have that $\nu(2) = y$. This implies a contradiction as $\nu(2) P_2 \eta(2)$ does not hold for $\nu(2) = y$ and $\eta(2) = z$. Thus there are (at least) two allocations in $P - core(N, H, (c_i)_{i \in N}, \mu)$.

The same example demonstrates that the R -core might be empty. We know from the proof of Lemma 2 that the R -Pareto set of the problem is empty. Since any element of the R -core has to be R -Pareto optimal, the R -core itself must be empty.

To see that the R -core of a housing market contains at most one element, observe that for any $\eta \in R - core(N, H, (c_i)_{i \in N}, \mu)$ there exists no coalition $K \subset N$ and matching ν such that $\nu(K) = \mu(K)$, $\nu(i) R_i \eta(i)$ for all $i \in K$ and $\nu(i) \neq \eta(i)$ for some $i \in K$. This implies that for any set of linear selections $(R_i^*)_{i \in N}$ of $(R_i)_{i \in N}$, there exists no coalition $K \subset N$ and matching ν such that $\nu(K) = \mu(K)$, $\nu(i) R_i^* \eta(i)$ for some $i \in K$ and $\nu(i) = \eta(i)$ for all other $i \in K$. Consequently, η is also an element of $core(N, H, (R_i^*)_{i \in N}, \mu)$ for any set of linear selections $(R_i^*)_{i \in N}$ of $(R_i)_{i \in N}$. We know from Roth and Postlewaite [15] that $core(N, H, (R_i^*)_{i \in N}, \mu)$ is unique for any $(R_i^*)_{i \in N}$. This implies that $R - core(N, H, (c_i)_{i \in N}, \mu)$ contains at most one element.

The proof of $R - core(N, H, (c_i)_{i \in N}, \mu) \subset P - core(N, H, (c_i)_{i \in N}, \mu)$ exactly parallels the proof of $R - PO(\mathcal{E}) \subset P - PO(\mathcal{E})$ in Lemma 2 and is therefore omitted. \square

5 Allocation Mechanisms

An allocation **rule** ϕ maps any problem $\mathcal{E} = (N, H, (c_i)_{i \in N})$ to a unique allocation $\phi(\mathcal{E})$. In this section I will compare two of the most prominent house allocation mechanism: serial dictatorship and top trading cycles. The first mechanism concerns house allocation problems; allocations are established by letting agents choose houses in a predetermined order. This predetermined order is given by a bijection $\pi : N \rightarrow N$ that assigns position $\pi(i)$ to agent i , such a π is called a **hierarchy**. The latter mechanism concerns housing markets; it presumes

⁵This proves the assertion that the “set $P - PO(\mathcal{E})$ is always non-empty” made in Lemma 2 as $\emptyset \neq P - core(N, H, (c_i)_{i \in N}, \mu) \subset P - PO(\mathcal{E})$ for any initial endowment μ .

an initial endowment from which agents can trade away. Except for a small adjustment in the definition of top trading cycles, the following two definitions are standard.

Definition 3 *An allocation rule ϕ is a **serial dictatorship** if there exists an hierarchy π and it holds that*

$$\begin{aligned}\phi(\mathcal{E})(\pi^{-1}(1)) &= c_{\pi^{-1}(1)}(H), \\ \phi(\mathcal{E})(\pi^{-1}(i)) &= c_{\pi^{-1}(i)}(H / (\bigcup_{k=1}^{i-1} \{\phi(\mathcal{E})(\pi^{-1}(k))\})).\end{aligned}$$

I write ϕ^π for the serial dictatorship induced by the hierarchy π . The set of all allocations that can be achieved through the rule of serial dictatorship for a given problem \mathcal{E} is called $SD(\mathcal{E}) := \bigcup_{\pi} \phi^\pi(\mathcal{E})$.

Definition 4 *An allocation rule $\phi(\cdot) := \tau(\cdot, \mu)$ is a **top trading cycle mechanism with endowment** μ if it is arrived at by the following procedure: First calculate $c_i(H)$ for all $i \in N$. Define a function $f : N \rightarrow N$, mapping agent i to the agent who is assigned the house $c_i(H)$ under μ , formally $f(i) = \mu^{-1}(c_i(H))$. Let $\phi(\mathcal{E})(i) = c_i(H)$ for any i such that $f^n(i) = i$ for some finite n .⁶ Assign all agents i that are part of such a “cycle” the house $c_i(H)$, eliminate these agents and houses from consideration. Restart the same procedure with the remaining sets of agents and houses. Since N is finite all houses are assigned in finitely many steps. For a given problem \mathcal{E} define the set of all allocations that can be achieved through the top trading cycle mechanism for some endowment μ as $TTC(\mathcal{E}) := \bigcup_{\mu} \tau(\mathcal{E}, \mu)$.*

The definition of top trading cycles given here differs slightly from some definitions in the literature insofar as that the definition given above needs to insist on a particular order of elimination of agents. To see this more clearly, reconsider Example 2 provided in the proof of Lemma 2 together with the initial endowment (x, y, z) . Construct f and observe that $f(i) = i$ for $i = 1, 2, 3$. According to the present definition all three top trading cycles are eliminated simultaneously and $\tau(\mathcal{E}, (x, y, z)) = (x, y, z)$ obtains. If on the other hand only agent 2 and house y are eliminated in round 1, the recalculation of f for the second round as yields $f(1) = 3$ and $f(3) = 1$. Given the sequential elimination of top trading cycles the allocation (z, y, x) would be obtained. Since the order of the elimination of top trading cycles does not matter in the case of standard housing markets, definitions for this case do not need to specify a particular order of elimination. The simplest choice to obtain a well-defined rule for the present case of choice

⁶The function $f^n : N \rightarrow N$ is defined as the n -th repetition of f such that $f^n(i) = f(f(\dots f(i)))$ for all i . Since N is finite we have that $f^n(i) = i$ for some i, n .

functions that are not necessarily rationalizable is to mandate that *all* top trading cycles be eliminated at once. The next Lemma describes a class of problems \mathcal{E} for which the top trading cycles mechanism determines the same allocation for any initial allocation μ . This observation will be of use in some of the later proofs.

Lemma 3 *Let $(N, H, (c_i)_{i \in N}, \mu)$ be a market such that $c_i(H) \neq c_{i'}(H)$ for all agents $i \neq i'$. Then $TTC(\mathcal{E}) = (c_i(H))_{i \in N}$.*

Proof Observe that $\tau(\mathcal{E}, \mu) = (c_i(H))_{i \in N}$ holds if every agent is assigned a house in the first round of elimination of top trading cycles. So suppose that for some μ not all agents are eliminated in the first round of top trading cycles. Partition the set of agents N into N_1 , the (non-empty) set of agents such that $f^n(i) = i$ for some $n \in \mathbb{N}$ where f is defined as in the definition of the top trading cycles mechanism, and $N_2 = N/N_1$, the remainder. Since N_2 is finite, for any $i \in N_2$ one can find an $n' \in \mathbb{N}$ such that $f^{n'}(i) \in N_1$. There must, in particular, exist a agent $i' \in N_2$ such that $f(i') = i'' \in N_1$; for this agent we have that $c_{i'}(H) = \mu(i'')$. But since $f^n(i'') = i''$, there also exists an agent $i''' \in N_1$ for whom $c_{i'''}(H) = \mu(i'')$ holds, a contradiction with the assumption that $c_i(H) \neq c_{i'}(H)$ for all agents $i \neq i'$. \square

For the standard case, a rule is considered Pareto optimal if $\phi(\mathcal{E})$ is Pareto-optimal for all problems \mathcal{E} . Similarly, in the present study, a rule is considered P - (R -)Pareto optimal if $\phi(\mathcal{E}) \in P - PO(\mathcal{E})$ ($R - PO(\mathcal{E})$) for all problems \mathcal{E} . While it is relatively easy to show that serial dictatorship and the top trading cycles mechanism are Pareto optimal mechanisms for standard problems, Abdulkadiroglu and Sonmez [1] show a much stronger result for this case: they show that serial dictatorship and top trading cycles lead to the “same” allocations in the following sense. They first define two random mechanisms which determine allocations based on equiprobable lotteries. According to the first mechanism, *random serial dictatorship*, each hierarchy is selected with equal probability, and allocations are then determined by serial dictatorship. According to the second mechanism, the *core from random endowments*, each initial endowment is selected with equal probability, and allocations are then determined as the corresponding cores. Abdulkadiroglu and Sonmez [1] show that random serial dictatorship and the core from random endowments are equivalent mechanisms in the sense that they imply the same distribution over allocations. They furthermore show that the support of each of these mechanism is the entire set of Pareto optimal allocations for a problem \mathcal{E} .

We already know from Theorem 1 that the P -core of a market $(N, H, (c_i)_{i \in N}, \mu)$ need not be a singleton, whereas the R -core of the same market might be empty. The core from random endowments is therefore not even well-defined for the case of non-rationalizable choice functions.

Consequently Abdulkadiroglu and Sonmez's result cannot hold for the present context. However, for the case of rationalizable choice functions, it does hold that the unique core allocation of a market $(N, H, (c_i)_{i \in N}, \mu)$ can be determined via the top trading cycles mechanism (see Roth and Postlewaite [15]). The top trading cycles mechanism also leads to a unique allocation for every market $(N, H, (c_i)_{i \in N}, \mu)$. Therefore one might find an equivalence result between random serial dictatorship and **markets with random endowments**, in which each initial endowment is selected with equal probability and allocations are then determined through the top trading cycles mechanism.

This modification of Abdulkadiroglu and Sonmez' [1] result does not hold either for the case of non-rationalizable choice functions. In fact, not even the necessary condition that $TTC(\mathcal{E}) = SD(\mathcal{E})$ holds for the case discussed in this paper. In the context of non-rationalizable preferences, markets with random endowments and random serial dictatorship need not have the same support. To contrast the rationalizable with the non-rationalizable case, I first state the following Lemma from Abdulkadiroglu and Sonmez' [1] (the proof can be found in their article).

Lemma 4 *Fix a standard housing problem $(N, H, (\succsim_i)_{i \in N})$. Then $R - PO(\mathcal{E}) = P - PO(\mathcal{E}) = TTC(\mathcal{E}) = SD(\mathcal{E})$.*

Lemma 4 contrasts with Theorem 2 in that neither $SD(\mathcal{E})$ nor $TTC(\mathcal{E})$ need to coincide with either one of the two Pareto sets, $R - PO(\mathcal{E})$ and $P - PO(\mathcal{E})$. Furthermore, they need not coincide with each other, in fact they do not even need to be nested.

Theorem 2 *Fix a housing problem $(N, H, (c_i)_{i \in N})$. $R - PO(\mathcal{E}) \subset SD(\mathcal{E}) \cap TTC(\mathcal{E})$, $SD(\mathcal{E}) \cup TTC(\mathcal{E}) \subset P - PO(\mathcal{E})$. The inverse inclusions need not hold. The sets $SD(\mathcal{E})$, $TTC(\mathcal{E})$ need not be nested.*

Proof Let $\mu \in R - PO(\mathcal{E})$. Define $f : N \rightarrow N$ as in the description of the top trading cycles mechanism. Define $n^* = \min_n f^n(i) = i$ (by the finiteness of the problem this is well-defined). Suppose that $n^* > 1$. Define $K := \{i : f^n(i) = i \text{ for some } n < \infty\}$ and μ' by $\mu'(i) = c_i(H)$ for $i \in K$ and $\mu'(i) = \mu(i)$ for $i \notin K$. Observe that μ' R -improves on μ as $\mu'(i) R_i \mu(i)$ for all $i \in K$, $\mu'(i) \neq \mu(i)$ for all $i \in K$ and $\mu'(i) = \mu(i)$ for all $i \notin K$. So μ cannot have been in $R - PO(\mathcal{E})$. Therefore it must be true that $n^* = 1$, which implies there exists an i^* such that $\mu(i^*) = c_{i^*}(H)$. Define $\pi(i^*) = 1$. Eliminate i^* and $c_{i^*}(H) = \mu(i^*)$ from consideration. Next, by the same argument as above we can find an i' such that $\mu(i') = c_{i'}(H/\mu(i^*))$. Let $\pi(i') = 2$. The procedure yields a hierarchy π such that $\phi^\pi(\mathcal{E}) = \mu$ and therefore $R - PO(\mathcal{E}) \subset SD(\mathcal{E})$.

To show that $R-PO(\mathcal{E}) \subset TTC(\mathcal{E})$, let $\mu \in R-PO(\mathcal{E})$ and calculate $\tau(\mathcal{E}, \mu)$. Observe that no top trading cycle can contain more than 1 agent. If such a cycle existed it would be possible to R -improve upon μ , a contradiction with the assumed R -Pareto optimality of μ . Since there must exist at least one such trading cycle, there must exist an i^* such that $\mu(i^*) = c_{i^*}(H)$. Eliminate all the agents and houses that form top trading cycles (of length 1). Repeat the argument with the remaining sets H' and N' until no more agents and houses remain to obtain $\mu = \tau(\mathcal{E}, \mu)$ and therefore $R-PO(\mathcal{E}) \subset TTC(\mathcal{E})$.

To see that the inclusion $R-PO(\mathcal{E}) \subset SD(\mathcal{E}) \cap TTC(\mathcal{E})$ might be strict, reconsider Example 2. The set of R -Pareto optimal allocations for this problem is empty. On the other hand, it holds that $(x, y, z) = \tau(\mathcal{E}, (x, y, z))$ and $(x, y, z) = \phi^\pi(\mathcal{E})$ for $\pi(1) = 1$, $\pi(2) = 3$ and $\pi(3) = 2$. The subset relation holds strictly as $(x, y, z) \in SD(\mathcal{E}) \cap TTC(\mathcal{E})$.

To see that $SD(\mathcal{E}) \subset P-PO(\mathcal{E})$ holds, suppose there existed an allocation $\mu \in SD(\mathcal{E})/P-PO(\mathcal{E})$. So we can find an alternative allocation μ' and a coalition $K \subset N$ such that $\mu'(i)P_i\mu(i)$ for all $i \in K$ and $\mu(i) = \mu'(i)$ for $i \notin K$. Let $\mu = \phi^\pi(\mathcal{E})$ and let $i^* = \operatorname{argmin}_{i \in K} \pi(i)$. When it is player i^* 's turn to choose $\mu(i^*)$, $\mu'(i^*)$ are still available. But $\mu'(i^*)P_{i^*}\mu(i^*)$ implies that i^* 's choice will never be $\mu(i^*)$ when $\mu'(i^*)$ is available. So μ cannot have been generated by a serial dictatorship.

To see that $TTC(\mathcal{E}) \subset P-PO(\mathcal{E})$ holds, suppose $\mu = \tau(\mathcal{E}, \eta)$ for some initial endowment η . So we can partition the set of agents into $N = N_1 \cup N_2 \cup N_3 \cup \dots \cup N_k$ such that the agents in N_j trade their houses as part of the j 'th round of top trading cycles. Now suppose that $\mu \notin P-PO(\mathcal{E})$, so there exists an allocation μ' and a coalition $K \subset N$ such that $\mu'(i)P_i\mu(i)$ for all $i \in K$ and $\mu(i) = \mu'(i)$ for all $i \notin K$. Let i^* be such that $i^* \in N_{j^*} \cap K$ for $j^* = \min\{j : N_j \cap K \neq \emptyset\}$. This agent i^* is assigned house $\mu(i^*)$ when houses $\mu(N_{j^*} \cup N_{j^*+1}, \dots \cup N_k)$ are still available. So we must have that $c_{i^*}(\mu(N_{j^*} \cup N_{j^*+1}, \dots \cup N_k)) = \mu(i^*)$, a contradiction with $\mu'(i^*)P_{i^*}\mu(i^*)$ as $\mu'(i^*) \in \mu(N_{j^*} \cup N_{j^*+1}, \dots \cup N_k)$.

To see that $P-PO(\mathcal{E})$ need not be a subset of $TTC(\mathcal{E}) \cup SD(\mathcal{E})$ and that $TTC(\mathcal{E})$ and $SD(\mathcal{E})$ need not be nested, consider the following example.

Example 3 Let $c_1(\{x, y, z\}) = x$, $c_2(\{x, y, z\}) = y$ and $c_3(\{x, y, z\}) = z$. Assume that the choice behavior on all other sets can be rationalized by preference orders \succsim_i that rank $c_i(\{x, y, z\})$ lowest for each of the 3 agents. Lemma 3 implies that $TTC(\mathcal{E}) = \{(x, y, z)\}$. To see that $(x, y, z) \notin SD(\mathcal{E})$, suppose we had $(x, y, z) = \phi^\pi(\mathcal{E})$ for some hierarchy π . Say the first person to choose is agent i , who chooses $c_i(\{x, y, z\})$. Next observe that the second agent, agent j , would never choose $c_j(\{x, y, z\})$, as j would rank $c_j(\{x, y, z\})$ lowest for every strict subset of $\{x, y, z\}$. Therefore, since $SD(\mathcal{E}) \neq \emptyset$, the two sets $SD(\mathcal{E})$ and $TTC(\mathcal{E})$ are not nested. Finally assume

in addition that $c_1(\{y, z\}) = y$, $c_2(\{z, x\}) = z$ and $c_3(\{x, y\}) = x$. Observe that the allocation $\mu = (y, z, x)$ is an element of $P - PO(\mathcal{E})$ as for no $i \in N$ does there exist an $\alpha \in \{x, y, z\}$ such that $\alpha P_i \mu(i)$. On the other hand, μ is not in $TTC(\mathcal{E}) \cup SD(\mathcal{E})$ as $c_i(H) \neq \mu(i)$ for all agents $i = 1, 2, 3$.

□

Theorem 2 holds some good and some bad news for the two most prominent allocation mechanism for housing problems when choice functions of agents are not necessarily rationalizable. First of all, any R -Pareto optimal allocation can be reached via serial dictatorship or the top trading cycles mechanism given the appropriate hierarchy π or initial allocation μ respectively. Given that the set of R -Pareto optima might be empty, it comes as no surprise that neither one of the two mechanisms is R -Pareto optimal; such mechanisms simply cannot exist. On the one hand, it should be satisfying to know that both mechanisms are P -Pareto optimal. It might, on the other hand, be of some concern that serial dictatorship and the top trading cycles mechanism yield different allocations and that there are P -Pareto optimal allocations that cannot be reached via either one of the two mechanisms. This brings up three related questions: Does Theorem 2 continue to hold if we restrict our attention to certain “minimal” deviations from rationalizability? Maybe any P -Pareto optimal allocation can be reached via one of the two mechanisms if we only consider a particular subclass of non-rationalizable choice functions. In Section 7 I give a mostly negative answer to this question. Secondly, we might search for some alternative mechanism X through which all possible P -Pareto optimal allocations could be reached. I discuss this question further in my conclusion. Finally we might develop a notion of “ Q -preference” that stands in between the two notions proposed here and would consequently yield a Q -Pareto set that is nested between the R - and the P -Pareto set. This Q -Pareto set might coincide with the set of allocations that is reachable via either one of the two mechanisms. This approach to mitigate the problem is related to the first approach in that one would probably need more assumptions on the derivation of choice functions to define such a ranking Q . In the following section I show that serial dictatorship and the top trading cycles mechanism fare much worse when judged by the criterion of consistency. In fact, I show the stronger result that there is no P -Pareto optimal mechanism that is consistent when considering problems with non-rationalizable choice functions.

6 Consistency

In this section I show that there are no P -Pareto optimal consistent allocation rules ϕ , where consistency is a requirement on the connection between allocations $\phi(\mathcal{E})$ for problems \mathcal{E} and allocations $\phi(\mathcal{E}')$ for subproblems \mathcal{E}' that only concern a subset of the houses and agents in the original problem \mathcal{E} . More precisely, an allocation rule ϕ is called consistent if agent i gets assignment $\phi(\mathcal{E})(i)$ also when ϕ is applied to a subproblem $\mathcal{E}' = (N', H', (c'_i)_{i \in N'})$ of \mathcal{E} that differs from \mathcal{E} only insofar as that the agents N/N' and their assignments under ϕ : $\phi(N/N')$ are missing in \mathcal{E}' , so $H' = H/\phi(N/N')$. A rule ϕ is consistent if its application to a problem yields the same result as the sequential procedure according to which the first application of the rule is used to determine the assignments of a subset of agents, before the rule is reapplied to the remaining houses and agents to determine the final allocation. Some more formalism is needed to give precise definitions.⁷

A problem $\mathcal{E}' = (N', H', (c'_i)_{i \in N'}) = r_{N'}^\mu(\mathcal{E})$ is called a **reduced problem of \mathcal{E} with respect to $N' \subset N$ at μ** if $H' = \mu(N')$ and $c'_i = c_i|_{\mathcal{P}(H')}$ for all $i \in N'$. In this section I consider a fixed grand sets of agents and houses N^* and H^* , and the grand set \mathfrak{E} of all problems \mathcal{E} that can be constructed using agents and houses from these sets N^* and H^* . The preferences P and R are defined with respect to the choice functions on the set H^* . I furthermore define the set \mathfrak{F} as the set of all bijections from subsets N of N^* to subsets H of H^* . Formally these two definitions can be expressed as:

$$\begin{aligned} \mathfrak{E} &:= \{ \mathcal{E} = (N, H, (c_i)_{i \in N}) \mid \mathcal{E} \text{ a problem, } N \subset N^* \text{ and } H \subset H^* \} \\ \mathfrak{F} &:= \{ \mu : N \rightarrow H \mid N \subset N^*, H \subset H^* \text{ and } \mu \text{ a bijection} \}. \end{aligned}$$

Observe that any reduced problem \mathcal{E}' of some problem $\mathcal{E} \in \mathfrak{E}$ is a member of the grand set of problems. A rule maps the set of problems to the set of allocations such that only allocations that are feasible for a particular problem are being selected. Formally a rule is a function $\phi : \mathfrak{E} \rightarrow \mathfrak{F}$ such that $\phi(\mathcal{E})$ is an allocation for \mathcal{E} .

Definition 5 A rule $\phi : \mathfrak{E} \rightarrow \mathfrak{F}$ is **consistent** if for any problem $\mathcal{E} = (N, H, (c_i)_{i \in N}) \in \mathfrak{E}$, and any $\emptyset \neq N' \subset N$ one has $\phi(\mathcal{E})(i) = \phi(\mathcal{E}')(i)$ for all $i \in N'$ and for $\mathcal{E}' = r_{N'}^{\phi(\mathcal{E})}(\mathcal{E})$. A rule ϕ is called **pairwise consistent** if the above holds for all two agent subsets $N' \subset N$.

⁷A comprehensive discussion of the consistency condition and its applications can be found in Thomson [20].

The fact that the set of all choice functions on H contains the set of all rationalizable choice functions entails some strong restrictions on P -Pareto optimal and pairwise consistent rules for subproblems with only two agents. The following Lemma uses this idea to derive an observation on two agent subproblems which is used in turn to show that there are no P -Pareto-optimal and pairwise consistent rules when considering the larger set of all choice functions.

Lemma 5 *Let ϕ be a pairwise consistent and P -Pareto optimal decision rule. Let $\mathcal{E} = (N, H, (c_i)_{i \in N}) \in \mathfrak{E}$ be such that $N = \{j, j'\}$, $H = \{\alpha, \beta\}$, $c_j(\{\alpha, \beta\}) = \alpha$ and $c_{j'}(\{\alpha, \beta\}) = \beta$. Then $\phi(\mathcal{E})(j) = \alpha$ and $\phi(\mathcal{E})(j') = \beta$ must hold.*

Proof Define a problem $\mathcal{E}^* = (N^*, H^*, (c_i^*)_{i \in N^*})$ such that all choice functions c_i are rationalizable and $c_i^*(H^*) \neq c_{i'}^*(H^*)$ for all $i \neq i'$. Also let $c_j^*(H^*) = \alpha$ and $c_{j'}^*(H^*) = \beta$. Lemmata 3 and 4 together with the P -Pareto optimality of ϕ imply that $\phi(\mathcal{E}^*) = (c_i(H^*))_{i \in N^*}$. Pairwise consistency implies that $\phi(\mathcal{E}^*)(i) = \phi(\mathcal{E}')(i)$ for $\mathcal{E}' = r_{N'}^{\phi(\mathcal{E}^*)}(\mathcal{E}^*)$ and any $i \in N' \subset N^*$ with $|N'| = 2$. Letting $N' := \{j, j'\}$, pairwise consistency implies in particular that $\phi(\mathcal{E})(j) = \alpha$ and $\phi(\mathcal{E})(j') = \beta$ as desired. \square

Theorem 3 *Let $|N^*| = |H^*| \geq 3$. There exists no P -Pareto optimal pairwise and pairwise consistent rule $\phi : \mathfrak{E} \rightarrow \mathfrak{F}$.*

Proof

Reconsider the problem defined in Example 2 as an element of \mathfrak{E} . For ϕ to be P -Pareto Optimal it must hold that $\phi(\mathcal{E}) \in P - PO(\mathcal{E}) = \{(x, y, z), (x, z, y), (z, y, x)\}$. If $\phi(\mathcal{E}) = (x, y, z)$, then pairwise consistency would imply that $\phi(\mathcal{E}')(1) = x$ for $\mathcal{E}' = r_{\{1,3\}}^{\phi(\mathcal{E})}(\mathcal{E})$. As $c_1(\{x, z\}) = z$ and $c_3(\{x, z\}) = x$ Lemma 5 forces us to conclude, on the other hand, that $\phi(\mathcal{E}')(1) = z$ which is a contradiction. Similarly, if $\phi(\mathcal{E}) = (x, z, y)$, then pairwise consistency would imply that $\phi(r_{\{1,2\}}^{\phi(\mathcal{E})}(\mathcal{E}))(1) = x$, whereas Lemma 5 together with $c_1(\{x, z\}) = z$ and $c_2(\{x, z\}) = x$ implies that $\phi(r_{\{1,2\}}^{\phi(\mathcal{E})}(\mathcal{E}))(1) = z$, a contradiction. Finally, if $\phi(\mathcal{E}) = (z, y, x)$, then pairwise consistency and Lemma 5 together with $c_1(\{y, z\}) = y$ and $c_2(\{y, z\}) = z$ imply that both $\phi(\mathcal{E}''')(1) = z$ and $\phi(\mathcal{E}''')(1) = y$ would have to hold for $\mathcal{E}''' = r_{\{1,2\}}^{\phi(\mathcal{E})}(\mathcal{E})$, a contradiction. \square

Theorem 3 stands in strong contrast with results on consistent rules for the standard case of housing problems in which linear orders are assumed as primitives. Ergin [6] shows that simple dictatorships are not just pairwise consistent but consistent in that case. He shows furthermore that serial dictatorships are the only rules that are Pareto optimal, pairwise consistent and “neutral” in the sense that the names of houses do not matter. Serial dictatorship turns out

to be the only rule of normative or practical interest in the the set of rules satisfying pairwise consistency and neutrality. He interprets this latter result as a negative one, saying that “dropping efficiency does not allow us to recover rules having other properties of normative interest” - as long as we require pairwise consistency and neutrality. The case discussed in the present study looks markedly different: there is no rule satisfying pairwise consistency and P -Pareto optimality. Given that the choice functions of agents are themselves in a sense not consistent, pairwise consistency turns out to be a very strong requirement on rules. This brings up again the question whether a restriction to moderate deviations from rationalizability suffice to generate this negative result. In the next section I show that Theorem 3 remains valid when restricting attention to grand sets of problems \mathfrak{E} that only allow for the most minimal deviations from rationalizability.

7 Minimal Non-Rationalizability

The results of Sections 3 through 6 are mostly of a negative nature. In these sections I did show that many of the most basic results for house allocation problems and housing markets do not extend to the case of choice functions that are not rationalizable. Most proofs therefore rely on the construction of counter-examples in which at least some agents’ choice functions are not rationalizable. In the following section I investigate the question whether “extreme” deviations from rationalizability are needed to construct these examples. To do so, some measures of the degree of non-rationalizability need to be introduced and discussed.

7.1 Multiple Rationales

Kalai, Rubinstein and Spiegler [10] introduced the notion of *rationalization by multiple rationales* according to which the set of all choice problems is partitioned into different subgroups, and every subgroup of choice problems is decided by a different predetermined linear order (rationale). Any choice function can be rationalized by multiple rationales as one can simply define a rationale per choice problem. Rationalizable choice functions are distinguished by the fact that a single rationale suffices to derive the choices from all sets. Kalai, Rubinstein and Spiegler propose a notion of choice functions being more or less “rational”, which is based on the minimal number of rationales needed to derive the choice function. In accordance with their definition let me say that a choice function is *KRS-irrational of degree $m \geq 2$* if at least m rationales are needed to derive the choice function. Kalai, Rubinstein and Spiegler demonstrate

that any choice function on a set with n elements is at most KRS-irrational of degree $n - 1$. All examples used in the prior sections concern choice sets with just 3 elements. Therefore the minimal degree of KRS-irrationality suffices to generate all of the above deviations from standard results on mechanism design. Using KRS-irrationality as the criterion, the most moderate deviation from rationalizability suffices to obtain the results of the prior sections. Note that the degree of KRS-irrationality is a purely formal measure. Other authors considered deviations from rationalizability that derive from particular decision procedures or particular contexts. In the following subsections 7.2 through 7.4 I discuss such deviations from rationalizable behavior.

7.2 Sequential Rationalizability

In this section I consider a decision-maker who chooses by sequentially applying a set of fixed asymmetric binary relations to eliminate inferior alternatives. Choice functions that arise from such procedures have been studied by Tadenuma [18], Manzini and Mariotti [12], Houy [8], Apesteguia and Ballester [2] and Tadenuma and Houy [19], among others. There are many contexts in which this kind of behavior would appear natural.

Consider a classical housing problem with 100 houses. Suppose that each agent has easy access to some short descriptions of all houses. An agent who needs to choose one out of the 100 houses might use an incomplete ranking based only on the short descriptions to eliminate a good deal of the houses from consideration. The agent might then make house visits to develop a complete ranking over the remaining houses. He uses this ranking to pick the best out of the remaining houses, call this house z . Suppose there is a house called y which is rejected immediately since it is dominated by a house called x according to the short description. Also assume that according to the short description y is only dominated by x . Now consider the choice problem in which all of the houses in the prior problem except for x are available. The information contained in the short descriptions no longer suffices to eliminate y from consideration right away. Thus the agent decides to include y in the set of visited houses, in the choice problem with 99 houses. It might result that the visit reveals y to be a gem and therefore chosen out of the set of 99 houses. Clearly such behavior is inconsistent with rationalizability.

Similar stories could be told for other typical mechanism design problems. As mentioned in the introduction, financial restrictions might force doctors in kidney exchange problems to reject some kidneys based on only a few cheaper tests, they might conduct all possible tests of compatibility only on a small set of kidneys that are not immediately rejected by the preliminary tests. Participants in school assignment mechanisms might not be able to invest the same amount

of resources to develop a full preference ranking on all possible schools. In short, procedures of sequential rationalization might be particularly relevant to many typical “housing problems”. In this subsection I investigate whether a restriction to choice functions that can be derived from such procedures would overturn some of the results of the prior sections. To formally define sequential rationalizability, let $\max_P(S)$ denote the set of maximal elements in S according to P , that is $\max_P(S) := \{x \in S : \nexists y \in S : yPx\}$.

Definition 6 *A choice function $c : \mathcal{P}(H) \rightarrow H$ is a **rational shortlist method (RSM)** if there exist two asymmetric relations P^1 and P^2 such that $c(S) = \max_{P^2}(\max_{P^1}(S))$ for all $S \subset H$. A choice function is **k -sequentially rationalizable** if there exist asymmetric and transitive relations P^1, \dots, P^k such that $c(S) = \operatorname{argmax}_{P^k}(\operatorname{argmax}_{P^{k-1}}(\dots \operatorname{argmax}_{P^1}(S)))$.*

The definition of a rational shortlist method is taken straight from Manzini and Mariotti [12], who provide an axiomatic characterization of all choice functions that are RSMs. Different from Manzini and Mariotti’s notion of k -sequential rationalizability, the present notion of k -sequential rationalizability requires not only that the relations P^j be asymmetric but also that they be transitive. The present definition is, therefore, more in tune with Tadenuma [18], who studied the sequential application of two binary relations, namely Pareto domination and a criterion of equity. The assumption that each of the sequentially applied rationales is transitive also fits the motivational stories better. While the quick and cheap tests on the match of different kidneys might not yield a full ranking on all kidneys the ranking they do provide should be transitive. The same holds for the ranking of apartments based solely on newspaper listings. Transitivity is often interpreted as “the” basic characteristic of rationality; consequently, the present definition of k -sequential rationalizability represents a more moderate deviation from rationalizability than does Manzini and Mariotti’s. Note that any choice function that is 2-sequentially rationalizable is also an RSM; the converse does not hold.

It should be noted that subscripts of binary relations continue to denote different agents, whereas the superscripts denote the different rationales applied by an agent. The statement xP_2^3y should be read to mean that agent 2 ranks x above y according to his third rationale. Note that RSMs as well as k -sequentially rationalizable choice functions can either be defined through complete lists of choices $c(S)$ for all subsets $S \subset H$ or through the statement of the rationales P^j . If H contains more than 3 elements, the former approach turns out to be more cumbersome than the latter. Therefore I will describe RSMs and k -sequentially rationalizable choice functions on sets H with $|H| > 3$ by a list of the rationales P^j . To further save on notation the preference statements $\alpha P\beta$, $\beta P\gamma$ and $\alpha P\gamma$ are sometimes summarized as $\alpha P\beta P\gamma$. The following preliminary remarks are in order:

Remark 1 • Let c be an RSM or k -sequentially rationalizable. If xP^1y holds for some x, y then xPy holds; the inverse does not hold.

- Let $c : \mathcal{P}(\{x, y, z\}) \rightarrow H$ be an RSM or k -sequentially rationalizable. Then c is either rationalizable or we have that $c(\{\alpha, \beta, \gamma\}) = \alpha$, $c(\{\alpha, \beta\}) = \alpha$, $c(\{\alpha, \gamma\}) = \gamma$ and $c(\{\beta, \gamma\}) = \beta$ for $\{\alpha, \beta, \gamma\} = \{x, y, z\}$. The unique k -sequential rationalization of c is $\beta P^1 \gamma$, $\gamma P^2 \alpha P^2 \beta$.

The following Lemma provides a preliminary step towards the characterization of P -Pareto optima when all agents choice functions are RSMs or k -sequentially rationalizable.

Lemma 6 *Let $\mu \in P - PO(\mathcal{E})$ and let all $(c_i)_{i \in N}$ be RSMs. Then there exists an i such that $\mu(i) = c_i(H)$. If the choice functions $(c_i)_{i \in N}$ are k -sequentially rationalizable with $k > 2$ rationales, then no such i needs to exist.*

Proof Let all c_i be RSMs. Suppose there existed a $\mu \in P - PO(\mathcal{E})$ such that $\mu(i) \neq c_i(H)$ for all agents i . For agent i this implies that there either exists an $\alpha \in H$ such that $\alpha P_i^1 \mu(i)$ or some $\beta \in H$ such that $\beta P_i^2 \mu(i)$ where $\gamma P_i^1 \beta$ does not hold for any $\gamma \in H$. In the first case, $\alpha P_i \mu(i)$ holds; in the latter case, $\beta P_i \mu(i)$ holds. So for each agent i , a house $\nu(i)$ can be found such that $\nu(i) P_i \mu(i)$. Define a function $f : N \rightarrow N$ that maps an agent i to the agent that is assigned the house $\nu(i)$, formally $f(i) = \mu^{-1}(\nu(i))$. The finiteness of the problem implies that $f^n(i) = i$ for some $i \in N$ and $n < \infty$. Also observe that n must be greater or equal 2 as no agent can P -prefer his own assignment to his own assignment. Observe that the alternative allocation μ' with $\mu'(i) = \nu(i)$ for all i such that $f^n(i) = i$ for some n and $\mu(i) = \mu'(i)$ otherwise P -improves upon μ which constitutes a contradiction with μ being P -Pareto optimal

To see that the claim need not hold for 3-sequentially rationalizable choice functions, consider the following example:

Example 4 Let $H = \{x, y, z, w\}$ and let the agent 1's choice function be given by the following (transitive) sequential rationales wP_1^1z , zP_1^2y , $yP_1^3xP_1^3w$, $yP_1^3xP_1^3z$, yielding the following choice function c_1 :

$$c_1(\{x, y, z, w\}) = y, \quad c_1(\{x, y, z\}) = x, \quad c_1(\{x, y, w\}) = y, \quad c_1(\{y, z, w\}) = y, \quad c_1(\{x, z, w\}) = x, \\ c_1(\{x, y\}) = y, \quad c_1(\{x, z\}) = x, \quad c_1(\{x, w\}) = x, \quad c_1(\{y, z\}) = z, \quad c_1(\{y, w\}) = y, \quad c_1(\{z, w\}) = w.$$

Observe that $x \neq c_1(H)$ while for all $\alpha = y, z, w$ there exists a set $S \subset H$ such that $\alpha, x \in S$ and $x = c_1(S)$; therefore, there does not exist an $\alpha \in H$ such that $\alpha P_1 x$.

Assume that the choice functions of the other three agents differ from agent 1's choice function only insofar that house x switches roles with houses y, z, w respectively for agents 2, 3 and 4. Consider $\mu = (x, y, z, w)$. By construction, $\alpha P_i \mu(i)$ does not hold for any $i = 1, 2, 3, 4$ and $\alpha = x, y, z, w$; therefore, the allocation μ is P-Pareto optimal. Also observe that by the construction of the choice functions, $\mu(i) \neq c_i(H)$ for all $i = 1, 2, 3, 4$.

□

Observe that Example 3 used in the proof of Theorem 2 uses choice functions that are neither RSMs nor k -sequentially rationalizable. Based on this proof it cannot be claimed that a “minimal” deviation from rationalizability suffices to show that $SD(\mathcal{E}) \neq P-PO(\mathcal{E}) \neq TTC(\mathcal{E})$. I investigate this question further in the next set of results. First I show that all statements of Theorem 2 except for $TTC(\mathcal{E}) \cup SD(\mathcal{E}) \subsetneq P-PO(\mathcal{E})$ hold for the case in which there are only 3 houses and agents and all these agents' choice functions are RSMs.

Lemma 7 *Let $(N, H, (c_i)_{i \in N})$ be such that $|N| = |H| = 3$ and c_i is an RSM for $i = 1, 2, 3$. Then $R-PO(\mathcal{E}) \subset SD(\mathcal{E}) \cap TTC(\mathcal{E})$, $SD(\mathcal{E}) \subset P-PO(\mathcal{E})$, and $TTC(\mathcal{E}) \subset P-PO(\mathcal{E})$. The inverse inclusions need not hold. The sets $SD(\mathcal{E})$, $TTC(\mathcal{E})$ need not be nested. Also $P-PO(\mathcal{E}) = SD(\mathcal{E}) \cup TTC(\mathcal{E})$.*

Proof The following example shows that $TTC(\mathcal{E}) \not\subseteq SD(\mathcal{E}) \subsetneq P-PO(\mathcal{E})$.

Example 5 Let the choice functions of agents 1 and 2 be given by

$$\begin{aligned} c_1(\{x, y, z\}) &= x, & c_1(\{x, y\}) &= x, & c_1(\{y, z\}) &= y, & c_1(\{x, z\}) &= z \\ c_2(\{x, y, z\}) &= y, & c_2(\{x, y\}) &= y, & c_2(\{y, z\}) &= z, & c_2(\{x, z\}) &= x. \end{aligned}$$

Let agent 3's choice function be rationalizable by $x \succ_3 y \succ_3 z$. Observe that the allocation $\mu = (x, y, z)$ is P-Pareto optimal as there exists no $\alpha \in \{x, y, z\}$ such that either $\alpha P_1 x$ or $\alpha P_2 y$. Next observe that $(x, y, z) = \tau(\mathcal{E}, (x, y, z))$, so $(x, y, z) \in TTC(\mathcal{E})$. Finally, to see that $(x, y, z) \notin SD(\mathcal{E})$, suppose there existed a hierarchy π such that $\phi^\pi(\mathcal{E}) = (x, y, z)$. Observe that $\pi(3) = 1$ cannot hold as agent 3 would choose x if he was to choose first. Observe that $\pi(1) = 1$ cannot hold as neither agent 2 nor agent 3 would pick their assignment under μ as second movers since $c_2(\{y, z\}) = z$ and $c_3(\{y, z\}) = y$. The case that $\pi(2) = 1$ fails by the same reasoning.

The following example shows that $SD(\mathcal{E}) \not\subseteq TTC(\mathcal{E}) \subsetneq P-PO(\mathcal{E})$.

Example 6 Let the choice functions of agents 1 and 2 be rationalizable by $x \succ_1 y \succ_1 z$ and $z \succ_2 y \succ_2 x$ respectively. Let agent 3's choice function be given by

$$c_3(\{x, y, z\}) = y, \quad c_3(\{x, y\}) = y, \quad c_3(\{y, z\}) = z, \quad c_3(\{x, z\}) = x.$$

Observe that the allocation (x, y, z) is P -Pareto optimal as agent 1 obtains his most preferred house and houses y, z are not ranked by P_3 . Since $c_i(H) \neq c_j(H)$ holds for all $i \neq j$, Lemma 3 implies that $TTC(\mathcal{E}) = \{(c_1(H), c_2(H), c_3(H))\} = \{(x, z, y)\}$ and therefore $(x, y, z) \notin TTC(\mathcal{E})$. Finally observe that $\phi^\pi(\mathcal{E}) = (x, y, z)$ for $\pi(i) = i$.

To show the equality $P-PO(\mathcal{E}) = SD(\mathcal{E}) \cup TTC(\mathcal{E})$, suppose that $\mu = (\alpha, \beta, \gamma) \in P-PO(\mathcal{E})$. We know from Lemma 6 that $\mu(i) = c_i(H)$ for at least one agent i . Assume without loss of generality that $\mu(1) = \alpha$. Next for $(\alpha, \beta, \gamma) \notin TTC(\mathcal{E})$ both $c_2(H) \neq \beta$ and $c_3(H) \neq \gamma$ need to hold. For $(\alpha, \beta, \gamma) \notin SD(\mathcal{E})$ both $c_2(\{\beta, \gamma\}) \neq \beta$ and $c_3(\{\beta, \gamma\}) \neq \gamma$ need to hold. Together, these negations imply that $\gamma P_2 \beta$ and $\beta P_3 \gamma$, which in turn implies that $(\alpha, \beta, \gamma) \notin P-PO$, a contradiction. So we must have that $P-PO(\mathcal{E}) \subset SD(\mathcal{E}) \cup TTC(\mathcal{E})$. The inverse inclusion follows from Theorem 2. \square

The main positive result of the preceding Lemma, $SD(\mathcal{E}) \cup TTC(\mathcal{E}) = P-PO(\mathcal{E})$, is very fragile. Lemma 6 shows that not even the necessary condition for $P-PO(\mathcal{E}) \subset TTC(\mathcal{E}) \cup SD(\mathcal{E})$, namely that at least one agent i is assigned the house $c_i(H)$ under a P -Pareto optimal μ , has to hold for choice functions that are k -sequentially rationalizable for $k \geq 3$. This implies that Example 3 in the proof of Theorem 2 could well be replaced by an example that uses only choice functions that are 3-sequentially rationalizable. The question whether the result that $P-PO(\mathcal{E}) \subset TTC(\mathcal{E}) \cup SD(\mathcal{E})$ survives the minimal deviation from rationalizability to 2-sequential rationalizability with more than 3 houses and agents implies that is addressed next.

Lemma 8 Fix a housing problem $(\{1, 2, 3, 4\}, H, (\succsim_i)_{i \in N})$. Let $(c_i)_{i=1,2,3,4}$ be 2-sequentially rationalizable. Then $P-PO(\mathcal{E}) = TTC(\mathcal{E}) \cup SD(\mathcal{E})$ need not be true.

Proof The following housing problem with 4 houses and agents contains a P -Pareto optimal allocation μ that cannot be reached either via top trading cycles or via serial dictatorship.

Example 7 Let the choice functions of agents 1 and 4 be rationalizable by $x \succ_i y \succ_i z \succ_i w$ for $i = 1, 4$. Let the choice functions of agents 2 and 3 be sequentially rationalizable by

$$\begin{aligned} x P_2^1 y, \quad x P_2^1 w P_2^1 z, \quad z P_2^2 y P_2^2 w \\ y P_3^1 z, \quad y P_3^1 x P_3^1 w, \quad w P_3^2 z P_3^2 x. \end{aligned}$$

Observe that the allocation $\mu = (x, y, z, w)$ is P -Pareto optimal. To see this, suppose that some μ' did improve upon μ . Since agent 1 obtains $\mu(1) = x$, his most preferred house, it must be true that $\mu(1) = \mu'(1) = x$. Next observe that the only house that agent 2 would P -prefer to y , his assignment under μ , is x . Therefore we must also have that $\mu(2) = \mu'(2) = y$. Finally an exchange of the assignments of agents 3 and 4 does not P -improve upon μ as agent 3 does not P -rank houses z and w .

Next observe that μ cannot be obtained through a serial dictatorship. To see this observe first of all that $c_i(H) = \mu(i)$ only holds for agent 1, so he would have to go first. Next observe that $c_i(\{y, z, w\}) = \mu(i)$ only holds for agent 2 so he would have to go second. Finally observe that neither $c_3(\{w, z\}) = z$ nor $c_4(\{w, z\}) = w$ holds, so neither of these two agents would choose $\mu(i)$ as the third mover. Consequently, μ cannot be obtained as the result of a serial dictatorship. The argument that $\mu \notin TTC(\mathcal{E})$ follows along very similar lines: Suppose $\mu = (x, y, z, w)$ was reachable via top trading cycles. Since $c_i(H) = \mu(i)$ only holds for $i = 1$, agent 1 is the only agent who can leave the mechanism in the first round. By the same argument the second round of top trading cycles must consist of agent 2 and house y . Finally, since $c_3(\{w, z\}) = w$ and $c_4(\{w, z\}) = z$, neither one of the two would be assigned $\mu(i)$ in the third round. For the example we have that $SD(\mathcal{E}) \cup TTC(\mathcal{E}) \not\subseteq P - PO(\mathcal{E})$.

□

It needs to be concluded that the restriction to sequentially rationalizable choice functions does only minimally change the results of Sections 3 through 6. Nearly all examples used in the proofs of the results of these sections are 2-sequentially rationalizable, in other words, nearly all examples deviate only minimally from rationalizability, in the sense that the sequential application of at most 2 (transitive) rationales suffices to construct these examples. Only the proof of Theorem 2 needed a minor revision. In Lemma 7 I showed that the subset relation $TTC(\mathcal{E}) \cup SD(\mathcal{E}) \subset P - PO(\mathcal{E})$ is never strict when restricting attention to 3-house 3-agent problems in which all choice functions are RSMs. This result already fails for housing problems with 4 agents. A restriction to sequentially rationalizable choice functions is therefore not sufficient to obtain the equality $P - PO(\mathcal{E}) = TTC(\mathcal{E}) \cup SD(\mathcal{E})$. To obtain mechanisms that would make the entire P -Pareto set reachable, we would either have to impose some stronger restrictions on the deviation from rationalizability or consider some alternatives to serial dictatorship and top trading cycles when sticking to sequential rationalizability. In the following two subsections I consider some more deviations from rationalizability. In the conclusion I quickly discuss the latter approach.

7.3 Rationalizability by Game Trees

Xu and Zhou [21] introduce and axiomatically characterize the notion of rationalizability by game trees. A choice function is rationalizable by a game tree if there exists an extensive form game with multiple players such that every choice can be explained as the subgame perfect outcome of that game.

Definition 7 *A choice function c is **rationalizable by game trees** if there exists a game tree G such that*

$$c(S) = SPNE(G|S; (\succsim_j)_{j \in J}) \text{ for all } S \subset H$$

where $G|S$ is the reduced tree of G that retains all branches of G leading to terminal nodes in S , J is a set of players, \succsim_j is player j 's linear order over all terminal nodes in G , and $SPNE(\cdot)$ stands for the subgame perfect outcome of the game.

Choice functions that are rationalizable by game trees might be very relevant for the analysis of mechanisms with collectives as agents. Consider standard housing problems with families as agents. If these families do not succeed in establishing transitive preference orders over all houses, but if their decision can instead be analyzed as the subgame perfect outcomes of games played within the family, then it would be appropriate to consider choice functions that are rationalizable by game trees. The same holds for the case of kidney exchange mechanisms when entire teams of doctors need to decide on which kidney should be chosen for which patient. It is therefore important to check if and how the results of the current paper would change if one was to restrict attention to choice functions that are rationalizable by game trees.

Apesteguia and Ballester [2] show that any choice function that is rationalizable by a game tree is sequentially rationalizable by k rationales, where sequential rationalizability by k rationales as defined by Apesteguia and Ballester differs from k -sequential rationalizability as it is defined here in that the rationales P^j are not required to be transitive. It is important to note that the one type of non-rationalizable choice function for a set H containing just 3 alternatives that can be explained as an RSM or a 2-sequentially rationalizable choice function is also rationalizable by a game tree. Consequently the gist of the results in this paper also holds for choice functions that are rationalizable by game trees. Lemmata 2 and 3 as well as Theorems 1 and 3 directly translate to this case. Theorem 2 only holds in its modified form given in Lemmata 7 and 8. It should be immediately apparent that Lemma 7 also holds for the case of choice functions that are rationalizable by game trees since this results only concerns the case of $|H| = 3$.

Given that an example with 4 houses was needed to show Lemma 8, it is not immediately apparent whether all P -Pareto optimal allocations are reachable through either serial dictatorship or top trading cycles when considering only choice functions that are rationalizable by game trees. To see that the result also holds for the more restricted set of choice functions it needs to be shown that the choice functions used in Example 7 are rationalizable by game trees. Observe first of all that the choice functions of agents 1 and 2 are rationalizable and therefore rationalizable by game trees. Observe next, that the other two agents choice functions are identical up to a renaming of alternatives. It is therefore sufficient to show that agent 2's choice function is rationalizable by a game tree. To see that this holds true consider a game tree with two nodes, as given in Figure 1. Player a first gets to choose whether to end the game with x or y as the final outcome or to continue the game. If he continues the game player b can end the game with either w or z as the final outcome. The player's preferences over the outcomes are given by the linear orders: $x \succ_a z \succ_a y \succ_a w$ and $w \succ_b z \succ_b x \succ_b y$. It is easy to check that this tree generates the same choices as do the sequential rationales xP_2^1y , $xP_2^1wP_2^1z$, $zP_2^2yP_2^2w$ given in Example 7. What is more, this game tree has only two players and two nodes. While I don't intend to formally define yet another measure of the distance between rationalizable choices and choices that are rationalizable by game trees, it should be clear that however one defines such a measure, the minimal deviation from rationalizability would have to allow for (at least) two players and two nodes. The notion of rationalizability by game trees is more restrictive than the notion of k -sequential rationalizability. This reinterpretation of Example 7 disappoints this hope, that all P -Pareto optimal allocations would be reachable through either serial dictatorship or top trading cycles when restricting attention only to choice functions that are rationalizable by game trees. Game trees with only 2 nodes and 2 players suffice to construct housing problems that have P -Pareto optimal allocations that cannot be reached by either one of the two discussed mechanisms.

7.4 Other Deviations from Rationalizability

In this section I provide reasons why some deviations from rationalizability do not appear in the present analysis even though they have gained prominence in the literature on decision theory. Baigent and Gaertner [3], for instance, characterize the behavior of agents that always choose the second best according to some predetermined linear order. This might potentially lead to different results as there is exactly one kind of choice function over a set of 3 houses $\{x, y, z\}$ which is given by (up to renaming of houses)

$$c(\{x, y, z\}) = y, \quad c(\{y, z\}) = z, \quad c(\{x, z\}) = z, \quad c(\{x, y\}) = y.$$

This choice function is not sequentially rationalizable. What is more, the one type of choice function used in the above examples that is not sequentially rationalizable (in Example 3) cannot be explained as the choices of an agent who always chooses the second best.

There are two reasons why I did not investigate how the results of the main body of the paper would change if we would restrict attention to agents that always choose the second best. First of all Baigent and Gaertner [3] motivate their study by various examples according to which agents derive some form of social esteem from not picking the best option: An agent who decides to pick the largest piece of cake runs the risk to appear greedy. “Pupils in class may be reluctant to offer an answer to a question first; party goers may be reluctant to be the most elegantly dressed or present their host with the most expensive bottle of wine; residents may not wish to drive the most luxurious car in the street.” (Baigent and Gaertner [3] p. 241) I could not think of convincing stories that would make such a desire to appear modest relevant in typical housing problems. Moreover, in many typical housing problems there is no underlying shared order over all alternatives that could guide such considerations of modesty. Take kidney allocation mechanisms as an example. Kidneys of younger and healthier donors are probably always preferred to these of older and less healthy donors. But the much more important question is one of fit. A doctor that has the right to choose first in a serial dictatorship mechanism does not necessarily render a service to the remaining patients by choosing the second best kidney for his own patient; the first best kidney for one patient could be the worst for another. In any case, the relevance of the social esteem motivation is questionable within the framework of anonymous mechanisms. In short, it is doubtful whether this behavior is very relevant for typical mechanisms.

The second reason not to extend the present results on the properties of mechanisms in the case when agents choose the second best is that this choice behavior would allow for much more stringent definitions of preference and Pareto optimality than the ones used in the main body of the text. If the underlying orders are indeed shared by all agents, as is in the case of objectively smaller or bigger pieces of the cake, the question for Pareto optimality even becomes trivial; any allocation would then be Pareto optimal.

It also has to be mentioned that there is a large set of choice theories that do not necessarily yield choice functions but rather multi-valued choice correspondences. Eliaz and Ok [5], for instance, provide an axiomatic characterization of choice correspondences that can be rationalized by transitive but not necessarily complete binary orders. While I consider the study of mechanisms when agents have incomplete preferences very worthwhile and relevant this subject goes beyond the scope of the present paper. In the present paper I generalized housing problems and

housing markets in exactly one dimension by replacing the assumption of linear orders on all available houses with the assumption of choice functions. Allowing for choice correspondences would amount to yet another step of generalization. It would then be appropriate to compare the results of standard housing problems with preferences that allow for indifferences to some results for housing problems with general choice correspondences. The study of this important question should go hand in hand with the further development of the theory of housing problems with preferences that are complete and transitive without necessarily being asymmetric.

8 Conclusion

One of the questions that remains open is whether there exist any practicable mechanisms that can reach the entire P -Pareto optimal set of a housing problem. In Section 7 I showed that a restriction to sequentially rationalizable choice functions or even choice functions that are rationalizable by game trees does not help matters for problems with more than 3 houses: some P -Pareto optimal allocations cannot be reached by either serial dictatorship or top trading cycles.

What about other mechanisms? Papai [14], for instance, defines a large set of sequential choice mechanisms that differ by the initial ownership of houses and the rules for the inheritance of these houses; serial dictatorship and top trading cycles are both embedded in this set. Any allocation that is reachable by a mechanism in this set has the feature that at least one agent is allocated his choice out of the grand set of houses. Lemma 6 can, therefore, be used to show that there are P -Pareto optimal allocations that cannot be reached by any of the mechanisms in the large set of mechanisms defined by Papai, even when restricting attention to k -sequentially rationalizable choice functions: While there are P -Pareto optimal allocations such that no agent is assigned his choice out of the grand set, no such allocation can be reached through one of the mechanisms suggested by Papai.

Alternatively one could find a notion of Pareto optimality that is nested between R - and P -Pareto optimality. Note that the concepts of R - and P - preference were defined without any reference to particular decision procedures or contexts. When restricting attention to particular deviations from rationalizability such as sequential rationalizability or rationalizability by game trees, it might be possible to use some of the information about these procedures to derive an intermediate notion of preference. It could well be that all the Pareto optima according to this notion of preference are reachable through either serial dictatorship or top trading cycles.

Next observe that no results on strategy-proofness of allocation mechanisms, such as the results by Roth [16] and Ma [11], appear in the present paper. Standardly a mechanism is

considered strategy-proof if no agent can obtain a better assignment by misrepresenting his preference. The first problem with the question whether a mechanism is strategy proof in the present context is that it is not clear what would constitute a “better assignment”. We could probably again use the notions of P - and R - preference to define this notion. The bigger problem lies in the interpretation of non-rationalizable behavior. In this paper I advocate the idea that non-rationalizable behavior should particularly be expected if agents do not have the time or resources to come up with a complete choice function or a complete ranking on all available alternatives. The leading example was that of some doctors who use a procedure to pick a kidney from a set without establishing a complete and transitive ranking on all kidneys. A translation of the notion of strategy-proofness to the present context would require that no agent could get a P - (or R -)preferred assignment by misrepresenting his choice function. So that any such agent should not only *know* his choice function, which as argued above, is maybe too strong an assumption for the present context. What is more, in addition to knowing the entire choice function the agent should understand how the outcome of the mechanism would change for all possible alternative statements of his choice function. In short, the assumptions on self-knowledge and rationality embodied in question of strategy-proofness do clash with one of the main motivations for non-rationalizable behavior underlying the present study

Probably the most interesting venue for research is to take the intuition behind the introductory story about the doctor’s decision process more seriously and to assume that some kidneys are objectively better than others for a patient but that it is costly to precisely learn the ranking over all the kidneys. In more general terms, it could be assumed that agents have (complete and transitive) preferences over objects but that it is costly to discover these preferences. The allocation mechanism would then interact with some strategic information acquisition. Interesting results on the comparison between serial dictatorship and top trading cycles should be obtainable. Suppose agents can choose to acquire costly information at any stage in the mechanism. In a serial dictatorship, agents can very economically tailor their information acquisition investment to establish choices that are relevant; each agent is called to choose only once and when he is called to choose he knows the entire set of alternatives he can choose from. Conversely, in the top trading cycles mechanism there is a danger for wasteful information acquisition. According to that mechanism each agent is called to state a choice from the grand set of alternatives, but not every agent is assigned that choice. If agents invest into information acquisition in the first stage some agents will overinvest in information acquisition. Consider the scenario in which some objects α and β are assigned in the first top trading cycle and in which agent i still stands empty handed after that first cycle. If agent i has acquired some costly information about the

relative ranking of α and β then this information was of no use to him as he was not assigned either one of them in the first cycle. Additionally, the information is of no further use as the two objects are no longer available in any of the following cycles. Conversely, if no agent acquires any information about the ranking of objects in the first top trading cycles, the agents that are assigned an object in the first cycle might later regret not having invested in some information acquisition.

Many interesting and related questions arise very quickly: how do the two mechanisms compare if preferences of agents are correlated? How do they compare if all information on preference rankings has to be acquired before the mechanisms are played? What is the optimal mechanism given that information about rankings is acquired endogenously?

References

- [1] Abdulkadiroglu, A. and T. Sonmez: “Random Serial Dictatorship and the Core from Random Endowments in House Allocation Problems”, *Econometrica*, 66, (1998), pp. 689-701.
- [2] Apestegua J. and M. Ballester: “A Characterization of Sequential Rationalizability”, *mimeo*, Universidad Autonoma de Barcelona, 2008.
- [3] Baigent, N. and W. Gaertner: “Never Choose the Uniquely Largest, a Characterization” *Economic Theory*, 8, (1996), pp. 239-249.
- [4] Bernheim D. and A. Rangel: “Beyond Revealed Preference: Choice Theoretic Foundations for Behavioral Welfare Economics”, NBER, working paper 13737, 2008.
- [5] Eliaz, K. and E. Ok: “Indifference or Indecisiveness? Choice-Theoretic Foundations of incomplete Preferences”, *Games and Economic Behavior*, 56, (2006), pp. 61-86.
- [6] Ergin, H.: “Consistency in House Allocation Problems”, *Journal of Mathematical Economics*, 34, (2000), pp. 77-97.
- [7] Gaertner, W. and Y. Xu: “On Rationalizability of Choice Functions: A Characterization of the Median”, *Social Choice and Welfare*, 16, (1999), pp. 629-638.
- [8] Houy, N.: “Rationality and Order Dependent Sequential Rationality”, *Theory and Decision*, 62, (2007), pp. 119-134.
- [9] Hylland, A. and R. Zeckhauser: “The Efficient Allocation of Individuals to Positions”, *Journal of Political Economy*, 87, (1979), pp. 293-314.

- [10] Kalai, G. A. Rubinstein and R. Spiegler: “Rationalizing Choice Functions by Multiple Rationales”, *Econometrica*, 70, (2002), pp. 2481-2488.
- [11] Ma, J.: “Strategyproofness and the Strict Core in a Market with Indivisibilities”, *International Journal of Game Theory*, 23, (1995), pp. 75-83.
- [12] Manzini, P. and M. Mariotti: “Sequentially Rationalizable Choice”, *American Economic Review*, 97, (2007), pp. 1824-1839.
- [13] Ok, E.: “Real Analysis with Economic Applications”, Princeton University Press, Princeton, (2007).
- [14] Papai, S.: “Strategyproof Assignment by Hierarchical Exchange”, *Econometrica*, 68, (2000), pp. 1403-1433.
- [15] Roth, A. and A. Postlewaite: “Weak versus Strong Domination in a Market with Indivisible Goods”, *Journal of Mathematical Economics*, 4, (1977), pp. 131-137.
- [16] Roth, A.: “Incentive Compatibility in a Market with Indivisibilities” *Economics Letters*, 9, (1982), pp. 127-132.
- [17] Shapley, L. and H. Scarf: “On Cores and Indivisibility”, *Journal of Mathematical Economics*, 1, (1974), pp. 23-37.
- [18] Tadenuma, K.: “Egalitarian-Equivalence and the Pareto Principle for Social Preferences”, *Social Choice and Welfare*, 24, (2005), pp. 455-473.
- [19] Tadenuma, K. and N. Houy: “Lexicographic Composition of Multiple Criteria for Decision Making”, *mimeo*, Hitotsubashi University no 2007-13, 2007.
- [20] Thomson, W.: “Consistency and its Converse, an Introduction”, Rochester Center for Economic Research, Working paper no.448, 1998.
- [21] Xu, Y. and L. Zhou: “Rationalizability of Choice Functions by Game Trees”, *Journal of Economic Theory*, 134, (2007), pp. 548-556.