Interim Self-Stable Decision Rules

Daeyoung Jeong* Semin Kim†

June 16, 2017

Abstract

This study identifies a set of interim self-stable decision rules. In our model, individual voters encounter two separate decisions sequentially: (1) a decision on the change of a voting rule they are going to use later and (2) a decision on the final voting outcome under the voting rule which has been decided from the prior procedure. A given decision rule is self-stable if any other possible rule does not get enough votes to replace the given rule under the given rule itself. We fully characterize the set of interim self-stable decision rules among weighted majority rules with given weights.

JEL Classification: C72, D02, D72, D82.

Keywords: Weighted majority rules, decision rules, self-stability

*Economic Research Institute, The Bank of Korea; daeyoung.jeong@gmail.com
†School of Economics, Yonsei University, South Korea; seminkim@yonsei.ac.kr

Contents

1 Introduction

Since different voting rules may result in different voting outcomes, the welfare of individuals does not just depend on individuals' preferences over possible voting outcomes, but also on a voting rule itself. Hence, each individual voter's preference over different voting rules would be induced by the potential differences in the decision outcomes. It is natural that a group of individuals may want to change the voting rule according to their interests. When the change on voting rule is possible in a certain way, some rules may survive longer than others, but not all rules would stably survive for a long time.

This study identifies a set of interim self-stable decision rules. In our model, individual voters encounter two separate decisions sequentially: (1) a decision on the change of a voting rule they are going to use later and (2) a decision on the final voting outcome under the voting rule which has been decided from the prior procedure. A given decision rule is self stable if any other possible rule does not get enough votes to replace the given rule under the given rule itself. Unlike the previous studies (Barberà and Jackson, 2004; Azrieli and Kim, 2016) on self-stable decision rule, which assume that the decision for changing the decision rule takes place before the individuals' preferences over possible outcomes have been realized, we assume that individuals' preferences have been realized even before making a decision about changing the decision rule.[1]

On many occasions, even before voters decide on a voting rule, they have been aware of their own preferences over possible voting outcomes. For example, a legislature may have a chance to change their decision rule even after they have been aware of the characteristics of the legislation proposed beforehand or that of an upcoming proposal they are going to vote on eventually. In this situation, voters may make decisions on changing the voting rule strategically based on their preferences in the proposal they are going to vote on.

The primary purpose of this study is to characterize the set of interim self-stable decision rules and compare it with the set of ex-ante self stable decision rules, which has

[1] In this sense, our study is an obvious generalization of Holmström and Myerson (1983), which assumes individuals with realized preferences over possible voting outcomes make decisions on changing the decision rule not under the given decision rule, but only under unanimity.

been identified by previous studies. By comparing two different self-stability concepts, we hope to investigate the welfare effect of the timing of the decision on changing voting rules, and provide policy implications of it. We first start with the simplest possible case by restricting our attention to the set of qualified majority rules where all voters have the same voting powers.[2] We then move on to a set of weighted majority rules where individual voters may have different voting powers. In either case, we show that some voting rules which are not ex-ante self-stable could be interim self-stable. We would like to generalize this result and show that if a weighted majority rule is ex ante self-stable, then it is interim self-stable.

We also intend to generalize our analysis further by considering a framework with a constitution: a constitution is a pair of voting rules, one of which is for the decision on changing the constitution and the other is for the decision on the final outcome. This generalized framework will allow us to study more realistic situations for any forms of legislatures, and compare our results directly with the results of Barberà and Jackson (2004).

1.1 Related Literature

The two papers, Barberà and Jackson (2004) and Holmström and Myerson (1983) motivate this project. Barberà and Jackson (2004) introduce the ex-ante self-stability of voting rules and focus on the qualified majority rules. Unlike them, we define the interim self-stability of voting rules and study not only the qualified majority rules but also general voting rules. The interim self-stability is similar to the durability of decision rules defined by Holmström and Myerson (1983) in that an agent utilizes the preferences information in the interim stage. While they use the unanimous rule to choose between rules, we start with the given rule itself and try to extend the argument with the various rules. It can show the effects of those variations on the set of stable rules.

In our model, agents' preferences over voting rules are endogenously determined from their assessments regarding their preferences over alternatives. Such a model was first

[2]Typical examples of qualified majority rules are unanimity and simple majority. Any super or sub majority rule is also a qualified majority rule.

suggested in early papers by Rae (1969), Badger (1972), and Curtis (1972). While these papers only consider anonymous voting rules with the same weight to all agents, we study weighted majority rules which allow the heterogenous weights for agents.

The seminal book of Neumann and Morgenstern (1953, Section 5) theoretically investigates weighted majority rules. The main interest of the book is the measures of the voting power of agents under the rule. A common scenario leading to heterogeneous voting weights is that of a representative democracy with heterogenous district sizes. An early paper on this topic is Penrose (1946). Recently, Barberà and Jackson (2006) and Fleurbaey (2008) point out the advantage of weighted majority rules from a utilitarian point of view. Also, Azrieli and Kim (2014) show that, in a standard mechanism design setup, weighted majority rules naturally arise from considerations of efficiency and incentive compatibility. We investigate another property, the stability of weighted majority rules.

The idea that the same voting rule used to choose between alternatives is also used to choose between voting rules can be found in the social choice literature. Koray (2000) introduces the concept of self-selection for social choice functions. See also Barberà and Beviá (2002) and Semih Koray (2008).

2 Environment

A society faces a binary decision whether to implement the Reform (R) or to keep the Status-quo (S), so the set of alternatives is $A = \{R, S\}$. The set of agents in the society is $N = \{1, 2, ..., n\}$ with $n \geq 2$. Each agent can either prefer R or S, which indicates the type of the agent, $t_i \in T_i = \{r, s\}$. The probability of agent i being a type t_i is $p_i(t_i)$ and $p_i(t_i = r) + p_i(t_i = s) = 1$. We assume that there is no agent who is indifferent between R and S and that $p_i(t_i) > 0$ for any $t_i \in T_i$. Let $T = T_1 \times \cdots \times T_n$ be the set of type profiles. We assume that types are independent across agents, so we denote $P(t) = \prod_{i \in N} p_i(t_i)$ for the probability of a type profile $t \in T$. For the technical convenience, we abuse the notation, $P(t_{-i}) = \frac{p(t)}{p_i(t_i)}$ for the probability of a type profile of other agents

excluding agent i.

The utility of each agent depends on the chosen alternative and on his own type, $u_i : A \times T_i \to \mathbb{R}$. We normalize the utiliy such that $u(R,r) = a$, $u(R,s) = -1$, and $u(S,r) = u(S,s) = 0$. Thus a society can be characterized by the pair (p_r, a) ,where $p_r = (p_1(r), ..., p_n(r))$. Since randomization over alternatives will be considered, we need to extend each $u_i(\cdot, t_i)$ to $\triangle(A)$. Simply we identify $\triangle(A)$ with the probabilty $x \in [0,1]$ which corresponds to the probability that R is chosen. That is, $u_i(x, t_i) = xu_i(R, t_i) + (1-x)u_i(S, t_i)$.

A voting rule is any mapping $f : T \to [0,1]$, with the interpretation that, $f(t)$ is the probability that R is chosen when the type profile of agents is t. We mainly focus on weighted majority rules.

Definition 1. The voting rule f is a *Weighted Majority Rule* if there are non-negative weights $w = (w_1, ..., w_n)$ and a quota $0 \leq q < \sum_{i \in N} w_i$ such that

$$f(t) = \begin{cases} 1 & if \quad \sum_{\{i:t_i=r\}} w_i > q \\ 0 & if \quad \sum_{\{i:t_i=r\}} w_i \leq q. \end{cases}$$

We write $f = (w, q)$ if f can be represented by these weights and quota.

3 Interim Self-stability

Roughly, we would like to say a weighted majority rule f is interim self-stable, if, for any alternative voting rule g, there is a Nash equilibrium of a voting game in which the alternative g is defeated by f for any type profile $t \in T$ if the decision is made by the incumbent rule f itself.[3] Hence, in order to define this concept formally, we need to define the two stage voting game.

Timing of the game is as follows. In the first stage, agents observe their own type t_i. Then under the incumbent rule $f = (w, q)$, agents play a simultaneous voting game

[3]In a general definition, g can be any indirect mechanism of $g : A_1 \times \cdots \times A_n \to [0,1]$, where each A_i is a nonempty finite set.

whether to keep the incumbent rule f or to choose the alternative rule g. The alternative rule g would be implemented if $\sum w_i > q$, where the sum is taken over all agents who vote for g, and f would be maintained otherwise. In the second stage, agents make a decision on $A = \{R, S\}$ by the rule chosen in the first stage. Let $\sigma_i(t_i)$ be the probability that individual i would vote for g in the first stage when her type is t_i.

To reject the alternative rule g all the time, the probability that g gets sufficient weighted votes should be zero for all $t \in T$. In other words, the alternative g is always rejected if and only if

$$\sum_{\{j:\sigma_j(t_j)>0\}} w_j \leq q, \quad \forall t \in T. \tag{3.1}$$

If Equation (3.1) holds, then honest behavior in f and g (we consider incentive compatible f and g) , together with the voting strategies in the first stage, $\sigma = (\sigma_1, \ldots, \sigma_n)$ form a Nash equilibrium if and only if

$$\sum_{t_{-i}} P\left(t_{-i}\right) \gamma_i(t_{-i}) \left(u_i(f(t), t_i) - u_i(g(t), t_i)\right) \geq 0 \quad \forall i, \ \ \forall t_i \in T_i, \tag{3.2}$$

where

$$\Phi_i = \{H_i \subseteq N \setminus \{i\} |\, q - w_i < \sum_{j \in H_i} w_j \leq q\},$$

and

$$\gamma_i(t_{-i}) = \sum_{H_i \in \Phi_i} \left(\prod_{j \in H_i} \sigma_j(t_j) \right) \left(\prod_{j \in N/(H_i \cup \{i\})} (1 - \sigma_j(t_j)) \right).$$

Denote by Φ_i the collection of the set of other agents $H_i \subseteq N \setminus \{i\}$ such that the agent i is pivotal if the agents in H_i vote for the alternative rule g and all others $j \notin H_i$ vote for the given rule f.[4] We can interpret that $\gamma_i(t_{-i})$ is the probability of pivotal event for agent i at t_{-i} given σ. Equation (3.2) characterizes the condition which guarantees

[4]Also denote Ψ^f the set of minimal winning coalitions under a decision rule f. Note that, when f is a qualified majority rule, $\Phi_i = \{C \setminus \{i\} : C \in \Psi^f \text{ and } i \in C\}$.

the rejection of g in a Nash equilibrium: either agent i with t_i is never pivotal, or she is weakly better off under f than g given that she is pivotal.

However, in a simultaneous voting game, there generally exists a trivial Nash equilibrium in which $\sigma_i(t_i) = 0$ for all i and t_i, unless one individual has the dictatorial power in f. In such an equilibrium, where the condition (3.1) and (3.2) are satisfied, g is always defeated by f. If we do not exclude such trivial Nash equilibria properly, all weighted majority rules would be interim self-stable. Therefore, in order to define a reasonable concept of interim self-stability, we need to refine the equilibria of the game Γ further. Holmström and Myerson (1983) refine equilibria by requiring that if, conditional on an agent is pivotal, the agent would get higher expected utility under the alternative rule than under the current rule, then she must vote for the alternative rule. Similarly, we require a type of sequential rationality for an agent voting strategy σ_i given that she is pivotal.

We first characterize a posterior distribution given that an agent i is pivotal as follows.[5]

$$\mu_i(t_{-i}) = \lim_{k\to\infty} \frac{P(t_{-i}) \sum_{H_i \in \Phi_i} \left(\prod_{j\in H_i} \sigma_j^k(t_j)\right) \left(\prod_{j\in N/(H_i\cup\{i\})} (1-\sigma_j^k(t_j))\right)}{\sum_{\hat{t}_{-i}\in T_{-i}} P(\hat{t}_{-i}) \sum_{H_i \in \Phi_i} \left(\prod_{j\in H_i} \sigma_j^k(\hat{t}_j)\right) \left(\prod_{j\in N/(H_i\cup\{i\})} (1-\sigma_j^k(\hat{t}_j))\right)} \tag{3.3}$$

$\forall i, \forall t_i \in T_i, \forall t_{-i} \in T_{-i},$

where

$$\sigma_j^k(t_j) > 0 \quad \forall k, \forall j, \forall t_j \in T_j$$
$$\sigma_j(t_j) = \lim_{k\to\infty} \sigma_j^k(t_j) \quad \forall j, \forall t_j \in T_j$$

Since the denominator of the equation (3.3) could be zero, we characterize the distribution

[5]The posterior distribution $\mu_i(t_{-i})$ is not exactly the posterior beliefs in a concept of sequential equilibrium. However, an agent i's decision on the changing rules is only relevant when she is pivotal. Hence, we characterize an agent's posterior distribution given that she is pivotal, and then require a rational behavior at the first stage given the posterior distribution.

in the style of the trembling hand model. Given this distribution or belief, we require that, for any type t_i of any individual i,

$$\text{if } \sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t), t_i) < \sum_{t_{-i}} \mu_i(t_{-i}) u_i(g(t), t_i), \tag{3.4}$$

$$\text{then } \sigma_i(t_i) = 1.$$

This condition imposes that, conditional on that the agent i is pivotal, if an agent i with type t_i is expected to be better off under the alternative rule g than under the rule f, then she should vote for g.

Now, we have all the conditions to construct a reasonable equilibrium concept to compare a pair of decision rules f and g. We define the concepts of equilibrium rejection and endurance for the comparison.

Definition 2 (Equilibrium rejection)**.**
Consider a weighted majority rule f. A strategy profile and a belief (σ, μ) consists an equilibrium rejection of g if and only if the conditions (3.1) through (3.4) are all satisfied.

Definition 3 (Endurance)**.**
Consider a weighted majority rule f. f endures g if and only if there exists some equilibrium rejection of g.

Now, we formally define the interim self-stability of a weighted majority rule.

Definition 4 (Interim Self-stability)**.**
Consider a weighted majority rule f. f is interim self-stable if and only if f endures every alternative rule $g : T \to [0, 1]$.

4 Anonymity constraint

In this section, we focus on anonymous weighted majority rules which are called qualified majority rules similarly to Barberà and Jackson (2004). The current and alternative rules are qualified majority rules which can be represented by the special type of weighted

majority rules where $w_i = 1$ for all $i \in N$ and $q \in \{0, 1, ..., n-1\}$. In the literature, they are classified according to the quota: a simple majority rule ($q^s = \frac{n}{2}$ if n is even and $q^s = \frac{n-1}{2}$ if n is odd), a sub majority rule ($q < q^s$), and a super majority rule ($q > q^s$).

Proposition 1 (Interim Self Stable Qualified Majority Rules)**.**

A qualified majority rule f is interim self-stable among qualified majority rules if and only if it is a simple or super majority rule.

Proof of Proposition 1.

(Only if part)

Assume that the current rule f is sub majority rule with the quota q and that it is interim self-stable. Consider the unanimous rule as the alternative rule g with $\frac{n-1}{n} \leq q$. By the assumption, there exists an equilibrium rejection of g, (σ, μ). Fix agent i with $\bar{t}_i = s$. Define $\bar{T}_{-i} \equiv \{t_{-i} \in T_{-i} : f(t_{-i}, \bar{t}_i) = R\}$. In the equilibrium rejection (σ, μ), for the agent i the left hand side of Equation (3.4) is

$\sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t_{-i}, \bar{t}_i), \bar{t}_i) = -\sum_{t_{-i} \in \bar{T}_{-i}} \mu_i(t_{-i})$ and the right hand side of Equation (3.4) is zero.

We claim that $\sum_{t_{-i} \in \bar{T}_{-i}} \mu_i(t_{-i}) > 0$ in any equilibrium rejection. Note that under qualified majority rule, at most q agents vote for g with a positive probability for any $t \in T$ in any equilibrium rejection. In other words, $n-q$ agents never vote for g. There are two cases regarding the probability of agent i being pivotal for any equilibrium rejection. First there exists a $t_{-i} \in T_{-i}$ such that $\gamma(t_{-i}) > 0$. It implies that exactly q agents vote for g with a positive probability at t_{-i}. Fix these agents and we can find a $\bar{t}_{-i} \in \bar{T}_{-i}$ such that $\gamma(\bar{t}_{-i}) > 0$ since the number of other agents is $n - q - 1 > q$ and they decide $f(\bar{t}) = R$ by themselves. Then by Bayes theorem, $\sum_{t_{-i} \in \bar{T}_{-i}} \mu_i(t_{-i}) > 0$. Second for any $t_{-i} \in T_{-i}$, $\gamma(t_{-i}) = 0$. We can find a $\tilde{t}_{-i} \in T_{-i}$ such that $\mu_i(\tilde{t}_{-i}) > 0$. With the similar trick of the previous case, fix agents in H_i at $\tilde{t}_{-i}$ and we can find a $\bar{t}_{-i} \in \bar{T}_{-i}$ such that $\lim_{k \to \infty} \frac{(\prod_{j \in H_i} \sigma_j(\bar{t}_j))(\prod_{j \in N/(H_i \cup \{i\})} (1 - \sigma_j(\bar{t}_j)))}{(\prod_{j \in H_i} \sigma_j(\tilde{t}_j))(\prod_{j \in N/(H_i \cup \{i\})} (1 - \sigma_j(\tilde{t}_j)))} = \frac{(\prod_{j \in H_i} \sigma_j(\bar{t}_j))}{(\prod_{j \in H_i} \sigma_j(\tilde{t}_j))} = 1$. Then, $\mu_i(\bar{t}_{-i}) > 0$ which proves the claim. By the claim, the condition (3.4) implies that $\sigma_i(\bar{t}_i) = 1$. The argument is valid for any agent i with $t_i = s$. However, at the type profile $\bar{t}$ with $| \{i : \bar{t}_i = s\} | > q$, this equilibrium rejection contradicts (3.1).

(If part)

We only show that the simple majority rule is interim self-stable because the proof for super majority rules is almost the same. Among alternative rules, we consider the two extreme qualified majority rules, $q = n-1$ and 0. When the alternative rule is the unanimous rule, i.e., $q = n-1$, consider the strategy profile and a belief (σ, μ) such that $\sigma_i(t_i) = 0$ for $\forall t_i \in T_i$, $\sigma_i^k(s) = \frac{1}{k}$, and $\sigma_i^k(r) = \frac{1}{k^2}$ for $\forall i \in N$. This trivial strategy profile simply satisfies the conditions (3.1) and (3.2). By (3.3), we can derive $\mu_i(t_{-i}) > 0$ for t_{-i} such that $\mid \{i : t_i = s\} \mid = q^s$ and $\mu_i(t_{-i}) = 0$ otherwise. The right hand side of the equation in (3.4) is weakly less than the left for any type and any agent. Thus, the pair (σ, μ) is an equilibrium rejection of g. When the alternative rule is the other extreme case of $q = 0$, we can similarly find an equilibrium rejection (σ, μ) such that $\sigma_i(t_i) = 0$ for $\forall t_i \in T_i$, $\sigma_i^k(s) = \frac{1}{k^2}$, and $\sigma_i^k(r) = \frac{1}{k}$ for $\forall i \in N$. For any intermediate qualified majority rule with $0 < q < n-1$, the same argument is valid, which proves that the simple majority rule is interim self-stable. □

5 Non-anonymity constraint

We consider non-anonymous rules. A current rule f is a weighted majority rule and an alternative rule g can be any voting rule.

Denote $\Psi^{\hat{f}}$ the set of minimal winning coalitions(MWCs) under a decision rule $\hat{f}$.

For the technical convenience, we define, for a decision rule $\hat{f}$ and a type t_i, $T_{t_i}^{\hat{f}} \equiv \left\{ t_{-i} : \hat{f}(t_i, t_{-i}) = R \right\}$. We also define, for a set of type profile $\tilde{T} \subseteq T_{-i}$, $\mu_i(\tilde{T}) \equiv \sum_{t_{-i} \in \tilde{T}} \mu_i(t_{-i})$.

Lemma 1 (Necessary Condition: No Veto Group)**.**

Consider a given rule f. If any minimal winning coalition C has a mutually exclusive minimal winning coalition $C' \in \Psi^f$ such that $C \cap C' = \emptyset$, f is not interim self stable.

Proof of Lemma 1.

Consider a unanimous rule g. And suppose there exists an equilibrium rejection of g, (σ, μ). We only consider a case with $\gamma_i(t_i) = 0$ since a case with $\gamma_i(t_i) > 0$ is only easier

to prove. We know there exists a minimal winning coalition $\bar{C} \in \Psi^f$ such that for all $i \in \bar{C}$, $w_i \geq w_j$ for any $j \in N \setminus \bar{C}$.

Fix i such that $w_i \geq w_j$ for any $j \in N$. Consider $t_i = s$.

We first argue that if, for a type profile $\tilde{t}_{-i} \in T_{-i}$ and some set of agents $\tilde{H} \in \Phi_i$,

$$\lim_{k\to\infty} \left(\prod_{j\in\tilde{H}} \sigma_j^k(\tilde{t}_j) \right) \left(\prod_{j\in N\setminus(\tilde{H}\cup\{i\})} (1 - \sigma_j^k(\tilde{t}_j)) \right) \tag{5.1}$$

goes to zero in a speed that is no faster than for any other $H \in \Phi_i$ and type profile $t_{-i} \in T_{-i}$, then all $j \in N \setminus (\tilde{H} \cup \{i\})$ should have $\sigma_j(s) = \sigma_j(r) = 0$. Suppose not. So there is some agent $j \in N \setminus (\tilde{H} \cup \{i\})$ with $\sigma_j(s) + \sigma_j(r) > 0$. Define $\hat{N}^+ \equiv \{j \in N \setminus (\tilde{H} \cup \{i\}) | \sigma_j(s) + \sigma_j(r) > 0\}$. Also define $\tilde{N}^+ \equiv \{j \in \tilde{H} | \sigma_j(\tilde{t}_j) > 0\}$. We know $w(\hat{N}^+ \cup \tilde{N}^+ \cup \{i\}) \leq q$ and $w(\tilde{H} \cup \{i\}) > q$, which imply that there exists some subset $\eta \subseteq (\tilde{H} \cup \{i\}) \setminus \tilde{N}^+$ such that $w(\hat{N}^+ \cup \tilde{N}^+ \cup \eta) > q$ and $w((\hat{N}^+ \cup \tilde{N}^+ \cup \eta) \setminus \{j\}) \leq q$ for any $j \in \eta$. Since $w_i \geq w_j$ for any $j \in \eta$, there should be some $H' \in \Phi_i$ such that $(\hat{N}^+ \cup \tilde{N}^+) \subset H'$, $H' \subset (\hat{N}^+ \cup \tilde{H})$, and H' with some $\tilde{t}'_{-i}$ makes Equation (5.1) goes to zero in a speed that is slower than for $\tilde{H}$ with some $\tilde{t}_{-i}$. Contradiction. So, all $j \in N \setminus (\tilde{H} \cup \{i\})$ should have $\sigma_j(s) = \sigma_j(r) = 0$.

Second we argue that $\sigma_i(s) = 1$. By construction, there exists a minimal winning coalition C' such that 1) $i \notin C'$, 2) for all $j \in C' \setminus (\tilde{H} \cup \{i\})$ we have $\sigma_j(s) = \sigma_j(r) = 0$, and 3) for all $j \in C' \cap (\tilde{H} \cup \{i\})$ we have $\sigma_j(r) + \sigma_j(s) > 0$. Hence, $\mu_i(t'_{-i}) > 0$ for some t'_{-i} such that $f(t_i = s, t'_{-i}) = R$. Then, from Condition (3.4), we should have $\sigma_i(s) = 1$, since $\sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t), t_i) < \sum_{t_{-i}} \mu_i(t_{-i}) u_i(g(t), t_i) = 0$.

A similar argument can be applied for all $i \in \bar{C}$. So, for all $i \in \bar{C}$, $\sigma_i(s) = 1$, which contradict the fact that σ is an equilibrium rejection of g.

Therefore, there exists no equilibrium rejection of the unanimous rule g. □

Lemma 2 (Sufficient Condition 1: Veto Agent)**.**

If, under the existing rule f, there exists a veto agent such that $i \in C$ for all $C \in \Psi^f$, f is interim self-stable.

Proof of Lemma 2.

Consider a strategy profile and a belief system (σ, μ) such that, for all $j \in N$, $\sigma_{j\neq i}(r) = 1$, $\sigma_{j\neq i}(s) = 0$, $\sigma_i(s) = \sigma_i(r) = 0$ and $\sigma^k_{j\neq i}(s) > \sigma^k_i(s) > \sigma^k_i(r)$.

For the veto agent i, if $t_i = s$, $\sigma_i(s) = 0$ is justified since $f(t_i, t_{-i}) = S$ for any t_{-i}. If $t_i = r$, $\sigma_i(r) = 0$ is justified since i is pivotal only when the right enough number of other agents with type r vote for g, so only when $f(t_i, t_{-i}) = R$.

For all other agents $j \neq i$, $\sigma_j(s) = 0$ is justified since $\gamma_j(t_j) = 0$ for all t_j and $\mu_j(t_j, t_{-j})$ is positive only when $t_i = s$, so $f(t_j, t_{-j}) = S$. Also, $\sigma_j(r) = 1$ obviously satisfies Equation (3.2) since $\gamma_j(t_j) = 0$ for all t_j and does not violate Equation (3.4). □

Denote $G(w)$ the set of weighted majority rules with $w = (w_i)_{i=1}^n$.

Lemma 3 (Sufficient Condition in a Fixed-weights Environment)**.**

A weighted majority rule $f \in G(w)$ is interim self-stable among $G(W)$ if there exists a minimal winning coalition C which is not mutually exclusive with any other minimal winning coalition.

Proof of Lemma 3.

Suppose a minimal winning coalition C is not mutually exclusive with any other minimal winning coalition. So, $w(N \setminus C) \leq q$. For the convenience of notation, $w_i \geq w_j$ if and only if $i \geq j$. Find $\underline{i}$ such that $H_{\underline{i}} \equiv \{i \in N : i < \underline{i}\} \in \Psi_{\underline{i}}$ and $|C \setminus H_{\underline{i}}| = 1$. By construction, $H_{\underline{i}} \in \Psi_i$ for any $i \geq \underline{i}$. Also, we know either $w(N \setminus H_{\underline{i}}) \leq q$ or $(N \setminus H_{\underline{i}}) \in \Phi_f$.

Since we are focusing on the case where $w_f = w_g$ for any g, it is either $q_f > q_g$ or $q_f < q_g$.

First, consider the case where $q_f > q_g$. So, if $g(t) = S$, then $f(t) = S$. And if $f(t) = R$, then $g(t) = R$. Suppose a strategy profile (σ, μ) such that $\sigma_i(s) = 0$ for any i and $\sigma_j(r) = 0$ for any $j \notin H_{\underline{i}}$ and $\sigma_j(r) = 1$ for $j \in H_{\underline{i}}$, and any arbitrary $\sigma^k_i(t_i)$ which converges to $\sigma_i(t_i)$ for any i and t_i. $\sigma_i(s)$ is always justified, since $g(t) = S$ implies $f(t) = S$. For any $i \notin H_{\underline{i}}$, $\gamma_i(t_{-i}) > 0$ only when $t_j = r$ for all $j \in H_{\underline{i}}$. Then, for $t_i = r$, Condition (3.2) is satisfied, and so $\sigma_i(r) = 0$ is justified. For any $i \in H_{\underline{i}}$, $\gamma_i(t_{-i})$ is always zero, $\sigma_j(r) = 1$ it is okay.

Second, consider the case where $q_f < q_g$. So, if $f(t) = S$, then $g(t) = S$. And if $g(t) = R$, then $f(t) = R$. Suppose a strategy profile (σ, μ) such that $\sigma_i(r) = 0$ for any

i and $\sigma_j(s) = 0$ for any $j \notin H_{\underline{i}}$ and $\sigma_j(s) = 1$ for $j \in H_{\underline{i}}$, and any arbitrary $\sigma_i^k(t_i)$ which converges to $\sigma_i(t_i)$ for any i and t_i. $\sigma_i(r)$ is always justified, since $g(t) = R$ implies $f(t) = R$. For any $i \notin H_{\underline{i}}$, $\gamma_i(t_{-i}) > 0$ only when $t_j = s$ for all $j \in H_{\underline{i}}$. Then, for $t_i = s$, Condition (3.2) is satisfied, and so $\sigma_i(s) = 0$ is justified. For any $i \in H_{\underline{i}}$, $\gamma_i(t_{-i})$ is always zero, so $\sigma_j(s) = 1$ is okay. □

6 Discussion

In this paper, we have identified a set of interim self-stable decision rules. In contrast to the previous studies, which assume that the decision for changing the decision rule takes place before the individuals' preferences over possible outcomes have been realized, we assume that individuals evaluate decision rules after their preferences have been realized. Among anonymous weighted majority rules which are called qualified majority rules, a decision rule is interim self-stable if and only if it is a simple or super majority rule. We then move on to a set of weighted majority rules where individual voters may have different voting powers. We so far have shown that, if a decision rule is interim self-stable, there exists a minimal winning coalition which is not mutually exclusive with any other minimal winning coalition. Moreover, in the set of weighted majority rules with a fixed weights, the previous condition is necessary and sufficient condition of interim self-stability.

We also intend to generalize our analysis further by considering a constitution. Contrary to a simple decision rule with only one voting rule, a constitution consists of a pair of voting rules, one of which is for the decision on changing the constitution and the other is for the decision on the final outcome. In fact, on many occasions, society uses different voting rules for changing decision rules and for final decisions. Hence, it is important to generalize our analysis further by considering a framework with a constitution, as in Barberà and Jackson (2004).

References

Azrieli, Y. and Kim, S. (2014). Pareto efficiency and weighted majority rules. *International Economic Review*, 55(4):1067–1088.

Azrieli, Y. and Kim, S. (2016). On the self-(in)stability of weighted majority rules. *Games and Economic Behavior*, 100:376 – 389.

Badger, W. (1972). Political individualism, positional preferences, and optimal decision rules. In Niemi, R. G. and Weisberg, H. F., editors, *Probability Models of Collective Decision Making*. Merrill, Columbus, Ohio.

Barberà, S. and Beviá, C. (2002). Self-selection consistent functions. *Journal of Economic Theory*, 105(2):263 – 277.

Barberà, S. and Jackson, M. (2006). On the weights of nations: Assigning voting weights in a heterogeneous union. *Journal of Political Economy*, 114(2):317–339.

Barberà, S. and Jackson, M. O. (2004). Choosing how to choose: Self-stable majority rules and constitutions. *The Quarterly Journal of Economics*, 119(3):1011–1048.

Curtis, R. (1972). Political individualism, positional preferences, and optimal decision rules. In Niemi, R. G. and Weisberg, H. F., editors, *Probability Models of Collective Decision Making*. Merrill, Columbus, Ohio.

Fleurbaey, M. (2008). Weighted majority and democratic theory. *Mimeo*.

Holmström, B. and Myerson, R. B. (1983). Efficient and durable decision rules with incomplete information. *Econometrica*, 51(6):pp. 1799–1819.

Koray, S. (2000). Self-selective social choice functions verify arrow and gibbard- satterthwaite theorems. *Econometrica*, 68(4):981–995.

Neumann, J. V. and Morgenstern, O. (1953). *Theory of games and economic behavior,*. Princeton University Press, Princeton.

Penrose, L. S. (1946). The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57.

Rae, D. W. (1969). Decision-rules and individual values in constitutional choice. *American Political Science Review*, 63:40–56.

Semih Koray, A. S. (2008). Self-selective social choice functions. *Social Choice and Welfare*, 31(1):129–149.

A Appendix

A.1 Consistency of Definition

Here, we discuss the consistency of our definition of interim self stability with the ex-ante self-stability à la Azrieli and Kim (2016) and the durability à la Holmström and Myerson (1983).

Consider the "ex-ante environment" studied in Azrieli and Kim (2016), where agents vote on rule change before their types are realized. We rewrite our conditions and definition as follows.

To reject the alternative rule g all the time, the probability that g gets sufficient weighted votes should be zero. In other words, the alternative g is always rejected if and only if

$$\sum_{\{j:\sigma_j>0\}} w_j \leq q. \tag{A.1}$$

If Equation (A.1) holds, then honest behavior in f and g (we consider incentive compatible f and g), together with the voting strategies in the first stage, $\sigma = (\sigma_1, \ldots, \sigma_n)$ form a Nash equilibrium if and only if

$$\gamma_i \left(u_i(f) - u_i(g)\right) \geq 0 \quad \forall i, \tag{A.2}$$

where

$$u_i(f) = a \sum_{\{t\in T: t_i=r\}} p(t)f(t) - \sum_{\{t\in T: t_i=s\}} p(t)f(t),$$

$$\Phi_i = \{H_i \subseteq N/\{i\} \,|\, q - w_i < \sum_{j\in H_i} w_j \leq q\},$$

and

$$\gamma_i = \sum_{H_i\in\Phi_i} \left(\prod_{j\in H_i} \sigma_j\right) \left(\prod_{j\in N/(H_i\cup\{i\})} (1-\sigma_j)\right).$$

We require that, for any individual i,

$$\text{if } \; u_i(f) < u_i(g), \text{then } \; \sigma_i = 1. \tag{A.3}$$

This condition imposes that, if the expected utility of individual i in the alternative decision rule g would be higher than in the current rule f, then individual i should vote for g.[6]

$w(Y) := \sum_{i\in Y} w_i$ denotes the total weight of coalition Y.

Proposition 2. *For a given weighted majority rule f, $w(\{i : u_i(f) < u_i(g)\}) \leq q$ for any alternative rule g if and only if there exists a strategy profile σ that satisfies conditions (A.1), (A.2),*

[6] In the second stage, since f and g are incentive compatible, we simply assume that all individuals report their true types.

and (A.3).

Proof of Proposition 2.

(Only if part)

Suppose $w(\{i : u_i(f) < u_i(g)\}) \leq q$. Then, set $\sigma_i = 1$ for any $i \in \{i : u_i(f) < u_i(g)\}$ and $\sigma_i = 0$ for any $i \notin \{i : u_i(f) < u_i(g)\}$. The condition (A.1) and (A.3) are satisfied. For an individual i with $\sigma_i = 1$, $\gamma_i = 0$. For an individual i with $\sigma_i = 0$, $(u_i(f) - u_i(g)) \geq 0$ by construction. Therefore, the condition (A.2) is satisfied.

(If part)

Suppose not. That is, the conditions (A.1), (A.2), and (A.3) are all satisfied, but

$$w(\{i : u_i(f) < u_i(g)\}) > q.$$

Since we suppose the condition (A.3) is satisfied, $\{j : u_j(f) < u_j(g)\} \subseteq \{j : \sigma_j > 0\}$, which implies $w(\{i : u_i(f) < u_i(g)\}) \leq w(\{j : \sigma_j > 0\})$. Then, the condition (A.1) is violated, since $\sum_{\{j:\sigma_j>0\}} w_j \geq w(\{i : u_i(f) < u_i(g)\}) > q$. Contradiction. □

One may wonder why we don't use the simple condition as $w(\{i : u_i(f) < u_i(g)\}) \leq q$ in Azrieli and Kim (2016) to define interim self-stability. To do that, in our setting, we need to add up the weights of agents i's who have

$$\sum_{t_{-i}} \mu_i(t_{-i}) u_i(f(t), t_i) < \sum_{t_{-i}} \mu_i(t_{-i}) u_i(g(t), t_i),$$

which is a part of the condition (3.4). But as in the condition (3.3), the posterior belief μ_i can only be calculated with a strategy profile for the first stage voting game σ. That is, to define interim self-stability in a way analogous to Azrieli and Kim (2016), we need a complete characterization of a Nash equilibrium with a sequentially rational strategy profile and a consistent belief system.

A.2 Extra Lemmas and Propositions

Lemma 4.

There is a minimal winning coalition $\bar{C} \in \Psi^f$ where for any $i \in \bar{C}$, any minimal winning coalition $C_i \ni i$ has a mutually exclusive minimal winning coalition $C' \in \Psi^f$ such that $C_i \cap C' = \emptyset$ if and only if any minimal winning coalition in Ψ^f has a mutually exclusive minimal winning coalition in Ψ^f.

Proof.

(Only if)

Assume a minimal winning coalition $\bar{C} \in \Psi^f$ where for any $i \in \bar{C}$, any minimal winning coalition $C_i \ni i$ has a mutually exclusive minimal winning coalition $C' \in \Psi^f$ such that $C_i \cap C' =$

$\emptyset$. If there exists a minimal winning coalition $\tilde{C}$ which is not mutually exclusive with any other minimal winning coalition, then $\tilde{C}$ should not contain any $i \in \bar{C}$. So, $\tilde{C} \cap \bar{C} = \emptyset$. Contradiction.

(If)

It is obvious. □

Lemma 5.

If any minimal winning coalition has a mutually exclusive minimal winning coalition, there exists a pair of mutually exclusive minimal winning coalitions C and C' such that $w_i \geq w_{i'}$ for any $i \in C$ and $i' \in C'$.

Proof. There always exists a minimal winning coalition $\bar{C} \in \Psi^f$ such that for all $i \in \bar{C}$, $w_i \geq w_j$ for any $j \in N \setminus \bar{C}$. If the minimal winning coalition $\bar{C}$ has a mutually exclusive minimal winning coalition C', then $w_i \geq w_{i'}$ for any $i \in \bar{C}$ and $i' \in C'$. □

Lemma 6 (Necessary Condition 1: Single Agent Minimal Winning Coalition)**.**

If, under f, there exists an agent i who consists a minimal winning coalition by itself $C = \{i\}$ and a mutually exclusive minimal winning coalition $\tilde{C}$ such that $\tilde{C} \cap C = \emptyset$, f is not interim self stable.

Proof of Lemma 6.

By construction, the agent i is always pivotal, $\gamma(t_{-i}) = 1$ for all t_{-i}.

Consider an alternative rule g such that $w_i^g = w_i^f$ and $w_j^g = 0$ for all $j \neq i$.

Then, for $t_i = s$, Equation (3.2) is violated since

$$\sum_{t_{-i}} P(t_{-i})\, u_i(f(t), t_i) = \sum_{t_{-i} \in T^f_{t_i=s}} P(t_{-i})\, u_i(f(t), t_i) = -\sum_{t_{-i} \in T^f_{t_i=s}} P(t_{-i}) < 0.$$

So, there is no equilibrium rejection of g. □

Lemma 7 (Necessary Condition 2: Small Quota)**.**

If there exists a minimal winning coalition $C \in \Psi^f$ such that for any $i \in C$ and for any minimal winning coalition $C_i \ni i$, $w(N \setminus C_i) > 2q$, f is not interim self-stable.

Proof of Lemma 7.

Let an alternative rule g be the unanimous rule. By construction, for any agent $i \in C$, $T^f_{t_i=s} \neq \emptyset$ and $T^g_{t_i=s} = \emptyset$.

We prove by contradiction. Let's suppose there exists an equilibrium of g, (σ, μ).

For $t_i = s$, suppose $\sigma_i(s) \neq 1$. From Equation (3.4), $-\mu_i(T^f_s) \geq -\mu_i(T^g_s)$. We know $\mu_i(T^g_r) = 0$, we should have $\mu_i(T^f_s) = 0$.

So, if a type profile $\tilde{t}_{-i} \in T_{-i}$ and some set of agents $\tilde{H} \in \Phi_i$,

$$\lim_{k \to \infty} \left(\prod_{j \in \tilde{H}} \sigma_j^k(\tilde{t}_j) \right) \left(\prod_{j \in N \setminus (\tilde{H} \cup \{i\})} (1 - \sigma_j^k(\tilde{t}_j)) \right)$$

goes to zero in a speed that is no faster than for any other $H \in \Phi_i$ and type profile $t_{-i} \in T_{-i}$, $\tilde{t}_{-i}$ should not be in $T^f_{t_i=s}$. It means that, for a minimal winning coalition $C_i \ni i$ which is a subset of $\tilde{H}$, any $j \in \left(\tilde{H} \setminus C_i\right)$ have either $\sigma_j(s) > 0$ or $\sigma_j(r) > 0$. Also, there should not be any minimal winning coalition with all r types in $\tilde{t}_{-i}$. It means that $w(\left\{j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = r\right\}) \leq q$. Then, we should have enough number of agents $j \in N \setminus (\tilde{H} \cup \{i\})$ such that $\tilde{t}_j = s$ and

$$w(\left\{j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = s\right\}) \geq w(N \setminus (\tilde{H} \cup \{i\})) - q.$$

It implies that for such j with $\tilde{t}_j = s$ we should have $\sigma_j(r) = 1$. But, since $w(N \setminus C_i) > 2q$, $w(N \setminus C_i) = w(\tilde{H} \setminus C_i) + w(N \setminus (\tilde{H} \cup \{i\}))$ and

$$\begin{aligned} &w(N \setminus (\tilde{H} \cup \{i\})) \\ &\quad = w(\left\{j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = s\right\}) + w(\left\{j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = r\right\}), \end{aligned}$$

we have

$$w(\left\{j \in N \setminus (\tilde{H} \cup \{i\}) | \tilde{t}_j = s\right\}) + w(\tilde{H} \setminus C_i) > q.$$

The above result violates Equation (3.1). So, σ cannot be an equilibrium rejection. Hence, $\sigma_i(s)$ should be 1.

However, this is true for any $i \in C$. Contradiction. □

Lemma 8.

A weighted majority rule $f \in G(w)$ is interim self-stable among $G(W)$ if, for any individual i, there exists a minimal winning coalition including i which is not mutually exclusive with any other minimal winning coalition.

Proof.

Denote $\hat{C}_i$ a minimal winning coalition which is not mutually exclusive with any other minimal winning coalition. Since we are focusing on the case where $w_f = w_g$ for any g, it is either $q_f < q_g$ or $q_f > q_g$.

First, consider the case where $q_f < q_g$. So, if $f(t) = S$, then $g(t) = S$. And if $f(t) = R$, then $g(t)$ could be either S or R. Suppose a strategy profile (σ, μ) such that $\sigma_i(t_i) = 0$ for all i and $t_i \in T_i$, $\sigma_i^k(s) = k^{-\frac{1}{w_i}}$ and $\sigma_i^k(r) = k^{-\frac{2}{w_i}}$. Pick an agent i. Suppose for a minimal winning coalition C_m and a type profile t_{-i}, the convergence speed of

$$\lim_{k\to\infty} \left(\prod_{j \in C_m \setminus \{i\}} \sigma_j^k(t_j) \right) \left(\prod_{j \in N \setminus C_m} (1 - \sigma_j^k(t_j)) \right)$$

is slower than for any other minimal winning coalition. By construction, $W(C_m) \geq W(\hat{C}_i)$ and $t_j = s$ for $j \in C_m \setminus \{i\}$. For such t_{-i}, $f(t) = S$ if $t_i = s$, and $f(t)$ could be either S or R if $t_i = r$, because C_m has no mutually exclusive minimal winning coalition. Thus the right hand

side of Equation (3.4) is weakly less than the left for any type and any agent. This is true for any i. The strategy profile σ and the derived belief system μ is an equilibrium rejection of g.

Second, consider the case where $q_f > q_g$. So, if $f(t) = S$, then $g(t)$ could be either S or R. And if $f(t) = R$, then $g(t) = R$. Suppose a strategy profile (σ, μ) such that $\sigma_i(t_i) = 0$ for all i and $t_i \in T_i$, $\sigma_i^k(s) = k^{-\frac{2}{w_i}}$ and $\sigma_i^k(r) = k^{-\frac{1}{w_i}}$. Pick an agent i. Suppose for a minimal winning coalition C_m and a type profile t_{-i}, the convergence speed of

$$\lim_{k\to\infty} \left(\prod_{j \in C_m \setminus \{i\}} \sigma_j^k(t_j) \right) \left(\prod_{j \in N \setminus C_m} (1 - \sigma_j^k(t_j)) \right)$$

is slower than for any other minimal winning coalition and type profile. By construction, $W(C_m) \geq W(\hat{C}_i)$ and $t_j = r$ for $j \in C_m \setminus \{i\}$. For such t_{-i}, $f(t)$ could be either S or R if $t_i = s$, and $f(t) = R$ if $t_i = r$, because C_m has no mutually exclusive minimal winning coalition. Thus the right hand side of Equation (3.4) is weakly less than the left for any type and any agent. This is true for any i. The strategy profile σ and the derived belief system μ is an equilibrium rejection of g. □