Similarity-Nash Equilibria in Statistical Games*

Rossella Argenziano[†] and Itzhak Gilboa[‡]

23 July 2019

Abstract

A *statistical game* is a strategic form game accompanied by a prediction problem, where a characteristic x of the game may be used to predict an outcome y based on past values of (x, y). Players' strategies determine the outcome and each player's payoff depends on the others' moves only via this outcome. Players combine statistical and strategic reasoning, so that an estimate of y can be used for equilibrium selection. We focus on prediction by similarity-weighted frequencies, where the similarity function is learnt from the data. We prove results that capture the importance of precedence and the endogenous formation of sunspots.

*Gilboa gratefully acknowledges ISF Grant 1077/17 and the Investissements d'Avenir ANR -11- IDEX-0003 / Labex ECODEC No. ANR - 11-LABX-0047.

†Department of Economics, University of Essex. r_argenziano@essex.ac.uk

‡HEC, Paris and Tel-Aviv University. tzachigilboa@gmail.com

1 Introduction

1.1 A Motivating Example

The Soviet bloc started collapsing with Poland, which was the first country in the Warsaw Pact to break free from the rule of the USSR. Once this was allowed by the USSR, practically all its satellites in Eastern Europe underwent democratic revolutions, culminating in the fall of the Berlin Wall in 1989. The single precedent of Poland generated a "domino effect." This paper suggests a belief formation process that explains how a single precedent can have such a dramatic effect even in the absence of informational spillovers and strategic dependency among games.

Revolution attempts are typically modeled as coordination games: the expected utility derived from taking part in an uprising increases in the probability of it success, which in turn increases in the number of participants[1]. For a citizen trying to decide whether to join such an attempt, it is crucial to predict the outcome of the uprising. A natural piece of information to use for such a prediction is the outcome of past revolutions in similar contexts.[2] We suggest that the importance of the successful revolution in Poland didn't lie only in changing the relative frequency of successful revolutions, but also in changing the notion of *which* past revolution attempts were similar to current ones, hence relevant to predict their outcomes. Specifically, the case of Poland was the first revolution attempt after the "Glasnost" policy was declared and implemented by the USSR. Pre-Glasnost attempts in Hungary in 1956 and in Czechoslovakia in 1968 had failed. In 1989, one might well wonder, has Glasnost made a difference? Is it a new era, where older cases of revolution attempts are no longer relevant to predict the outcome of a new one, or is it "Business as usual", and Glasnost doesn't change much more than does, say, a leader's proper name, leaving pre-Glasnost cases relevant for prediction?

If the revolution attempt in Poland were to fail as did previous ones, it would seem that the variable "post-Glasnost" does not matter for pre-

[1]See, for example, Edmond, 2013.

[2]Steiner and Stewart, 2008, Argenziano and Gilboa, 2012, and Halaburda, Jullien, and Yehezkel, 2017 provide models in which similarity-weighted frequencies of past cases are used to form beliefs in coordination games.

diction: with or without it, revolution attempts fail. As a result, when a person wonders what is the “right” way of judging similarity between past cases, she would likely be led to the conclusion that the variable “post-Glasnost” should be ignored, and that, consequently, the statistics are zero successes out of 3 revolution attempts. By contrast, because the revolution attempt in Poland succeeded, it had a double effect on the statistics. First, it increased the frequency of successful revolutions from 0:2 to 1:3. While $\frac{1}{3}$ is larger than 0, it still leads to pessimistic predictions about successes of future attempts. However, if people also learn how to judge similarity, the single case of Poland leads them to the conclusion that “post-Glasnost” is an important variable. Indeed, the frequency of successes post-Glasnost, 1:1, differs dramatically from the pre-Glasnost frequency, 0:2 . Once this is taken into account, pre-Glasnost events are not as relevant for prediction as they used to be. If we consider the somewhat extreme view that post-Glasnost attempts constitute a class apart, the relevant empirical frequency of success becomes 1:1 rather than 1:3. Correspondingly, other countries in the Soviet Bloc could be encouraged by this single precedent, and soon it wasn’t single any more.

1.2 Statistics and Equilibrium Selection

To capture the reasoning above in a formal model, we suggest three steps:

(i) First, we define the notion of a statistical game. The term refers to a strategic game accompanied by a statistical problem, of predicting a variable y based on an observed characteristic x and on past values of both x and y. The statistical problem interacts with the game in two ways: first, the value of y is determined by the players’ strategy choices (as well as by the current value of x); second, it affects the payoffs of the game. We assume that a player’s utility depends only on her own strategy and on the values of (x, y) in the current period. That is, once the value of y has been determined, a player need not guess what the other players are doing in order to make her optimal choice, and she certainly can ignore past values of the variables. In this sense, the current values of (x, y) are a “strategically-sufficient statistic” for the game.

In our motivating example, the game is a standard coordination game, in which each player needs to choose one of two strategies, namely whether to join the revolution attempt or not. The characteristic x denotes the current state of the polity, and it is potentially payoff relevant. The variable y indicates the success of the revolution attempt, and it depends on the players' choices (with the probability of $y = 1$ increasing in the number of players who join the revolution) and clearly affects their payoffs. However, given the realization of y, neither the characteristics x not the outcomes y of past revolutions affect players' payoffs.

(ii) Next, we propose that, when confronted with a statistical game, players combine statistical and strategic reasoning. Pure statistical reasoning would estimate the variable y (based on the observed current value of x and on past values of both x and y), and use it for optimal choice, ignoring the fact that other players are simultaneously deliberating their choices and thereby affecting y. Pure strategic reasoning, on the other hand, when commonly known, would focus on equilibria of the game given observations. Such equilibria may well ignore past values of the variables (x, y), which are payoff-irrelevant (as well as the current value of x if it is also payoff-irrelevant). Indeed, if a player believes that all other players ignore the past in making their choices, so should she, and the history of the statistical problem will become irrelevant. However, we propose that both modes of reasoning interact: even when all players are smart enough to realize that they are playing a game, and when this fact is common knowledge among them, our solution concept relies on the supposition that players use statistics as a starting point for strategic reasoning. As discussed below, our solution concept is compatible with many possible assumptions about rationality and higher order beliefs in rationality, spanning a gamut between the pure-statistical and pure-strategic modes of reasoning.

Generally speaking, the two modes of reasoning may be in conflict. For example, consider a centipede games in which x denotes the length of the game, and y describes the number of steps where players chose "pass". Assume that past information on x and y describes games played by other players and their observed outcomes, say, as reported in experimental papers, and let us assume that the unique equilibrium of each of the past games has not always been played. Facing a new game, a player might wonder whether she should follow strategic or statistical reasoning. Is it more rational to ignore information about past games played by others, and focus on the unique equilibrium of the current game, or to trust statistics even though it is incompatible with equilibrium reasoning?

In this paper we do not delve into this problem. Rather, we focus on the conceptually easier cases in which the two modes of reasoning can coexist. Specifically, we ask whether, starting with a statistical estimate of y, and choosing a best-response to it, players end up at a Nash equilibrium. We limit attention to games in which the answer is in the affirmative, and thus use the estimate of y as an equilibrium selection device: of all Nash equilibria, we single out the equilibrium that can also be justified by pure statistical reasoning. In games where such a selection is possible, the equilibrium selection problem is reduced to the choice of a statistical estimator of y.

In our motivating example, a revolution game has two pure strategy Nash equilibria: one in which citizens participate in the revolution, which therefore succeeds with high probability, and one in which they do not. A player who uses purely-strategic reasoning would realize that the past is payoff irrelevant. If this fact is commonly known, and if, furthermore, it is commonly known that all players ignore history in their choices, it might be difficult to decide which equilibrium will be played, or if indeed one will be. By contrast, we assume that the players are aware of statistical estimates, and that these serve as a natural coordination device. If we knew which statistical estimation method they employ, we would have a natural candidate for equilibrium selection. Clearly, in such simple games no tension arises between statistical and strategic reasoning.

(iii) Finally, we turn to ask which statistical estimation method may be a good candidate for modeling players' reasoning in the games we have in mind. Statistics and machine learning offer a wide range of estimation and learning techniques and, in principle, each of these can be used as a way to define coordination devices.[3] However, not all these methods can serve as reasonable models of the way most people think, as in the example of the revolution games. In this paper we select a very simple prediction method,

[3]Indeed, one may embed the game in a reasoning game, where each player first chooses a method of reasoning, and then chooses an act that is a best response to the estimate that this method generates. If the original game is a coordination game, so will be the reasoning game.

namely, estimating probabilities by empirical frequencies in similar cases in the past. This, however, begs the question, which cases are deemed similar? We focus here on a simple, qualitative version of this question: will x be used to predict y or will x be ignored? In other words, will x act as a coordination device? We focus on the simplest possible model in which both x and y are binary. Ignoring x would mean estimating the probability that y be 1 by the overall (unconditional) frequency of $y = 1$ in the past; by contrast, taking x into account would estimate it by the (conditional) empirical frequency of y in the sub-database in which x had the same value currently observed. Importantly, we assume that players choose a method that would have performed best had it been used in the past. This is a (very) special case of the "empirical similarity" as in Gilboa, Lieberman, and Schmeidler (2006) and Argenziano and Gilboa (2019).

The rest of the paper is organized as follows. In Section 2 we present the formal definition of statistical games. Section 3 discusses statistical and strategic reasoning, and formally defines equilibrium selection by statistical estimators. In Section 4 we introduce the estimators that are of interest to us and their associated equilibrium concepts, with the main one being "Similarity-Nash Equilibrium". Section 5 presents binary coordination games and states several propositions that formally capture our account of the motivating example. We also discuss how they apply to other examples such currency redenomination by a central bank. Section 6 discusses related literature, while Section 7 concludes with a general discussion.

2 Definition of Statistical Games

A *statistical problem* is a prediction problem of a variable y by an observable variable x. More precisely, at time t, one is given past values of all variables, $(x_i, y_i)_{1 \leq i < t}$, and current values of the *characteristic*, x_t, and is asked to guess the value of the *outcome* y_t. A statistical problem is thus defined by $\left((x_i, y_i)_{i<t}, x_t\right)$. We restrict attention to *binary* statistical problems, in which $x_i, y_i \in \{0, 1\}$ for all i.

A *statistical game* is a quintuple $G = \left(H, \left(A^h\right)_{h \in H}, \left(u^h\right)_{h \in H}, \left((x_i, y_i)_{i<t}, x_t\right), \nu\right)$ where:

(i) H is a (finite or infinite) non-empty set of players;

(ii) For $h \in H$, A^h is a finite and non-empty set of strategies;

(iii) For $h \in H$, $u^h : \{0,1\} \times A^h \times \{0,1\} \to \mathbb{R}$ is player h's payoff function, which depends on the characteristic x, the player's own strategy, and the outcome y;

(iv) ν is the distribution of y given the players' strategies and the current values of the characteristic: $\nu \in [0,1]^{\{0,1\}\times A}$ for $A \equiv \prod_{h\in H} A^h$.

The statistical game G is played in two stages, as follows.

• Stage 1: All players observe x_t and take simultaneous actions: player $h \in H$ selects an action $a^h \in A^h$, determining $a = \left(a^h\right)_{h\in H} \in A$;

• Stage 2: Nature selects a value for the outcome y_t according to the distribution $\nu(x_t, a) = \Pr(y_t = 1 \mid x_t, a)$. The game ends and player h's payoff is $u^h\left(x_t, a^h, y_t\right)$.

A statistical game differs from a standard game in two ways. First, it is augmented by a statistical problem $\left((x_i, y_i)_{i<t}, x_t\right)$. This problem is implicitly assumed to be commonly known to all players, as are the sets of players, their strategies, etc.[4] Second, the statistical problem summarizes the strategic aspect of the game: a player who knows what y_t is about to be may well ignore the strategy choices of the other players because her payoff does not depend directly on their strategies. Although y_t has a distribution that depends on all players' choices, its realization is a *strategically-sufficient statistic* for the game G.

If we were to allow y_t to assume values in A, rather than in $\{0,1\}$, any game could be embedded in a statistical game: one could simply set, for every $a = \left(a^h\right)_{h\in H}$ and every x_t, $\nu(x_t, a) = \Pr(y_t = a \mid x_t, a) = 1$, so that no loss of strategic information would be entailed by summarizing the play of the game by y_t, and then $u^h\left(x_t, a^h, y_t\right)$ can describe whatever payoff function player h has. A key feature of statistical games is that, like in aggregative games, each player's payoff only depends on a statistic of the players' choices, which in our case is limited to be a binary variable. On the other hand, statistical games are richer than standard games (and

[4] We implicitly assume that all the players encode information in the same way and that they agree on the meaning of statements such as "$x_i^j = 0$" or "$y_i = 1$". If, for instance, different players think of a given case as a "success" ($y_i = 1$) and others – as a "failure" ($y_i = 0$), without a 1-1 mapping between the different languages they use, we cannot assume a common process of statistical learning.

aggregative games) in that they are equipped with additional data, namely the current values of x_t, *and the past values of* x *and* y. These past values are payoff-irrelevant: u^h depends on $\left(x_t, a^h, y_t\right)$ but not on $(x_i, y_i)_{i<t}$. Yet, the existence of the latter as part of the description of the game allows the players to coordinate on its values.

3 Statistical-Strategic Reasoning

How does a player $h \in H$ choose her action in game G? There are at least two approaches to the player's problem. The first relies on the fact that the player's payoff does not depend on the others' choices beyond the realization of y_t. Thus, the player can ask herself what y_t is likely to be, and best-respond to her estimate of that variable. The estimation of y_t (given x_t and previous values $(x_i, y_i)_{i<t}$) is a statistical problem. Providing the best estimate of y_t based on a statistical technique, and responding to this estimate, would therefore be referred to as "statistical reasoning". Estimating y_t as if it were independent of the player's own current action is perfectly valid if the set of players is infinite, and each player's choice has no impact on the realization of y_t (as in Schmeidler, 1973). Otherwise, in particular, if the set of players is finite, ignoring one's influence on y_t is, strictly speaking, a logical fallacy. Yet, when the player has negligible impact on y_t it can be a reasonable approximation. In many cases of interest, such as the examples of revolutions we started out with, presidential elections, financial crises, or bank runs, each individual player can assume that her own impact on the variable y_t is negligible, estimate the probability that $y_t = 1$ by some statistic $\bar{y}_t\left((x_i, y_i)_{i<t}, x_t\right)$ and choose a^h to maximize

$$\bar{y}_t\left((x_i, y_i)_{i<t}, x_t\right) u^h\left(x_t, a^h, 1\right) + \left[1 - \bar{y}_t\left((x_i, y_i)_{i<t}, x_t\right)\right] u^h\left(x_t, a^h, 0\right)$$

as if $\Pr(y_t = 1)$ did not depend on a^h.[5]

Another approach to the player's problem involves reasoning about all the details of the strategic interaction. This means that the player takes

[5] The logical fallacy involved in ignoring one's own impact on y_t is akin to the logical fallacy that price-taking agents commit in the standard competitive equilibrium model (with finitely many agents): each agent can affect prices by changing her demand/supply in different markets, and yet, she selects her optimal bundle as if the prices were given.

into account not only the dependence of her payoff on y_t, but also the dependence of the latter on all the players' actions. The player is thus led to strategic reasoning, and will feel comfortable only with predictions that are equilibria of the game.

We hold that both modes of reasoning tend to coexist in people's minds. While the purely-statistical reasoning mode relies on a wrong assumption of causal independence (of y_t with respect to a^h), it is a natural process of induction that is hard to ignore. For example, asking oneself whether most people would drive on the right or on the left tomorrow morning, it is almost inevitable that one would think "most drivers will drive on the right, as they do every morning". By contrast, the purely-strategic mode of reasoning concedes that the problem is one of equilibrium selection, and that multiple equilibria exist. Indeed, it is possible that at some point all players would switch to another equilibrium, and, say, drive on the left on the following morning.

As mentioned in the Introduction, statistical and strategic reasoning modes might in general lead to different predictions. Moreover, it is not obvious that the strategic one, which fully incorporates all causal dependencies and assumes that rationality is commonly known among all players, is necessarily more accurate than the statistical one.[6] In this paper, however, we focus on cases where statistical and strategic reasoning can coexist. We assume that the players use statistics to estimate y_t by some $\bar{y}_t\left((x_i, y_i)_{i<t}, x_t\right)$, as if they were outside observers of the process, rather than active participants therein, and then play a best response to $\bar{y}_t\left((x_i, y_i)_{i<t}, x_t\right)$. We limit our analysis to cases in which these best-responses are also equilibria of the game G. Thus, we use the estimate $\bar{y}_t\left((x_i, y_i)_{i<t}, x_t\right)$ as an initial guess that the players make, and use it as an equilibrium-selection device. The resulting equilibrium will be compatible with different assumptions on the players' rationality and beliefs about it, as explained shortly. We first complete the formal definition.

Let an *estimation function* be

$$Y : \cup_{t\geq 1}\left[\{0,1\}^2\right]^{t-1} \times \{0,1\} \to [0,1]$$

[6]Selten (1978) discusses three different modes of reasoning in the context of his celebrated Chain Store Paradox.

Such a function takes past observations, $(x_i, y_i)_{i<t}$, and the current characteristic, x_t, and generates a number which is interpreted as the probability that $y_t = 1$.

Next, we augment the statistical reasoning embedded in Y with optimizing behavior. Let there be given a statistical game

$$G = \left(H, \left(A^h\right)_{h \in H}, \left(u^h\right)_{h \in H}, \left((x_i, y_i)_{i<t}, x_t \right), \nu \right).$$

Consider a strategy a^h for a player $h \in H$. If

$$a^h \in \arg\max_{\hat{a}^h \in A^h} \left[\bar{y}_t u^h \left(x_t, \hat{a}^h, 1 \right) + (1 - \bar{y}_t) u^h \left(x_t, \hat{a}^h, 0 \right) \right],$$

where $\bar{y}_t = Y\left((x_i, y_i)_{i<t}, x_t \right)$, we refer to a^h as a Y*-best-response* strategy of player h. Finally, we consider the strategic aspect of the game by restricting attention to strategy profiles at which players best-respond not only to their own prediction of y (based on an estimation function Y), but also to others' choices. Formally, a strategy profile $\left(a^h\right)_{h \in H} \in A$ consisting of Y-best-response strategies for each h, which also happens to be a Nash equilibrium of G is called a Y*-Nash equilibrium* of G.

The notion of Y-Nash equilibrium thus uses the statistical method Y as a belief coordination device, pointing at an initial conjecture $\bar{y}_t$. For example, Y may suggest that outcome $y_t = 1$ will occur with probability 90%. Players' best responses to this conjecture will generate a probability of $(y_t = 1)$ that might differ from 90%. Indeed, it might be 1. Our requirement that the strategies a^h constitute a Nash equilibrium implies that these strategies are best responses not only to the original conjecture $\bar{y}_t$ but also to the equilibrium beliefs. Thus, the players are assumed to choose strategies that are best responses to each other, as in any equilibrium, and the requirement of best-responding to the initial conjecture $\bar{y}_t$ is an additional condition that is used for equilibrium selection.

Throughout most of the discussion we promote the perfectly rational interpretation, according to which statistics are only used as coordination devices. Specifically, each player has beliefs over the reasoning modes of all players, as well as all higher order beliefs about these reasoning modes. Any estimation function Y may be commonly known among the players and thus select an equilibrium of the game, namely, a Y-Nash equilibrium. However, the notion of Y-Nash equilibrium is also compatible with a wide range of less stringent assumptions about the players' rationality and about their higher order beliefs regarding other players' reasoning. Consider the following two examples:

(i) Suppose that only a proportion $\beta \in [0,1]$ of the players think strategically. We can assume that, among these players, the entire game, including the value of β, is commonly known. The fraction $(1-\beta)$ of non-strategic players use statistical reasoning to generate the estimate $\bar{y}_t$ and respond optimally to it. In the example above, they never update their belief that the probability of $(y_t = 1)$ is 90% . The others generate the same conjecture $\bar{y}_t$ but also perform strategic reasoning and infer what the actual probability of $(y_t = 1)$ will be in equilibrium (in the example above, this probability is 1). The case $\beta = 0$ is a purely behavioral model in which players' actions constitute an equilibrium even though the players fail to correctly estimate others' players moves: they believe that y_t will be 1 with probability of 90%, and fail to realize that this very belief implies that the probability is 100%. In the other extreme case, in which $\beta = 1$, the players' actions are in equilibrium, and they have perfectly accurate beliefs, and higher-order beliefs, about the equilibrium play. Indeed, in this case the initial statistical estimate is but a coordination device and there are other equilibria in which the players ignore it. Finally, if $\beta \in (0,1)$, the strategic and the non-strategic players have different beliefs, but they all play equilibrium actions: the non-strategic players respond to the prediction that y_t be 1 with probability of 90%, and as a result they play the same strategies that do the strategic players.

(ii) Suppose next that each player is capable of Level-K reasoning for a given $K \leq \infty$ (see Nagel, 1995, Stahl and Wilson, 1995). There may be players at level $K = 0$, who are incapable of strategic reasoning, estimate $P(y_t = 1)$ by $\bar{y}_t$ and respond optimally to this estimate. There are others who are at level $K = 1$, who compute $\bar{y}_t$ as well as the best response to this estimate, and believe that this best response would be the choice made by all the other players. And then there are players who assume that all the others are performing level $K = 1$ reasoning and best respond to that assumption, and so forth. Eventually we may find also players of Level-∞

reasoning, who can compute equilibria. These players may also be sophisticated enough to have beliefs over the distribution of levels of reasoning in the population. A Y-Nash equilibrium consists (by definition) of strategies that are best response to the initial guess, $\bar{y}_t$, and to themselves, and thus all levels of reasoning would lead to the same choices, namely the equilibrium strategies. As in the case $\beta = 1$, above, if all players are capable of Level-∞ reasoning, and if this fact is commonly known among them, the equilibrium selection by best-responding to $\bar{y}_t$ might appear rather arbitrary. But if there are some players of lower levels of reasoning, then it makes sense for Level-∞ reasoning players to play this equilibrium. Similarly, even if all the players are in fact capable of Level-∞ reasoning, but this fact is not common knowledge among them, we are led again to the Y-Nash equilibrium.

Thus, the equilibrium selection proposed by an estimation function Y is rather robust to assumptions about rationality and common belief thereof. However, the discussion above implicitly assumes that the estimation function Y is sufficiently natural and conspicuous to be a good model of the way people think about uncertainty. We now turn to the choice of such a function.

4 Similarity Nash Equilibria

We assume that prediction is made by empirical frequencies of past outcomes. In the spirit of fictitious play (Brown, 1951), where players best respond to the empirical frequency of moves by the other players, it is natural to assume that they would respond to the empirical frequency of past values of y. However, not all past observations are necessarily equally relevant. We present an estimation function which captures a process of "first-order" induction: "from causes [x] which appear similar, we expect similar effects [y]", (Hume, 1748). More precisely, we consider similarity-weighted empirical frequencies, defined as follows: let there be given a similarity function $s : X \times X \to \mathbb{R}_+$. Define the estimation function Y^s by:

$$Y^s\left((x_i, y_i)_{i<t}, x_t\right) = \overline{y}_t^s = \frac{\sum_{i<t} s(x_i, x_t) y_i}{\sum_{i<t} s(x_i, x_t)} \tag{1}$$

if $\sum_{i<t} s(x_i, x_t) > 0$ and $\overline{y}_t^s = 0.5$ otherwise. In case there is a unique Y^s-Nash equilibrium for a game, we refer to it as the *Nash equilibrium selected by the function* s.

Similarity-weighted relative frequencies are formally equivalent to kernel estimation of probabilities (Akaike, 1954, Rosenblatt, 1956, Parzen, 1962; see Silverman, 1986) and they are also reminiscent of exemplar learning in psychology (Shepard, 1957, 1987, Medin and Schaffer, 1978, Nosofsky, 1984, 1988). The formula has also been axiomatized in Billot, Gilboa, Samet, and Schmeidler (2005) (if y takes at least three values), and in Gilboa, Lieberman, and Schmeidler (2006) (for the case of two values discussed here).

In this paper we focus on the simplest case in which the variables x, y are binary, and further assume that so if the function s. Further, we limit attention to two possibilities: the simple, *unconditional* empirical frequencies where the similarity is defined by

$$s_0 (x_i, x_t) = 1$$

for all x_i, x_t, and the *conditional* ones where the similarity is defined by

$$s_x (x_i, x_t) = \mathbf{1}_{\{x_i = x_t\}}. \tag{2}$$

Thus, while the general model of similarity-weighted empirical frequencies allows for a gamut of similarity functions between s_0 and s_x, we here limit attention only to these two extremes.

Steiner and Stewart (2008), Argenziano and Gilboa (2012), Halaburda, Jullien, and Yehezkel (2016) deal with Nash equilibria selected by appropriately defined similarity functions. As opposed to this literature, in this paper we do not assume that a similarity function is given a priori, but that it is learned from the data itself. This process is referred to as "second order induction" in Gilboa, Lieberman, and Schmeidler (2006) and Argenziano and Gilboa (2019), because the *way* one learns from the past about the future, that is, the similarity function, is also learned from data. This idea also appeared both in the statistical literature (Hardlë and Marron, 1985) and in the psychological one (Nosofsky, 2011).

Which similarity functions would players use? Which of the two similarity functions best fits the data? We use a leave-one-out cross-validation technique, and define an "empirically optimal similarity" as a similarity function that, had it been used to predict the existing data points, where each is estimated based on the others, would have performed best.[7] We consider this similarity selection method as an obviously-idealized model of the process people implicitly undergo in learning similarity from data.

Formally, for a similarity function s, and $i < t$, define

$$\overline{y}_i^s = \frac{\sum_{r \neq i} s(x_r, x_i) y_r}{\sum_{r \neq i} s(x_r, x_i)}$$

and consider the sum of squared errors,

$$SSE\left(s\right) = \sum_{i=1}^{t-1} \left(\overline{y}_i^s - y_i\right)^2$$

We assume that players choose a similarity function (between s_0 and s_x) that minimizes the SSE. In case of a tie we assume that s_0 is selected, as it appears to be a simpler theory, using less variables in the similarity judgment.[8]

When will players find that using the variable x reduces the sum of squared errors? That is, when will we have $SSE\left(s_x\right) < SSE\left(s_0\right)$? Intuitively, splitting the database according to the realization of x can make predictions more accurate because it relates a current case only to cases that are more similar to it. On the other hand, it exposes the predictor to the "curse of dimensionality": some observations might be "isolated" and have too few other cases that are considered similar to them. Therefore, a larger difference between the relative frequencies of $y_i = 1$ (vs. $y_i = 0$) in the sub-database of $x = 0$ and of $x = 1$ suggests that including x involves greater gain in accuracy; by contrast, a low number of observations overall indicates a more severe curse of dimensionality.

The function $s \in \{s_0, s_x\}$ that minimizes the SSE (with the tie-breaking rule above) is *the empirically optimal similarity function.* It defines an es-

[7] See Argenziano and Gilboa (2019) for similar definitions in a continuous model.

[8] The preference for fewer variables is similar to the simplicity criteria implicit in the adjusted R^2, Lasso, the Akaike Information Criterion etc.

timation function

$$Y^{ES}\left((x_i,y_i)_{i<t}\,,x_t\right)=Y^s\left((x_i,y_i)_{i<t}\,,x_t\right)$$

Observe that in this expression $s=s\left((x_i,y_i)_{i<t}\right)$, that it, it is the similarity function that was learned from the database. Y^{ES} thus does not equal Y^s for any fixed s, because different databases are likely to result in different empirically optimal similarity functions s.

Our focus is on Nash equilibria selected by this estimation function, and we dub a Y^{ES}-Nash equilibrium of G a *Similarity-Nash equilibrium (SNE)* of G.

5 SNE in Binary Coordination Games

5.1 Binary Coordination Games

A *statistical binary coordination game* is a statistical game G with the following properties:

(i) For every $h\in H$, $A^h=\{0,1\}$;

(ii) λ is a measure on 2^H ;[9]

(iii) For every $h\in H$, the utility function is given by[10]

$u^h\left(x_t,a^h,y_t\right)$	$y_t=1$	$y_t=0$
$a^h=1$	1	0
$a^h=0$	$d\left(x_t\right)$	$c\left(x_t\right)$

with $0\le d\left(x_t\right)<c\left(x_t\right)\le 1$ for $x_t\in\{0,1\}$;

(iv) There exists a strictly increasing function $f:[0,1]\to[0,1]$ such that

$$\nu\left(x_t,a\right)=\Pr\left(y_t=1\,|x_t,a\right)=f\left(\alpha\right)$$

where α is the λ-measure of players (in H) that chose $a^h=1$. We also assume that $f\left(0\right)=\varepsilon$ and $f\left(1\right)=1-\varepsilon$ for some $\varepsilon\in\left(0,0.5\right)$.

[9] λ_t measures the size of a set of players that choose a given strategy. If H is finite, it is most natural to define, for $H'\subseteq H$, $\lambda_t\left(H'\right)=\left|H'\right|/\left|H\right|$. If $H=[0,1]$, we can define λ to be a finitely additive extension of Lebesgue's measure. At equilibrium the relevant sets of players will typically be measurable.

[10] We allow x to be payoff relevant but the results in this section hold also if it is not.

In the revolution game example, $a^h = 1$ stands for player h deciding to join the revolution attempt, and $y_t = 1$ – for the success of this attempt. In this example, the characteristic x, is the indicator of "post-Glasnost".

Consider statistical reasoning first. If a player h knew y, her best response would be $a^h = y$. The player attempts to predict y_t and, given an estimate $\overline{y}$, her best response is $a^h = 1$ iff $\overline{y} \geq \frac{c(x_t)}{1-d(x_t)+c(x_t)}$. Therefore, if players' common estimate is $\overline{y} \geq \frac{c(x_t)}{1-d(x_t)+c(x_t)}$, they all play $a^h = 1$, otherwise they all play $a^h = 0$. Next, consider strategic reasoning: given $f(\alpha)$, $\left(a^h = 1\right)_h$ and $\left(a^h = 0\right)_h$ are pure strategy Nash equilibria of the period-game G_t. Therefore, for every estimation function Y, G has at least one *Y-Nash equilibrium.* We will focus on Y^{ES} and study the properties of SNE of these games.

5.2 SNE and the Evolution of Coordination Devices

Consider the game in section 5.1. If players use estimation function Y^{s_0}, they ignore x in forming their prediction of y and choosing their strategy. If instead they use estimation function Y^{s_x}, they use x as a coordination device: they form predictions about y looking at its relative frequency in all, and only, the past games where the realization of x was the same as in the current one. We assume that players *learn,* through second order-induction, whether x should be used as a coordination device. In other words, they use estimation function Y^{ES}. We ask under what circumstances x will become, or remain, a coordination device. We use the results to argue that the SNE, i.e., the Nash equilibria selected by second order induction, have some intuitively appealing properties.

It will be convenient to use the following notation: there are $(t-1)$ points in the database, and they are divided into four types, according to the values of x and of y. Let the number of cases of each type be given by the following case-frequency matrix:

# of cases	$x=0$	$x=1$
$y=0$	L	l
$y=1$	W	w

In the motivating example of subsection 1.1, let $y = 1$ (or zero) denote the success (or failure) of a revolution attempt (w for "win", and l for "lose"), while $x = 1$ (or zero) – whether or not it occurred post-Glasnost. Consider citizens in Hungary in 1989. They lived in a post-Glasnost world, i.e., $x = 1$. After the successful revolution in Poland, they observed two failed revolutions pre-Glasnost, and a successful one post-Glasnost: $(L, W, l, w) = (2, 0, 0, 1)$. At this point, if they had ignored x they would have predicted failure as the most likely outcome of a revolution attempt (with a relative frequency of 2/3) and therefore found it optimal not to take part in one. Instead, by taking into account x they would have considered only the case of Poland as relevant for their predictions, expected a success, and therefore participated in the attempt. Second-order induction is consistent with the fact that Glasnost was indeed considered relevant for predictions and a revolution was therefore attempted (successfully) in Hungary. Ignoring x yields $SSE\,(s_0) = 1.5$ while taking it into account reduces the sum of squared errors to $SSE\,(s_x) = 0.25$. Thus, the single case of a successful revolution made the variable "post-Glasnost" informative enough to enter the similarity judgment. Note that, had the case of Poland ended in a failure, $SSE\,(s_0) = 0$ would hold and the empirically optimal similarity would not attach any importance to the post-Glasnost variable.

In the rest of this section we will focus on larger databases, assuming that there is a non-trivial history in which $x = 0$. Specifically, we assume throughout that $L, W > 2$. This assumption means that (i) history contains a non-trivial number of cases overall, and that (ii) the prediction of the variable in question, y, is a non-trivial task: there are a few (at least three) cases with $y = 0$ as well as with $y = 1$.

5.2.1 A New Value

We start by looking at SNE of statistical games for which the history does not contain any case with $x = 1$. There's a non-trivial history of cases with different outcomes but the characteristic x had a constant value $x = 0$ in all of them. In terms of frequencies: $L, W > 2$ and $l = w = 0$. Consider classical examples of coordination games such as a revolutionary attempt, a bank run, or a currency attack. Suppose that, in a sequence of such games, $x = 1$ is observed for the first time. Thus, $x_t = 1$ might denote the appearance of a new political leader, or the announcement of a new policy. History includes cases with various outcomes of analogous attempts to attack a government, a bank, or a currency. Some succeeded, some failed. But in all these cases, the new leader or policy was not in place (x was constantly equal to zero). As a result, x doesn't have any predictive power in the existing database, hence the first time that $x = 1$ appears, it is ignored.[11] The natural question then is: what will it take for players to start paying attention to the new feature? Starting from a clean slate, what does it take for a new leader or policy to be taken seriously, to be considered something that separates history into two periods: a past regime, which is not relevant anymore, not similar to the current game in our terminology, and a new regime which contains cases relevant for predicting the outcome of the current game?

Our first two results answer exactly this question. Proposition 1 says that even a single case is sufficient to convince players that they are under a new regime, if and only if the observed outcome y is the one which had been less frequently observed in the past. This result is rather intuitive: in order to be noticed, one needs to be different.

Proposition 1 *Let $L, W > 2$. If $(l, w) = (1, 0)$, any[12] SNE is selected by s_x if $L < W$ and by s_0 otherwise. Symmetrically, if $(l, w) = (0, 1)$, any SNE is selected by s_x if $L > W$ and by s_0 otherwise.*

Thus a new feature (leader, policy, etc.) that results in the modal outcome cannot be selected by the empirically optimal similarity function. However, if it is consistently the case that $x = 1$ is associated with a particular value of y, we would expect the similarity function to "notice" this regularity by taking x into account. The following results corroborates

[11] Observe that, since all past cases have $x = 0$, the variable does not affect their similarity to each other. Thus, one obtains exactly the same in-sample predictions whether one the variable x or not. This means that $SSE(s_0) = SSE(s_x)$ and the empirically-optimal similarity function will by s_0 because of the tie-breaking rule which favours the simplest theory.

[12] Recall that for each similarity function the corresponding Nash equilibria are generically unique. In our setup there is always a unique empirically-optimal similarity function (either $s_\varnothing$ or $s_{\{1\}}$), and non-uniqueness can only follow from ties.

this intuition and shows that "consistently" need not be more than twice, provided that there are no counter-examples:

Proposition 2 *Let $L, W > 2$. If either ($l > 1$ and $w = 0$) or ($l = 0$ and $w > 1$), then any SNE is selected by s_x .*

The importance of this proposition lies in the comparison of case-based and rule-based reasoning: while our model does not equip players with the language in which general rules can be stated, learned, or acted upon, the empirically-optimal similarity function can mimic this type of reasoning. If it so happens that the associative rule "If $x_i = 1$ then $y_i = b$" (for $b \in \{0, 1\}$) is valid in the database, the players will notice this regularity: the empirically optimal similarity function will include the variable x, and therefore in any SNE of the game, if $x = 1$, players will expect $y = b$ and play $a^h = b$. By contrast, if $x = 0$, they will expect y to be equal to the average value of y in the past cases with $x_i = 0$ and play accordingly.

As an example, consider a central bank which redenominates its currency in an attempt to restrain inflation. Inflation is an equilibrium phenomenon: an economic agent who expects others to raise prices of goods and services would be wise to do so herself. Thus, one can think of the inflation game as a price-setting game with multiple equilibria, and redenomination as an attempt to switch from a hyperinflation equilibrium to a low inflation equilibrium[13]. If x denotes the new currency, then $x_i = 0$ throughout all cases in history ($i < t$), and setting $x_t = 1$ is an attempt to signal a new regime, and to coordinate on the non-inflationary equilibrium. Will economic agents use the new variable in their belief formation, or will they dismiss it as a "cosmetic change" and believe that inflation will continue to run high? Proposition 1 suggests that the answer depends on the first period: if, in this period, inflation is low – namely, y takes the value that was less frequent in the past – the variable will be used for prediction and a new, low-inflation equilibrium can be reached. By contrast, if in the first period the inflation rate continues to be high, the variable will be

[13]See Mosley (2005): "...redenominations often occur after economic crises, as governments attempt to convince citizens and markets that hyperinflation is a thing of the past. In some cases, the timing is correct, in that redenomination caps off high levels of inflation. In other cases, governments are not able to reign in inflation immediately after redenomination, and they may make multiple efforts....".

judged irrelevant, and the entire history will be used for prediction, making it very difficult to convince economic agents that the future will differ from the past. Israel switched from a Lira to a Shekel (worth 10 Liras) in 1980 and then to a New Shekel (worth 1,000 Shekels) in 1985. In 1980 the change was not accompanied by fiscal policy changes, and inflation spiraled into hyper-inflation. By contrast, the change in 1985 was accompanied by budget cuts, and inflation was curbed in the following years. These two examples seem to corroborate the intuition behind Proposition 1: a change of currency appeals to a payoff-irrelevant but perceptually-conspicuous difference that might change the equilibrium selected; whether it succeeds in doing so depends on the realization of a payoff-relevant variable (y). In these examples psychological considerations suggest potential sunspots; but rational learning of the similarity function implies that economic outcomes will determine which sunspots are used for coordination and which get ignored.

5.2.2 The Power of a Single Precedent

Suppose now that after a non-trivial history ($L, W > 2$) of cases with $x = 0$, a new leader appeared, $x = 1$, and has established herself as relevant for prediction either through a series of consistent outcomes, as in Proposition 2 or through a single, "surprising" outcome, as in Proposition 1. The next proposition asks what would it take for the new leader to *lose* its role as a coordination device. Would a single inconsistency, a single precedent with the opposite outcome, be enough for the players to stop paying attention to the variable x? The result is rather intuitive: a single precedent can make a variable irrelevant for prediction if the number of consistent outcomes of the opposite sign that have established its relevance is not too large.

Proposition 3 *Let $L, W > 2$. If either ($l = 1$ and $0 < w \leq \lfloor \frac{W}{L} \rfloor + 1$) or, symmetrically, ($w = 1$ and $0 < l \leq \lfloor \frac{L}{W} \rfloor + 1$), then any SNE is selected by s_0.*

Consider the first statement (the second is symmetric): if relevance for prediction had been established with a single surprising outcome, i.e., if $W < L$ and $w = 1$, a single case ($l = 1$) makes the variable irrelevant again. Similarly, it makes it irrelevant if relevance had been established with multiple, but not too many, outcomes of the type most frequent in the past, i.e., if $W > L$ and $1 < w \leq \lfloor \frac{W}{L} \rfloor + 1$. Finally, note that, if $W > L$ and $w = 1$, we already know by Proposition 1 that the empirical similarity is s_0 for $l = 0$, and Proposition 3 shows that this is the case also for $l = 1$: if in the first case in which the new leader was in office the outcome of the game was the one most frequent in the past, the new leader does not become a coordination device, and that is still true even if a second case ends up having the opposite outcome.

5.2.3 The General Case

We now turn to the more general case, where a new leader ($x = 1$) appeared after a non-trivial history with $L, W > 2$, and outcomes of both types have been observed: $l, w > 0$. We ask what will it take for players to take into account the change in leadership when they form their beliefs. The basic intuition is, again, rather simple: if the ratio w/l is close to W/L, the change of leadership will seem immaterial and players will ignore it when forming beliefs: the empirically optimal similarity is s_0. If, however, the relative frequency of $y = 1$ in the sub-database corresponding to $x = 1$ is very different from that corresponding to $x = 0$, players will be convinced that they are under a "new regime" and the empirically optimal similarity will be s_x.

Proposition 4 starts from a scenario in which the sub-database with $x = 1$ has, up to integrality constraint, the same ratio of cases with $y = 0$ and $y = 1$ as the sub-database with $x = 0$. In this case x is irrelevant for predicting y (part (i) of the Proposition 4). Suppose that we now increase w. We find that this improves the performance of the similarity function that includes the variable, up to a point where it outperforms the similarity function that does not include it (part(ii)). As could be expected, the minimum $w^* > \frac{lW}{L}$ for which this inequality holds increases in the number of cases with the opposite outcome, l (part (iii)). Moreover, up to details of integrality constraints, the number of additional cases needed to get to this minimum ($w^* - \frac{lW}{L}$) is also non-decreasing in l (part (iv)).

Formally, let $[\;] : R \to Z$ be the nearest integer function, selecting the ceiling in case of a tie. (That is, for all $x \in R$ and $z \in Z$, we have $[x] = z$ if $x = z + \varepsilon$ and $\varepsilon \in [-0.5, 0.5)$.) We prove the following:

Proposition 4 *Let L, W, l, w be any four integers such that $L, W > 2$, $l > 0$, and $w = \left[\frac{lW}{L}\right] \geq 0$. The following hold:*

(i) For databases (L, W, l, w) and $(L, W, l, w+1)$, the unique SNE is the one selected by s_0.

(ii) There exists an integer $w^*(L, W, l) \geq w + 2$ such that, for every $q \geq w$, the unique SNE is the one selected by s_0 for $q < w^*(L, W, l)$ and by s_x for $q \geq w^*(L, W, l)$. (Clearly, if such an integer exists it is unique.)

(iii) $w^*(L, W, l)$ is non-decreasing in l.

(iv) If W/L is an integer, $(w^*(L, W, l) - w)$ is non-decreasing in l.

Thus, our model captures the fact that it is harder to re-establish relevance than to establish it at the outset. Suppose that a new leader whose identity is characterized by $x = 1$ wishes to associate herself with successes, that is, to make others predict that $y = 1$ when $x = 1$. Let us assume that in the past successes were less frequent than failures ($W < L$) so that history is not very helpful; if the leader does not single herself out, players will expect failures and such beliefs will be self-fulfilling. On this background, Proposition 1 guarantees that starting off with a single success ($w = 1$, $l = 0$) suffices to establish relevance of x and thereby to place the leader in a class apart. Importantly, in the corresponding database (defined by $x = 1$) only $y = 1$ has been observed and thus the leader is associated with success.

However, if it so happens that one starts out with a failure ($w = 0, l = 1$) the task will be harder: Proposition 1 guarantees that the leader's identity won't be considered relevant after the initial failure and parts (i) and (ii) of Proposition 4 shows that for the leader to be noticed, and associated with successes, at least two or three successes will be needed (depending on how unusual successes were in the past). More generally, for any number of adverse outcomes $l > 0$ there is a sufficiently large number of successes w that would eventually make x a coordination device followed by the players (part (ii)), but the number of successes required (part (iii)), and even the *additional* number of such successes (part (iv)) weakly increase (up to integrality constraints in part (iv)). One does get a second chance to make a first impression, but it becomes costlier.

6 Related Literature

Statistical games are reminiscent of "Aggregative Games" (Selten, 1970) and of "Congestion Games" (Rosenthal, 1973, Schmeidler, 1973) in that a player's payoff depends only on a summary statistic of the others' choices. In the former, strategies are real numbers and the statistic is their sum. In the latter, there are typically finitely many strategies and the statistic is the relative frequencies of choice. In both, each player finds the others interchangeable. Similarly, in statistical games each players should only bother about the prediction of y, and the others' choices only matter to the extent that they affect y. The definition of statistical games brings the summary statistic y to the fore, allowing for a variety of ways in which it is determined by players' choices, encapsulated in the function ν.[14] The additional freedom allowed by making y (and ν) explicit means that the game need not be symmetric, and that it might be meaningless to consider the sum of the players' strategies, or the frequency of their choices of players. More importantly, statistical games are equipped with a history, that is, a database of past observation of x and y, which has no counterpart in the standard models of aggregative or congestion games.

Statistical games are similar to Correlated Equilibria (Aumann, 1974) in that we assume that Nature sends a signal to each player before the game is played. However, in our context the signal is commonly known. Thus, the correlation device x (coupled with the database $(x_i, y_i)_{i<t}$) selects an equilibrium but does not allow non-equilibrium plays. In this sense our correlating signal, x, brings to mind "sunspots" (Cass and Shell, 1983). In particular, if one imposes the additional assumption that in a statistical game x is payoff-irrelevant, it does function, like sunspots, as a mere public correlation device. Viewed thus, our suggestion to use second-order induction to find the similarity function can be considered a theory of sunspot

[14]Note that, if we were to allow y to assume values in larger spaces, aggregative games and congestion games could be embedded in our model (by allowing y to be real-valued, or a point in a corresponding simplex, respectively).

selection.

When considered as a method of equilibrium selection in coordination games, statistical games cannot fail to remind one of "Global Games" (Carlsson and van Damme, 1993), especially if one considers a sequence of statistical games that have a given prediction problem in common (as in the revolutions example). Like Global Games, our approach attempts to embed the game in context in order to predict equilibrium selection. However, in Global Games equilibria are chosen ex ante, simultaneously for all games, whereas in statistical games they are chosen sequentially, highlighting the role of statistical learning. Global Games rely on some uncertainty about the game played, while a statistical game is commonly known among its players, and the variable x only serves as a coordination device.

In a 2x2 symmetric coordination game, Similarity-Nash equilibria are related to risk-dominant equilibria (Harsanyi and Selten, 1988). Specifically, assume that there is no history to be considered ($t = 1$). Then our definition of Y^s yields an initial guess of $P(y = 1) = 0.5$. When players best respond to this guess, they will select the risk-dominant equilibrium. Indeed, even when a history $(x_i, y_i)_{i<t}$ is available, the players may choose to ignore it, use $P(y = 1) = 0.5$ as a starting point and select the risk-dominant equilibrium. By contrast, Similarity-Nash equilibria assume that the initial statistical estimate is a function of history, where the values of (x, y) are used for weighted averaging, as well as for determining the weights in the averaging formula.

One can also view Similarity-Nash equilibria as a possible formalization of Schelling's (1960) focal points: estimating y based on its past values, and finding the equilibrium that consists of best responses to this estimate can be viewed as a procedure to determine focality. In the simplest case, assume that a game is played repeatedly and that a given equilibrium is played in the vast majority of past observations. It then stands to reason that a statistical prediction function would estimate a value of y that gives rise to the same equilibrium played in the past. In this sense, SNEs capture "statistical focality".

7 Extensions

7.1 Iterative Best Response

One may generalize the statistical-strategic reasoning process in several ways. In particular, best responses can be used to generate conjectures, which are fed back into the best response operators, in a way that parallels Level-K reasoning (see Nagel, 1995, Stahl and Wilson, 1995). For the class of games we considered above, the algorithm as defined will suffice. However, in more general coordination games one needs more than one step of best-response reasoning to arrive at an equilibrium. For example, consider a modified version of the sequence of revolutions described in Section 5. Suppose that $f(\alpha) = \alpha^2$ and that there is a continuum of heterogeneous players where player h's payoff is given by

Payoff to h	$y_t = 1$	$y_t = 0$
$a^h = 1$	$1 + \varepsilon^h$	0
$a^h = 0$	0	$1 - \varepsilon^h$

and $\varepsilon^h \sim U(-1, 1)$, so that her best response is to join the revolution attempt if and only if she thinks that the probability of success is at least $\frac{1-\varepsilon^h}{2} \sim U(0, 1)$. For any initial belief $\Pr(y_t = 1) = p_0 \in (0, 1)$, the best response would be to join the revolution for a fraction $\alpha_0 = p_0$ of the population and not to join it for the remaining fraction. This in turn would generate beliefs $p_1 = f(p_0) = p_0^2 < p_0$, to which the best response would be to join the revolution for an analogous fraction p_1 of the population. A formal analysis of such a game would require a generalized notion of Similarity-Nash equilibrium, allowing for an iterative process of best-response to initial beliefs. Such an iterative process would converge to an equilibrium with $\alpha = 0$ for any initial belief $p \in (0, 1)$.

This process brings to mind Level-K reasoning, where one does not start the process with an arbitrary, say, uniform distribution, but with the statistical one obtained from the empirical similarity weighted frequencies.

Note that an iterative process of best responses is at the heart of equilibrium selection in Global Games (Carlsson and van Damme, 1994). Thus, an extension of our equilibrium selection to iterative best responses can simultaneously generalize Global Games (by allowing different games) and our analysis above.

7.2 A Continuous Multi-Dimensional Model

While the analysis in this paper focused on a single binary characteristic x, it is natural to extend it so that x assume continuous values, and be multi-dimensional, $x = (x^1, ..., x^m) \in \mathbb{R}^m$. Similarly, the outcome y can also be continuous, and, importantly, so can be the similarity function. This is indeed the model studied in Argenziano and Gilboa (2019), which focuses on existence, uniqueness, and complexity of the empirical similarity function, with a focus on asymptotic behavior as the number of observations and/or the number of characteristics grow.

7.3 Conspicuity of Variables

What makes a characteristic conspicuous enough to be considered by all players a potentially relevant predictor? When do all players notice a certain characteristic, and when can we assume, implicitly, that noticing this characteristic is commonly known among them? Part of the answer lies in perception theory. For example, people tend to notice a person's race and gender, and we can safely assume that they believe others do, and that they believe others believe others do, etc. Partly, conspicuity can be endogenous. For example, when Gorbachev announced a new political era, he made everyone aware of a new characteristic of the game (and this awareness was commonly known as the announcement was public). Thus, the language the players use to conceptualize the statistical game is partly determined by psychology, partly by norms and strategic choices. Our model does not address the language selection problem. Rather, it assumes that the language (of x and y) is given, and studies statistical equilibrium selection within the model defined by this language.

7.4 Role Models

Our analysis can explain how precedents affect the perceived importance of gender, ethnicity, or race. Consider, for example, Barack Obama being the first non-white candidate to be elected president of the US. There are many ways to model the game involved, which includes interaction among candidates, voters, donors, volunteers, etc. But, whatever is the formulation of the game analyzed, the probability of Obama being elected is crucial to players' strategic considerations. It stands to reason that this probability be estimated by past cases, and this raises the question: should one consider race as a relevant characteristic for this estimation?

In our model, we can think of the characteristic x denoting the fact that the candidate is non-white, and y – denoting success (being elected). Before 2008 there had been relatively few attempts by non-whites to win the presidency, and they all failed. At that point, using the notation of Section 5, L, W, which stand for failures and successes of white candidates, respectively, were large, as was L/W, whereas l was not too large and $w = 0$.[15] The proof of Proposition 2 suggests that under these condition the variable "being non-white" was considered relevant, and non-white candidates were perceived as having a very low probability of success, as they were compared to the history of non-whites, containing no successes. However, by the proof of Proposition 3, the single case of Barack Obama rendered the variable "non-white" unimportant. Thus, if one computes the SNEs of an appropriately defined game, before 2008 race was considered relevant for prediction and for equilibrium selection, but after Obama's election it ceased to matter. More generally, our results can explain the importance of role models, and why a few precedents in which females or minority groups obtain a certain position may signals to others of the same position that they have a better chance of obtaining it than their group's statistics indicate.[16]

[15] If by an "attempt" we consider a person running in the primaries of one of the major parties, L/W might be in the order of magnitude of 10, whereas l would be 5 or 6, depending how far one goes back in history.

[16] See Chung (2000), Bayer and Rouse (2016).

Appendix: Proofs

For the following proofs, it is useful to define $\Delta(L, W, l, w) \equiv SSE(s_x) - SSE(s_0)$, where $\Delta(L, W, l, w) > 0$ implies that the variable x should not be included in the empirically optimal similarity function, whereas $\Delta(L, W, l, w) < 0$ implies that it should. Clearly, $\Delta(L, W, l, w) = \Delta(W, L, w, l)$ and $\Delta(L, W, l, w) = \Delta(l, w, L, W)$, as the SSE calculations do not change if we switch between 0 and 1 either for a predictor x or for the predicted variable y[17].

Proof of Proposition 1:

We need to show that

(i) If $L < W$, $\Delta(L, W, 1, 0) < 0$ and $\Delta(L, W, 0, 1) > 0$;

(ii) If $L > W$, $\Delta(L, W, 1, 0) > 0$ and $\Delta(L, W, 0, 1) < 0$;

(iii) $\Delta(L, L, 1, 0), \Delta(L, L, 0, 1) > 0$.

We first show that $\Delta(L, W, 1, 0)$ is positive for $L \geq W$ and negative for $L < W$. By symmetry, this implies that $\Delta(L, W, 0, 1)$ is positive for $L \leq W$ and negative for $L > W$, together completing the proof.

The SSE's are given by $SSE(s_0) = W\left(1 - \frac{W-1}{L+W}\right)^2 + (L+1)\left(-\frac{W}{L+W}\right)^2$ and $SSE(s_x) = W\left(1 - \frac{W-1}{L+W-1}\right)^2 + L\left(-\frac{W}{L+W-1}\right)^2 + 0.25$ (where the sub-database for which $x = 1$ yields $SSE = \frac{1}{4}$).

It follows that $\Delta(L, W, 1, 0)$ is equal to:

$$\frac{L^4 + L^3(4W-2) + L^2(2W^2+2W+1) + L(2W-4W^3+6W^2) - 3W^4 + 2W^3 + 5W^2 - 4W}{4(L+W-1)^2(L+W)^2} \tag{3}$$

The denominator of expression (3) is positive. Let $a(L, W)$ denote the numerator. First, we observe that $a(L, L) = 4L(2L^2 + 2L - 1) > 0$.This establishes Part (iii), and will also be a useful benchmark for Parts (i) and (ii). Indeed, to prove that $a(L, W) > 0$ (and thus that $\Delta(L, W, 1, 0) > 0$) for $L > W$, we will consider the partial derivative of $a(L, W)$ relative to its

[17]Whenever needed, we use partial derivatives to derive inequalities. In doing so we obviously extend the definition of the function $\Delta(L, W, l, w)$ to all non-negative real numbers (L, W, l, w) by the function's algebraic formula, whenever well-defined.

first argument, and show that it is positive for $L \geq W$. (Clearly, $a(L, W)$ is a polynomial in its two arguments, and it is well-defined and smooth for all real values of (L, W).) To see this, observe that $\frac{\partial a(L,W)}{\partial L}$ is equal to:

$$4L^3 + (12W - 6)\, L^2 + \left(4W^2 + 4W + 2\right) L + \left(-4W^3 + 6W^2 + 2W\right). \quad (4)$$

Since $W > 2$ implies $12W - 6 > 0$, the only negative term in (4) is $-4W^3$. However, for $L \geq W$ it is true that $4LW^2 - 4W^3 \geq 0$ and thus, for $L \geq W$ we have $\frac{\partial a(L,W)}{\partial L} > 0$. Because, for $L \geq W$, $a\,(L, W)$ is strictly increasing in L and $a(L, L) > 0$, we also have $a(L, W) > 0$ for $L > W$.

We now turn to the case $L < W$, where expression (4) might be negative (and, indeed, will become negative if L is held fixed and $K \to \infty$.) Again the strategy of the proof is to use direct evaluation at a benchmark and partial derivative arguments beyond, though a few special cases will require attention. The benchmark we use is the case $W = L + 1$. Here direct calculations yield $a\,(L, L + 1) = -4L\,(2L^2 - 1) < 0$.

This time we consider the partial derivative of $a\,(L, W)$ w.r.t. to its second argument, and would like to establish that it is negative. If it were, increasing K from $(L + 1)$ further up would only result in lower values of $a\,(L, W)$, and therefore the negativity of $a\,(L, W)$ (and of $\Delta\,(L, W, 1, 0)$) for $L < W$ would be established.

Consider, then,

$$\begin{aligned}
\frac{\partial a\,(L, W)}{\partial W} &= 4L^3 + 4L^2W + 2L^2 - 12LW^2 + 12LW + 2L - 12W^3 + 6W^2 + 10W - 4 \\
&= 4L^3 + (4W + 2)\, L^2 + \left(12W - 12W^2 + 2\right) L + \left(6W^2 - 12W^3 + 10W - 4\right) \\
&< 4W^3 + (4W + 2)\, W^2 + 12W^2 + 2W - 12LW^2 + 6W^2 - 12W^3 + 10W - 4 \\
&< 4W^3 + (4W + 2)\, W^2 + 12W^2 + 2W + 6W^2 - 12W^3 + 10W - 4 \\
&= -4\left(-3W - 5W^2 + W^3 + 1\right) \qquad (5)
\end{aligned}$$

where the first inequality follows from the fact that $L < W$ and the second from the fact that $L, W > 0$.

We now observe that expression (5) is negative for $W \geq 6$, and thus the partial derivative $\frac{\partial a(L,W)}{\partial W}$ is indeed negative for all $W \geq 6$, $L < W$. Coupled with the fact that $a\,(L, L + 1) < 0$, we obtain $a\,(L, W) < 0$ for all $W \geq 6$ (and $2 < L < W$).

We now wish to show that $a(L,W) < 0$ holds also for lower values of W. However, as $W > L > 2$ only a few pairs of values (L,W) are possible: $(3,4)$,$(3,5)$,$(4,5)$. Direct calculation shows that $a(L,W)$ is negative for all these pairs. Specifically, $a(3,4) = -204$, $a(3,5) = -1,424$, and $a(4,5) = -496$. This concludes the proof of Parts (i) and (ii). □

Proof of Proposition 2:

Let there be given $l > 1$. We wish to prove that for any $L,W > 2$, $\Delta(L,W,l,0) < 0$ (where the case $l = 0$, $w > 1$ is obviously symmetric).

The SSE's are given by $SSE(s_0) = (L+l)\left(-\frac{W}{l+L+W-1}\right)^2 + W\left(1 - \frac{W-1}{l+L+W-1}\right)^2$ and $SSE(s_x) = L\left(-\frac{W}{L+W-1}\right)^2 + W\left(1 - \frac{W-1}{L+W-1}\right)^2$(where the sub-database for which $x^j = 1$ yields $SSE = 0$). Straightforward calculation yields

$$\Delta(L,W,l,0) = -Wl\frac{\left(L(W-2) + (W-1)^2\right)l + (L+W-1)(L(W-2) + W(W-1))}{(L+W-1)^2(l+L+W-1)^2}$$

which is clearly negative. □

For convenience, we prove Proposition 4 before Proposition 3.

Proof of Proposition 4

First, observe that if $w = \left[\frac{lW}{L}\right] = 0$, then the first result in part (i), namely that for databases $(L,W,l,0)$ the unique SNE is the one selected by s_0, follows directly from Proposition 1. To prove the rest of the Proposition, it will be convenient to extend the definition of Δ to real-valued arguments and use calculus. We will only resort to (first- and second- order) partial derivatives with respect to the last two arguments. Note that for positive integers L,W,l,w, the SSE formulae are

$$SSE(s_0) = (L+l)\frac{(W+w)^2}{(L+W+l+w-1)^2} + (L+l)^2\frac{W+w}{(L+W+l+w-1)^2}.$$

$$SSE(s_x) = LW\frac{L+W}{(L+W-1)^2} + lw\frac{l+w}{(l+w-1)^2}$$

It is therefore natural to define, for positive integers L,W, and any $l,w \in \mathbb{R}$,

$$\Delta(L,W,l,w) = LW\frac{L+W}{(L+W-1)^2} + lw\frac{l+w}{(l+w-1)^2}$$
$$-(L+l)\frac{(W+w)^2}{(L+W+l+w-1)^2} - (L+l)^2\frac{W+w}{(L+W+l+w-1)^2}$$

as long as $l+w \neq 1-(L+W)$ and $w \neq 1-l$. Clearly, the function Δ is a rational function in its four arguments, and apart from these points of singularity, it is well-defined and smooth. Note that we are interested in l,w that are positive integers, hence $l,w \geq 1$. In particular, $l+w \geq 2$ while $1-(L+W) < -3$ and $w \geq 1$ while $1-l \leq 0$, so that none of the two singular points of Δ is within or even on the boundary of the range of values that is of interest to the statement of the proposition, apart from the special case discussed in the first paragraph of this proof. However, these points will prove useful in analyzing the function.

Next, because our focus is on the behavior of Δ as we change its fourth argument, starting from the critical point $w = \frac{lW}{L}$, it will simplify notation if we shift the fourth variable to center it around that point. Formally, let $\omega \in R$ and define a function $b : \mathbb{Z}_+^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$b(L,W,l,\omega) = \Delta\left(L,W,l,\frac{lW}{L}+\omega\right).$$

The statements in the Proposition are about the value of the $\Delta(\cdot)$ function evaluated at points where the third argument is a positive integer and the fourth argument is an integer larger or equal than $\left[\frac{lW}{L}\right]$. It is therefore useful to notice that for any positive integers L,W,l, and integer z we can write

$$\Delta\left(L,W,l,\left[\frac{lW}{L}\right]+z\right) = \Delta\left(L,W,l,\frac{lW}{L}+\varepsilon+z\right) = b(L,W,l,z+\varepsilon) \tag{6}$$

where $\varepsilon = \left[\frac{lW}{L}\right] - \frac{lW}{L}$. Note that $\varepsilon \in [-0.5,0]$ if $\left[\frac{lW}{L}\right] = \left\lfloor\frac{lW}{L}\right\rfloor$ and $\varepsilon \in [0,0.5)$ if $\left[\frac{lW}{L}\right] = \left\lceil\frac{lW}{L}\right\rceil$.

We prove the proposition as follows:

(1) We first show that $b(L,W,l,\omega)$ is strictly decreasing in ω for $\omega \geq 1$ (Lemma 1);

(2) Next, we prove that $b(L, W, l, \omega)$ has a limit as $\omega \to \infty$ and that it is a negative number (Lemma 2);
(3) Direct calculation shows that $b(L, W, l, 1.5) > 0$, and from this we conclude that, as a function of ω, $b(L, W, l, \omega)$ has a unique root larger than 1.5 (Lemma 3);
(4) We prove that $b(L, W, l, \omega) > 0$ for $\omega \in [-0.5, 1.5]$ if $\left[\frac{lW}{L}\right] \geq 1$, and for $\omega \in [0.5, 1.5]$ if $\left[\frac{lW}{L}\right] = 0$ (Lemma 4);
(5) Next, we show that $\frac{\partial b(L,W,l,\omega)}{\partial l} > 0$ for $\omega \geq 2$ (Lemma 5);
(6) We then show that, for all $l' > l > 1$, $\widetilde{w} > \frac{l'W}{L}$, if $\Delta(L, W, l, \widetilde{w}) \geq 0$ then $\Delta(L, W, l', \widetilde{w}) \geq 0$ (Lemma 6).

Before we proceed to formally state and prove these lemmas, let us explain why they prove the result:

Part (i) follows from (4). For $\left[\frac{lW}{L}\right] \geq 1$ we need to show that (for all $L, W > 2$, $l > 0$), we have $\Delta(L, W, l, w), \Delta(L, W, l, w+1) > 0$. In terms of the function b, $\Delta(L, W, l, w) = b(L, W, l, \varepsilon)$ and $\Delta(L, W, l, w+1) = b(L, W, l, \varepsilon + 1)$. Thus we have to show that $b(L, W, l, \varepsilon), b(L, W, l, \varepsilon + 1) > 0$ where $\varepsilon = \left[\frac{lW}{L}\right] - \frac{lW}{L} \in [-0.5, 0.5)$. Clearly, this follows from Lemma 4. Similarly, for $\left[\frac{lW}{L}\right] = 0$ we need to show that (for all $L, W > 2$, $l > 0$), we have $\Delta(L, W, l, w+1) = b(L, W, l, \varepsilon + 1) > 0$, where $\varepsilon = \left[\frac{lW}{L}\right] - \frac{lW}{L} \in [-0.5, 0.5)$. Clearly, this also follows from Lemma 4.

Part (ii) follows from (1) and (3) because b is a smooth function of ω in the range $\omega \geq 1$.

Part (iii) follows from (6): If l' is such that $\left[\frac{l'W}{L}\right] \geq w^*(L, W, l) - 2$, the claim follows from the fact that $w^*(L, W, l') \geq \left[\frac{l'W}{L}\right] + 2$. Thus we focus on the case $\left[\frac{l'W}{L}\right] < w^*(L, W, l) - 2$.

Using part (i) and the definition of w^*, $\Delta(L, W, l, q) \geq 0$ for any integer q such that $0 \leq q \leq w^*(L, W, l) - 1$. Claim (6) implies that for the same values of q, $\Delta(L, W, l', q) \geq 0$. It follows that the smallest integer w'' ($w'' > \left[\frac{l'W}{L}\right]$) for which $\Delta(L, W, l', w'')$ becomes negative is greater or equal than $w^*(L, W, l)$ and thus $w^*(L, W, l') \geq w^*(L, W, l)$.

Finally, for Part (iv), assume that W/L is an integer, and consider integers $l' > l > 1$. Let $w = \left[\frac{lW}{L}\right]$ and $w' = \left[\frac{l'W}{L}\right]$, that is, $w = \frac{lW}{L}$ and $w' = \frac{l'W}{L}$ as these are integers. Lemma 5 implies that, if $b(L, W, l, \omega) = \Delta(L, W, l, w + \omega) > 0$ for $\omega \geq 2$, then $b(L, W, l', \omega) = \Delta(L, W, l', w' + \omega) > 0$ (for the same ω). It follows that the smallest integer $\omega > 1$ for which

$\Delta(L, W, l', w' + \omega)$ becomes negative is bigger than that for which $\Delta(L, W, l, w + \omega)$ becomes negative, thus $w^*(L, W, l') - w' \geq w^*(L, W, l) - w$.

We start by providing the explicit formula for $b(L, W, l, \omega)$:

$$b(L, W, l, \omega) = \frac{LW(L+W)}{(L+W-1)^2} + \frac{l(lW+L\omega)[l(L+W)+L\omega]}{[lW+L(l+\omega-1)]^2} - \frac{(l+L)(lW+LW+L\omega)(lL+L^2+lW+LW+L\omega)}{(-L+lL+L^2+lW+LW+L\omega)^2} \quad (7)$$

This is a rational function in ω, with two vertical asymptotes where either the denominator of the first term or the denominator of the third term in 7 vanishes. We denote these singular points by $\underline{\omega}$ and $\overline{\omega}$, respectively:

$$\overline{\omega} = 1 - \frac{l(L+W)}{L} = 1 - l - \frac{lW}{L} < 0$$
$$\underline{\omega} = 1 - \frac{(l+L)(L+W)}{L} < \overline{\omega}$$

Thus, for $\omega > \overline{\omega}$, $b(L, W, l, \omega)$ is a smooth function.

We can now establish:

Lemma 1 *$b(L, W, l, \omega)$ is strictly decreasing in ω for $\omega \geq 1$.*

Proof: Differentiate $b(L, W, l, \omega)$ with respect to ω:

$$\frac{\partial b(L,W,l,\omega)}{\partial \omega} = \frac{(2L(l+L)(lW+L(W+\omega))(l(L+W)+L(L+W+\omega)))}{(L^2+lW+L(-1+l+W+\omega))^3} - \frac{(L(l+L)(l(L+2W)+L(L+2(W+\omega))))}{(L^2+lW+L(-1+l+W+\omega))^2} + \frac{(lL^2(-2lW+l^2(L+W)+lL(-1+\omega)-2L\omega))}{(lW+L(-1+l+\omega))^3}$$

The above expression can be rewritten as

$$-\frac{L^3\left[z_0(L,W,l)+z_1(L,W,l)\omega+z_2(L,W,l)\omega^2+z_3(L,W,l)\omega^3+z_4(L,W,l)\omega^4\right]}{(lW+L(l+\omega-1))^3(L^2+lW+L(l+W+\omega-1))^3} \quad (8)$$

where $z_0(L, W, l), z_1(L, W, l), z_2(L, W, l), z_3(L, W, l), z_4(L, W, l)$ are defined as:

$$
\begin{aligned}
z_0(L,W,l) \;=\;& -2l^4(L-W)(L+W)^3 - l^2L^2(L+W)^2(6+L(2L-9)-2W^2) \\
& -2l^3L(L+W)^2(L(2L-3)-2W^2) + L^4\left[2W - L(L+W-1)\right] \\
& +lL^3\left[L(2+3(L-2)L)+4W+6(L-2)LW+3(+L-2)W^2\right]
\end{aligned}
$$

$$
z_1(L,W,l) = L\left\{\begin{array}{c} L^3\left[(2(l-1)^4+4(l-1)^3L+(3-4l+2l^2)L^2\right] \\ +W\left[\begin{array}{c} 6(l-1)l(2-l+l^2)L^2+6(2l-1)(1-l+l^2)L^3 \\ +3(1-2l+2l^2)L^4+6lL(l+L)(1+l^2+lL)W \\ +2l(l+L)(2l+l^2+L+lL)W^2 \end{array}\right] \end{array}\right\}
$$

$$
z_2(L,W,l) = 3L^2\left\{2l^3W^2 + L\left[\begin{array}{c} (-2+4l-4l^2+2l^3)L+L^2\left[2-4l+3l^2+(l-1)L\right] \\ +[(4l(1-l+l^2)+2L+l(6l-4)L+(2l-1)L^2]W \\ +(3l^2+lL)W^2 \end{array}\right]\right\}
$$

$$
z_3(L,W,l) = L^3\left[L^3+2l(3l-2)W+L^2(-4+6l+W)+L(6-8l+6l^2-2W+6lW)\right]
$$

$$
z_4(L,W,l) = L^4(-2+2l+L)
$$

First, notice that L^3 and the denominator of expression (8) are strictly positive, hence the sign of (8) is equal to the opposite sign of the polynomial in ω on its numerator. Second, notice that $z_1(L,W,l)$, $z_2(L,W,l)$, $z_3(L,W,l)$, and $z_4(L,W,l)$ are strictly positive for all admissible values of $\{L,W,l\}$. It follows that the derivative of the polynomial in ω on the numerator of (8) is strictly positive for positive values of ω. Hence, if we can show that the polynomial is positive for some positive value of ω, then it is positive for all larger values of ω as well. Finally, we evaluate the polynomial at $\omega = 1$ and show that it is positive.

$$
\begin{aligned}
& z_0(L,W,l) + z_1(L,W,l)\,(1) + z_2(L,W,l)\,(1) + z_3(L,W,l)\,(1) + z_4(L,W,l)\,(1) \\
=\;& 2l(l+L)(L+W)^3[L^2+l^2W+lL(2+W)] > 0
\end{aligned}
$$

This allows us to conclude that $\frac{\partial b(L,W,l,\omega)}{\partial \omega} < 0$ for all $\omega \geq 1$. □□

Lemma 2 $\exists \lim_{\omega\to\infty} b\left(L,W,l,\omega\right) < 0$.

Proof:

$$\lim_{\omega\to\infty} b(L,W,l,\omega) = \frac{LW(L+W)}{(L+W-1)^2} + l - l - L = \frac{-L(L-1)^2-(L-2)LW}{(L+W-1)^2} < 0. \ \square$$

Lemma 3 *$b(L,W,l,\omega)$ has exactly one root in $\omega \in (1.5,\infty)$.*

Proof: We know that the singular points of b are negative. This means that for $\omega \geq 0$, $b(L,W,l,\omega)$ is a smooth function. Further, algebraic calculations[18] show that $b(L,W,l,1.5) > 0$ for all $L,W > 2$, $l > 0$. Since we established that $b(L,W,l,\omega)$ is negative for ω large enough, it has to have a root at some $\omega > 1.5$. Further, it is unique because b is strictly decreasing in ω over this range. $\square$

Lemma 4 *$b(L,W,l,\omega) > 0$ for $\omega \in [-0.5,1.5]$ if $\left[\frac{lW}{L}\right] \geq 1$, and for $\omega \in [0.5,1.5]$ if $\left[\frac{lW}{L}\right] = 0$.*

Proof: We need to consider two cases.

Case 1: $l = 1$

In this case, the vertical asymptotes are at $\underline{\omega} = -\frac{W}{L} - (W+L)$ and $\overline{\omega} = -\frac{W}{L}$ so for $\omega \geq -\frac{W}{L}$ the function is smooth. Algebraic calculations[19] show that for $l = 1$ and for all $L,W > 2$, $\frac{\partial b(L,W,l,\omega)}{\partial \omega}$ is strictly negative for all $\omega \geq -\frac{W}{L}$. This, together with the fact that $b(L,W,l,1.5) > 0$, proves that $b(L,W,l,\omega) > 0$ for $\omega \in (-\frac{W}{L}, 1, 5]$. If $\left[\frac{lW}{L}\right] \geq 1$, the fact that $-\frac{W}{L} < 0.5$ proves that $b(L,W,l,\omega) > 0$ for $\omega \in [-0.5,1.5]$. Similarly, if $\left[\frac{lW}{L}\right] = 0$ the fact that $-\frac{W}{L} < 0$ proves that $b(L,W,l,\omega) > 0$ for $\omega \in [0.5,1.5]$.

Case 2: $l > 1$

Algebraic calculations[20] show that $b(L,W,l,-0.5) > 0$ for all $l > 1$, $L,W > 2$ such that $\left[\frac{lW}{L}\right] \geq 1$, and that $b(L,W,l,0.5) > 0$ for all $l > 1$, $L,W > 2$ such that $\left[\frac{lW}{L}\right] = 0$. Consider first the case $\left[\frac{lW}{L}\right] \geq 1$. To study the sign of $b(L,W,l,\omega)$ for $\omega \in [-0.5,1.5]$ we observe that it is positive at $\omega = -0.5$ and at $\omega = 1.5$, and that it is continuous on the interval. Thus, to prove that it is positive throughout the interval it suffices to show that it has no roots in it.

[18] Available upon request. (Part (a) in the Appendix for referees)
[19] Available upon request. (Part (c) in the Appendix for referees)
[20] Available upon request. (Part (b) in the Appendix for referees).

Observe that $b(L,W,l,\omega)$ is a rational function in ω with a fourth degree polynomial (in ω) in its numerator. Every root of b is a root of this polynomial, and thus b can have at most four real roots. We claim that it has at least one real root in each of the following intervals:

(a) $(\underline{\omega},\overline{\omega})$, (b) $(\overline{\omega},-0.5)$, (c) $(1.5,\infty)$.

To see that there is a root in (a), observe that

$$\lim_{\omega\to_{+}\overline{\omega}} b(L,W,l,\omega) = \lim_{\omega\to_{-}\overline{\omega}} b(L,W,l,\omega)$$
$$= \frac{LW(L+W)}{(L+W-1)^2} - \frac{L^2 l(l-1)}{0} - \frac{L(L+l)(L+LW-Ll)(L+W+1)}{L^2(L+W)^2} = -\infty$$

$$\lim_{\omega\to_{+}\underline{\omega}} b(L,W,l,\omega) = \frac{LW(L+W)}{(L+W-1)^2} + \frac{l[-L(L+W+l-1)][-L(L+W-1)]}{L^2(L+W)^2}$$
$$-\frac{-L^2[l(L+2l-1)+l(l-1)]}{0^{+}} = +\infty$$

Thus, b, which is continuous over $(\underline{\omega},\overline{\omega})$, goes from $+\infty$ to $-\infty$ and has to cross 0 over the interval.

As for interval (b), observe, again, that $\lim_{\omega\to_{+}\overline{\omega}} b(L,W,l,\omega) = -\infty$ and that $b(L,W,l,-0.5) > 0$. Finally, it was established in Lemma 3 that there is a root in (c).

We can now consider the interval of interest, $[-0.5,1.5]$. We know that b is positive at the two endpoints. If it were non-positive at some point over this interval, the numerator of b would have to have two roots in the interval – either two distinct roots or a multiple one. In either case, we would have a total of five real roots for a polynomial of degree 4, which is impossible, and thus we conclude that b is strictly positive throughout $[-0.5,1.5]$.

Next, consider the case $\left[\frac{lW}{L}\right] = 0$. We need to study the sign of $b(L,W,l,\omega)$ for $\omega \in [0.5,1.5]$. The proof is analogous the the one for the previous case. In particular, it has been shown that $b(L,W,l,\omega) > 0$ at the two endpoints of the interval and continuous over the interval. Moreover, $b(\cdot)$ has at least one real root in each of the following intervals: (a) $(\underline{\omega},\overline{\omega})$, (b)' $(\overline{\omega},0.5)$, (c) $(1.5,\infty)$. Since the numerator of $b(\cdot)$ can have at most four real roots, there are no roots in the interval $\omega \in [0.5,1.5]$ and the function is positive over the whole interval.□

Lemma 5 *$b(L,W,l,\omega)$ is strictly increasing in l for $\omega \geq 2$.*

Proof: The derivative of $b(L,W,l,\omega)$ w.r.t. l is:

$$L^3 \frac{\zeta_0(L,W,\omega)+\zeta_1(L,W,\omega)\,l+\zeta_2(L,W,\omega)\,l^2+\zeta_3(L,W,\omega)\,l^3}{(-L+lL+lW+L\omega)^3(-L+lL+L^2+lW+LW+L\omega)^3} \tag{9}$$

where $\zeta_0(L,W,\omega)$, $\zeta_1(L,W,\omega)$, $\zeta_2(L,W,\omega)$, $\zeta_3(L,W,\omega)$ are defined as:

$$\zeta_0(L,W,\omega) = L^3(\omega-1)\begin{pmatrix} L^3\omega^2 + W(4(\omega-1)^2\omega + W^2(2\omega-1) + 3W(1-3\omega+2\omega^2)) \\ +L^2(3(\omega-1)\omega^2 + W(2\omega(1+\omega)-1)) \\ +L\begin{pmatrix} 2(\omega-1)^2\omega(1+\omega) + W^2(\omega(4+\omega)-2) \\ +3W(1+\omega(-3+\omega+\omega^2)) \end{pmatrix} \end{pmatrix}$$

$$\zeta_1(L,W,\omega) = L^2\begin{pmatrix} W^2(12W(\omega-1)^2 + W^2(2\omega-3) + 6(\omega-1)^2(2\omega-1)) \\ +L^4(\omega-2)\omega + 3L^2(2(\omega-1)^2\omega^2 + 4W(\omega-1)^2(1+\omega) + W^2(\omega^2-3)) \\ +LW(-6+6W(\omega-1)^2(4+\omega) + 6\omega(4-4\omega+\omega^3) + W^2(-9+\omega(4+\omega))) \\ +L^3(6(\omega-1)^2\omega + W(-3+\omega(3\omega-4))) \end{pmatrix}$$

$$\zeta_2(L,W,\omega) = 3L(L+W)^2\begin{pmatrix} L\left(L(\omega-2)+2(\omega-1)^2\right)\omega \\ +W^2(2\omega-3) + W\left(4(\omega-1)^2 + L(\omega^2-3)\right) \end{pmatrix}$$

$$\zeta_3(L,W,\omega) = 2(L+W)^3(L(\omega-2)\omega + W(2\omega-3))$$

First, notice that L^3 and the denominator of expression (9) are strictly positive. Second, notice that $\zeta_0(L,W,\omega)$, $\zeta_1(L,W,\omega)$, $\zeta_2(L,W,\omega)$, $\zeta_3(L,W,\omega)$ are strictly positive for all admissible values of $\{L,W\}$ and $\omega \geq 2$. Since l is an integer, it follows that the polynomial in l on the numerator of (9) is strictly positive for $\omega \geq 2$. This allows us to conclude that $\frac{\partial b(L,W,l,\omega)}{\partial \omega} > 0$ for all $\omega \geq 2$. $\square$

Lemma 6 *For all $l' > l > 1$, $\widetilde{w} > \frac{l'W}{L}$, if $\Delta(L,W,l,\widetilde{w}) \geq 0$ then $\Delta(L,W,l',\widetilde{w}) \geq 0$.*

Proof: If $\widetilde{w} = \left[\frac{l'W}{L}\right]$ or $\widetilde{w} = \left[\frac{l'W}{L}\right]+1$, the conclusion $\Delta(L,W,l',\widetilde{w}) \geq 0$ follows from either Part (i) or Proposition 1.

Assume, then, that $\widetilde{w} \geq \left[\frac{l'W}{L}\right]+2 \geq \left[\frac{lW}{L}\right]+2$. Recall that $w = \left[\frac{lW}{L}\right]$ with $\varepsilon = \left[\frac{lW}{L}\right] - \frac{lW}{L}$ and denote $w' = \left[\frac{l'W}{L}\right]$, $\varepsilon' = \left[\frac{l'W}{L}\right] - \frac{l'W}{L}$. Next, let $\omega = \widetilde{w} - w$ and $\omega' = \widetilde{w} - w'$. Thus, $\widetilde{w} = w + \omega = lW/L + \varepsilon + \omega = w' + \omega' = l'W/L + \varepsilon' + \omega'$.

Clearly, as $l' > l$, we have $w' \geq w$ and therefore $\varepsilon' + \omega' \leq \varepsilon + \omega$. Note that $\omega, \omega' \geq 2$ and thus $\omega + \varepsilon, \omega' + \varepsilon' \geq 1$.

We assume that $\Delta(L, W, l, \widetilde{w}) = \Delta(L, W, l, w + \omega) = b(L, W, l, \omega + \varepsilon) \geq 0$ and need to show $\Delta(L, W, l', \widetilde{w}) = \Delta(L, W, l', w' + \omega') = b(L, W, l', \omega' + \varepsilon') \geq 0$. Indeed, $b(L, W, l, \omega + \varepsilon) \geq 0$, coupled with Lemma 5, implies that $b(L, W, l', \omega + \varepsilon) \geq 0$. Further, as $\omega' + \varepsilon' \leq \omega + \varepsilon$, Lemma 1 (with $\omega + \varepsilon, \omega' + \varepsilon' \geq 1$) implies that $b(L, W, l', \omega' + \varepsilon') \geq 0$, which completes the proof of the lemma. □□

Proof of Proposition 3

The proof relies on the analysis used to prove Proposition 4. Here, we prove only the first statement. The second holds by symmetry of the Δ function.

Let us denote by $\overline{w}$ the closest integer to $\frac{W}{L}$ ($= \frac{lW}{L}$ because we deal with the case $l = 1$), that is, $\overline{w} = \left[\frac{W}{L}\right]$.

We need to show that, for every $0 < w \leq \left\lfloor\frac{W}{L}\right\rfloor + 1$, $\Delta(L, W, 1, w) > 0$.

In (6) we had

$$\Delta\left(L, W, l, \left[\frac{lW}{L}\right] + z\right) = \Delta\left(L, W, l, \frac{lW}{L} + \varepsilon + z\right) = b(L, W, l, z + \varepsilon)$$

which, by setting $l = 1$, becomes

$$\Delta\left(L, W, 1, \left[\frac{W}{L}\right] + z\right) = \Delta\left(L, W, 1, \frac{W}{L} + \varepsilon + z\right) = b(L, W, 1, z + \varepsilon)$$

For $0 < w \leq \left\lfloor\frac{W}{L}\right\rfloor + 1$, denoting $z = w - \overline{w}$ we have $w = \overline{w} + z = \left[\frac{W}{L}\right] + z$. We can then write

$$\Delta(L, W, 1, w) = \Delta\left(L, W, 1, \left[\frac{W}{L}\right] + z\right) = \Delta\left(L, W, 1, \frac{W}{L} + \varepsilon + z\right) = b(L, W, 1, z + \varepsilon)$$

where $\varepsilon = \left[\frac{W}{L}\right] - \frac{W}{L} \in [-0.5, 0.5)$ and $z \in \{1 - \left[\frac{W}{L}\right], ..., 1\}$ if $\left[\frac{W}{L}\right] = \left\lfloor\frac{W}{L}\right\rfloor$ and $z \in \{1 - \left[\frac{W}{L}\right], ..., 0\}$ if $\left[\frac{W}{L}\right] = \left\lfloor\frac{W}{L}\right\rfloor + 1$.

Denoting the fourth argument of b by $\omega = z + \varepsilon$, we observe that, because $z \geq 1 - \left[\frac{W}{L}\right]$, $\omega \geq 1 - \frac{W}{L}$. Further, as $z \leq 1$ and $\varepsilon < 0.5$, $\omega < 1.5$. Thus, it suffices to show that $b(L, W, 1, \omega) > 0$ for $\omega \in \left[-\frac{W}{L} + 1, 1.5\right]$. We know that $b(L, W, 1, \omega)$ is continuous and differentiable for $\omega > -\frac{W}{L}$, that

$\frac{\partial b(W,L,1,\omega)}{\partial \omega} < 0$ for all $\omega \geq -\frac{W}{L}$, and that $b(L, W, 1, 1.5) > 0$. Therefore, $b(L, W, 1, \omega) > 0$ for all $\omega \in \left[-\frac{W}{L} + 1, 1.5\right]$. This concludes the proof. $\square$

References

[1] Akaike, H. (1954), "An Approximation to the Density Function", *Annals of the Institute of Statistical Mathematics*, **6**: 127-132.

[2] Argenziano, R. and I. Gilboa (2012), "History as a Coordination Device", *Theory and Decision*, **73**: 501-512.

[3] Argenziano, R. and I. Gilboa (2019), "Learning What is Similar: Precedents and Equilibrium Selection", *PNAS,* **116**.

[4] Aumann, R. (1974), "Subjectivity and Correlation in Randomized Strategies", *Journal of Mathematical Economics*, **1**: 67-96.

[5] Bayer, A. and Rouse, C. E. (2016) "Diversity in the Economics Profession: A New Attack on an Old Problem", *Journal of Economic Perspectives*, **30:** 221–242.

[6] Billot, A., I. Gilboa, D. Samet, and D. Schmeidler (2005), "Probabilities as Similarity-Weighted Frequencies", *Econometrica*, **73**: 1125-1136.

[7] Brown, G. W. (1951), "Iterative Solutions of Games by Fictitious Play", In *Activity Analysis of Production and Allocation*, T. C. Koopmans (Ed.), New York: Wiley.

[8] Carlsson, H. and Van Damme, E. (1993), "Global Games and Equilibrium Selection", *Econometrica*, **61**: 989-1018.

[9] Cass, D. and K. Shell (1983), "Do Sunspots Matter?", *Journal of Political Economy*, **91**: 193-228.

[10] Chung, K. S. (2000) "Role Models and Arguments for Affirmative Action", *American Economic Review*, **90**: 640-648.

[11] Edmond, C. (2013) "Information Manipulation, Coordination, and Regime Change", *Review of Economic Studies*, **80**: 1422-1458.

[12] Gilboa, I., O. Lieberman, and D. Schmeidler (2006), "Empirical Similarity", *Review of Economics and Statistics*, **88**: 433-444.

[13] Halaburda, H., Jullien, B., & Yehezkel, Y. (2016), "Dynamic Competition with Network Externalities", *working paper.*

[14] Hardlë, W. and J. S. Marron (1985), "Optimal Bandwidth Selection in Non-Parametric Regression Function Estimation", *The Annals of Statistics*, **13**: 1465-1481.

[15] Harsanyi, J. C. and R. Selten (1988), *A General Theory of Equilibrium Selection in Games.* Cambridge: MIT Press.

[16] Hume, D. (1748), *An Enquiry Concerning Human Understanding.* Oxford: Clarendon Press.

[17] Medin. D. L. and M. M. Schaffer (1978), "Context Theory of Classification Learning", *Psychological Review*, **85**: 207-238.

[18] Mosley, L. (2005), "Dropping Zeros, Gaining Credibility? Currency Redenomination in Developing Nations." *mimeo.*

[19] Nagel, R. (1995), "Unraveling in Guessing Games: An Experimental Study", *American Economic Review*, **85**: 1313-1326.

[20] Nosofsky, R. M. (1984), "Choice, Similarity, and the Context Theory of Classification", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**: 104-114.

[21] Nosofsky, R. M. (1988), "Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**: 700-708.

[22] Nosofsky, R. M. (2011), "The Generalized Context Model: An Exemplar Model of Classification", in *Formal Approaches in Categorization*, Cambridge University Press, New York, Chapter 2, 18-39.

[23] Parzen, E. (1962), “On the Estimation of a Probability Density Function and the Mode”, *Annals of Mathematical Statistics*, **33**: 1065-1076.

[24] Rosenblatt, M. (1956), “Remarks on Some Nonparametric Estimates of a Density Function”, *Annals of Mathematical Statistics*, **27**: 832-837.

[25] Rosenthal, R. W. (1973), “A Class of Games Possessing Pure-Strategy Nash Equilibria”, *International Journal of Game Theory*, **2**: 65–67.

[26] Schelling, T. C. (1960), *The Strategy of Conflict*. Cambridge: Harvard University Press

[27] Schmeidler, D. (1973), “Equilibrium Points of Nonatomic Games”, *Journal of Statistical Physics*, **7**: 295-300.

[28] Selten, R. (1970), *Preispolitik der Mehrproduktenunternehmung in der Statischen Theorie* (First ed.). Springer Verlag, Berlin.

[29] Selten, R. (1978), “The Chain Store Paradox”, *Theory and Decision*, **9**: 127-159.

[30] Shepard, R. N. (1957), “Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space”, *Psychometrika*, **22**: 325-345

[31] Shepard RN (1987), "Toward a universal law of generalization for psychological science", *Science* **237**:1317–1323.

[32] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. London and New York: Chapman and Hall.

[33] Stahl, D. O. and P. W. Wilson (1995), “On Players’ Models of Other Players: Theory and Experimental Evidence”, *Games and Economic Behavior*, **10**: 213-254.

[34] Steiner, J., and C. Stewart, C. (2008), “Contagion through Learning”, *Theoretical Economics*, **3**: 431-458.

Appendix for Referees

a) Calculation of $b\,(L, W, l, 1.5) > 0.$

We evaluate the function $b\,(L, W, l, \omega)$ at $\omega = 1.5$ and find the following expression:

$$\frac{L * g\,(L, W, l)}{(-1 + L + W)^2(L + 2lL + 2lW)^2(L + 2lL + 2L^2 + 2lW + 2LW)^2}$$

where $g\,(L, W, l)$ can be expressed as a polynomial in W :

$$\begin{aligned}
& g\,(L, W, l) \\
= \; & (32l^4 + 64l^3L + 32l^2L^2)W^5 \\
& +16l\left[2L^3 + l^3(8L - 1) + 2l^2L(1 + 8L) + lL^2(5 + 8L)\right]W^4 \\
& +4lL\left[12l^3(4L - 1) + 3lL^2(21 + 16L) + L^2(4 + 27L) + 8l^2(-1 + 3L + 12L^2)\right]W^3 \\
& +2L^2\left[\begin{array}{c} -3(L-2)L^2 + 8l^4(8L - 3) + 16l^3(-2 + 3L + 8L^2)+ \\ 2l^2(-6 - 6L + 69L^2 + 32L^3) + 2lL(6 - L + 33L^2) \end{array}\right]W^2 \\
& +\left[\begin{array}{c} 16l^4(2L - 1) + 3L(2 + 5L - 4L^2) + 32l^3(-1 + L + 2L^2) \\ +4l^2(-3 - 12L + 29L^2 + 8L^3) + 4l(-6 + 15L - 14L^2 + 17L^3) \end{array}\right]L^3W \\
& -3L^4\,(L - 1)^2\,(3 + 2L) + 12l^2L^4\,(L - 1)^2 + 12lL^4\,(L - 1)^3
\end{aligned}$$

Notice that for $W, L > 2$ and $l > 0$ the terms multiplying W^5, W^4, and W^3 are positive. The terms multiplying W^2 and L^3W and the constant are polynomials in l. For $l > 0$, all three are increasing in l, as the coefficients of the positive powers of l are positive. Moreover, all three are positive when evaluated are $l = 1$, hence for all $l > 1$ as well. . In particular, the coefficient of W^2 evaluated at $l = 1$ is equal to $-68+112L+270L^2+127L^3 > 0$. The coefficient of L^3W evaluated at $l = 1$ is equal to $-84 + 82L + 139L^2 + 88L^3 > 0$. Finally, the constant evaluated at $l = 1$ is equal to $3L^4\,(2L - 3)\,(L - 1)^2 > 0$.

We have proved that $g\,(L, W, l) > 0$. Since $\frac{L}{(-1+L+W)^2(L+2lL+2lW)^2(L+2lL+2L^2+2lW+2LW)^2} > 0$, this concludes the proof.

b) Calculation of $b(L, W, l, -0.5) > 0$ for $l > 1$ and $\left[\frac{lW}{L}\right] \geq 1$.

We evaluate the function $b(L, W, l, \omega)$ at $\omega = -0.5$ and find the following expression:

$$\frac{-L * h(L, W, l)}{(-1 + L + W)^2(-3L + 2lL + 2lW)^2(-3L + 2lL + 2L^2 + 2lW + 2LW)^2}$$

where $h(L, W, l)$ can be expressed as a polynomial in L :

$$\begin{aligned}
& h(L, W, l) \\
= \; & (20l - 18) L^7 + \left[-4l^2(8W - 5) + l(-76 + 92W) + 45 - 36W\right] L^6 \\
& + \left[-36 + 4(23 - 10l)l + (81 - 4l(90 + l(-75 + 16l)))W - 2(9 - 78l + 64l^2)W^2\right] L^5 \\
& + \left[\begin{array}{c} 9 - 36l + 20l^2 + (-54 + 388l - 528l^2 + 224l^3 - 32l^4)W \\ +(36 - 492l + 780l^2 - 256l^3)W^2 + (116l - 192l^2)W^3 \end{array}\right] L^4 \\
& + \left[\begin{array}{c} -4l(18 - 43l + 24l^2 - 4l^3) - 4l(-74 + 234l - 168l^2 + 32l^3)W \\ -4l(52 - 185l + 96l^2)W^2 - 4l(32l - 8)W^3 \end{array}\right] WL^3 \\
& -8l^2W^2 \left[-19 + 56W - 30W^2 + 4W^3 + l^2(-6 + 24W) + l(24 - 84W + 32W^2)\right] L^2 \\
& -16l^3W^3 \left[6 - 3l + (8l - 14)W + 4W^2\right] L - 16l^4W^4(2W - 1)
\end{aligned}$$

In what follows, we prove that $h(L, W, l) < 0$ for all $l > 0$ and $L, W > 2$. The constant term is negative. The coefficient of L is negative because it is the product of a negative term and a quadratic expression in W with a positive coefficient on the square which is positive and increasing at $W = 2$, hence for any larger W too. Similarly, the coefficient of L^2 is negative because it is the product of a negative term and a quadratic expression in l with a positive coefficient on the square which is positive and increasing at $l = 2$, hence for any larger l too.

The coefficient of L^3 is the product of W, which is positive, and a third degree polynomial in W which can be shown to be negative in the relevant range. In particular, the polynomial has a negative coefficient on the third and second power. At $W = 2$, this polynomial is equal to $-56l + 236l^2 - 288l^3 - 240l^4$ which is negative for all $l > 1$. Moreover, its derivative at $W = 2$ is equal to $-152l + 488l^2 - 864l^3 - 128l^4$ which is also negative for all $l > 1$. Finally, the fact that this derivative is negative $W = 2$ implies that it is also negative for all values of $W > 2$, because the negative coefficients on the third and second powers of W guarantee that the function is concave in W for positive W.

The coefficient of L^4 is a third degree polynomial in W which can be shown to be negative in the relevant range ($l > 1$, $W > 2$). The polynomial has a negative coefficient on the third power. Evaluated at $W = 2$, it takes value $45-300l+548l^2-576l^3-64l^4 < 0$ for all $l > 1$. Moreover, its derivative w.r.t. W evaluated at $W = 2$ is equal to $90 - 188l + 288l^2 - 800l^3 - 32l^4$ which is also negative for all $l > 1$. Finally, its second derivative w.r.t. W is equal to $-8\left(-9 + 123l - 195l^2 + 64l^3 + (144l - 87)lW\right)$ which is negative at $W = 2$ and decreasing in W for all positive values of W.

The coefficient of L^5 is a quadratic function of W with a negative coefficient on the square, which is negative and decreasing at $W = 3$, hence negative for all larger values of W too. The coefficient of L^6 is a quadratic function of l with a negative coefficient on the square, which is positive for $l = 2$ and negative for all larger values of l. The coefficient of L^7 is positive.

Since the coefficient L^7 is positive, and we want to prove that the whole polynomial in L is negative, we prove that the sum of the terms in L^7 and L^5 is negative.

First, notice that the condition $\frac{lW}{L} \geq \frac{1}{2}$ implies that $L \leq 2lW$, which in turn implies:

$$(20l - 18)\, L^7 < 4\,(20l - 18)\, L^5 l^2 W^2$$

which in turn implies that

$$\begin{aligned}
& (20l - 18)\, L^7 + \left[\begin{array}{c} -36 + 4(23 - 10l)l \\ +(81 - 4l(90 + l(-75 + 16l)))W - 2(9 - 78l + 64l^2)W^2 \end{array}\right] L^5 \\
< \;& 4\,(20l - 18)\, L^5 l^2 W^2 + \left[\begin{array}{c} -36 + 4(23 - 10l)l \\ +(81 - 4l(90 + l(-75 + 16l)))W - 2(9 - 78l + 64l^2)W^2 \end{array}\right] L^5 \\
= \;& \left[\begin{array}{c} (80l - 72)\, l^2 W^2 - 36 + 4(23 - 10l)l \\ +(81 - 4l(90 + l(-75 + 16l)))W - 2(9 - 78l + 64l^2)W^2 \end{array}\right] L^5 \\
= \;& \left[\left(92l - 40l^2 - 36\right) + \left(300l^2 - 64l^3 - 360l + 81\right) W + (-128l^2 + 236l - 90)W^2\right] L^5
\end{aligned}$$

The last expression is a quadratic in W which is negative for all $W > 2$. In particular, it has a negative coefficient on the square, hence it is concave. Evaluated at $W = 2$ it is equal to $-128l^3 + 48l^2 + 316l - 234 < 0$

for all $l > 1$. Moreover, its derivative evaluated at $W = 2$ is equal to $-64l^3 - 212l^2 + 584l - 279 < 0$ for all $l > 1$.

To conclude the proof that the whole polynomial in L is negative, we still need to address the fact that the coefficient of L^6 is positive at $l = 2$. In particular, we do so by proving that the sum of the terms in L^6 and L^4 is negative at $l = 2$. First, notice that the condition $\frac{lW}{L} \geq \frac{1}{2}$ implies that $L \leq 2lW$, which in turn implies:

$$\begin{aligned} & \left[-4l^2(8W-5) + l(-76+92W) + 45 - 36W\right] /_{l=2} L^6 \\ < \; & 4\left[-4l^2(8W-5) + l(-76+92W) + 45 - 36W\right] /_{l=2} L^4 l^2 W^2 \end{aligned}$$

which in turn implies that

$$\begin{aligned} = \; & \left[-4l^2(8W-5) + l(-76+92W) + 45 - 36W\right] /_{l=2} L^6 \\ & +L^4 \left[\begin{array}{c} 9 - 36l + 20l^2 + (-54 + 388l - 528l^2 + 224l^3 - 32l^4)W \\ +(36 - 492l + 780l^2 - 256l^3)W^2 + (116l - 192l^2)W^3 \end{array} \right] /_{l=2} \\ < \; & 4\left[-4l^2(8W-5) + l(-76+92W) + 45 - 36W\right] /_{l=2} L^4 l^2 W^2 \\ & +L^4 \left[\begin{array}{c} 9 - 36l + 20l^2 + (-54 + 388l - 528l^2 + 224l^3 - 32l^4)W \\ +(36 - 492l + 780l^2 - 256l^3)W^2 + (116l - 192l^2)W^3 \end{array} \right] /_{l=2} \\ = \; & \left(-216W^3 - 308W^2 - 110W + 17\right) L^4 < 0 \text{ for all } W > 2. \end{aligned}$$

This concludes the proof that $b(L, W, l, -0.5) > 0$ for $l > 1$.

Calculation of $b(L, W, l, 0.5) > 0$ **for** $l > 1$ **and** $\left[\frac{lW}{L}\right] = 0$.

We evaluate the function $b(L, W, l, \omega)$ at $\omega = 0.5$ and find the following expression:

$$\frac{L * \eta(L, W, l)}{(L + W - 1)^2 (-L + 2LW + 2lW + 2Ll + 2L^2)(-L + 2lW + 2Ll)}$$

where $\eta(L, W, l)$ can be expressed as a polynomial in L : *in* which all the coefficients, as well as the constant, are positive:

$$
\begin{aligned}
& \eta\left(L, W, l\right) \\
= \; & (12l - 2)\, L^7 + \left[W\left(4l - 4\right) + l\left(32W - 28\right) + 32l^2 W + 12l^2 + 3\right] L^6 \\
& + \left[64l^3 W + l^2 W\left(100W - 44\right) + l^2\left(28W^2 - 24\right) + W^2\left(6l - 2\right) + lW\left(30W - 72\right) + 20l + 7 \right. \\
& + \left[\begin{array}{c} 6l^4 W + l^3 W\left(156W - 96\right) + l^2 W^2\left(192W - 204\right) + 16l^2 W\left(l^2 - 1\right) + 12lW^3 \\ + lW^2\left(100l^2 - 60\right) + \left(44lW - 1\right) + l\left(12l - 4\right) + 2W\left(2W - 1\right) \end{array}\right] L^4 \\
& + \left[\begin{array}{c} l^4 W\left(128W - 16\right) + l^3 W^2\left(300W - 288\right) + 32l^3 W + l^2 W^3\left(128W - 228\right) \\ + 40l^2 W^2 + 20l^2 W + lW^3\left(84l^2 - 16\right) + lW\left(24W - 8\right) \end{array}\right] L^3 \\
& + \left[\begin{array}{c} l^4 W^2\left(192W - 48\right) + l^2 W^4\left(156l - 80\right) + l^3 W^3\left(100W - 288\right) + 64l^3 W^2 \\ + 32l^2 W^5 + 32l^2 W^3 + 8l^2 W^2 \end{array}\right] L^2 \\
& + \left[l^4 W^3\left(128W - 48\right) + l^3 W^4\left(64W - 96\right) + 32l^3 W^3\right] L + 16l^4 W^4\left(2W - 1\right)
\end{aligned}
$$

c) Calculation of $\frac{\partial b(L,W,l,\omega)}{\partial \omega} < 0$ for all $\omega \geq -\frac{W}{L}$ for the case $l = 1$

For $l = 1$, the $b\left(L, W, l, \omega\right)$ function and its derivative with respect to ω are

$$
\begin{aligned}
b\left(L, W, 1, \omega\right) = \; & \frac{LW(L+W)}{(L+W-1)^2} + \frac{L + W + L\omega}{W + L\omega} \\
& - \frac{(1+L)(W + LW + L\omega)(L + L^2 + W + LW + L\omega)}{(L^2 + W + LW + L\omega)^2}
\end{aligned}
$$

$$
\frac{\partial b\left(L, W, 1, \omega\right)}{\partial \omega} = \frac{-L^3 \phi\left(L, W, \omega\right)}{\left(W + L\omega\right)^2 \left(L^2 + W + LW + L\omega\right)^3}
$$

where $\phi\left(L, W, \omega\right)$ is the following cubic expression in ω in which all the coefficients, including the constant, are positive.

$$
\begin{aligned}
& \phi\left(L, W, \omega\right) \\
= \; & L^5 + 3L^3 W + 3L^4 W + 4LW^2 + 8L^2 W^2 + 4L^3 W^2 + 2W^3 + 4LW^3 + 2L^2 W^3 \\
& + \omega\left(3L^4 + 8L^2 W + 10L^3 W + 2L^4 W + 4LW^2 + 6L^2 W^2 + 2L^3 W^2\right) \\
& + \omega^2\left(4L^3 + 2L^4 + L^5 + 2L^2 W + 3L^3 W + L^4 W\right) + \omega^3 L^4
\end{aligned}
$$

The sign of the coefficients guarantees that the expression is positive, for all $\omega \geq 0$. To examine the sign of $\phi\left(L, W, \omega\right)$ for $w \in [-\frac{W}{L}, 0)$, notice that:

a) $\phi\left(L, W, -\frac{W}{L}\right) = L^2 (L + W)^3 > 0$

b) $\phi(L,W,0) = L^5 + 3L^3W + 3L^4W + 4LW^2 + 8L^2W^2 + 4L^3W^2 + 2W^3 + 4LW^3 + 2L^2W^3 > 0$

c)

$$\begin{aligned}
\frac{\partial\phi(L,W,\omega)}{\partial\omega} &= \left(3L^4 + 8L^2W + 10L^3W + 2L^4W + 4LW^2 + 6L^2W^2 + 2L^3W^2\right) \\
&\quad +2\omega\left(4L^3 + 2L^4 + L^5 + 2L^2W + 3L^3W + L^4W\right) + 3L^4\omega^2 \\
&\geq \left(3L^4 + 8L^2W + 10L^3W + 2L^4W + 4LW^2 + 6L^2W^2 + 2L^3W^2\right) \\
&\quad +2\omega\left(4L^3 + 2L^4 + L^5 + 2L^2W + 3L^3W + L^4W\right) \\
&> \left(3L^4 + 8L^2W + 10L^3W + 2L^4W + 4LW^2 + 6L^2W^2 + 2L^3W^2\right) \\
&\quad -2\frac{W}{L}\left(4L^3 + 2L^4 + L^5 + 2L^2W + 3L^3W + L^4W\right) \\
&= 3L^3(L+2W) > 0
\end{aligned}$$

where the first inequality follows from the fact that $3L^4\omega^2 \geq 0$ and the second from the fact that $\omega > -\frac{W}{L}$.

Hence we can conclude that $\phi(L,W,\omega)$ is positive and increasing in the whole interval $\left(-\frac{W}{L}, 0\right)$, hence the function $b(L,W,1,\omega)$ is decreasing for all $\omega > -\frac{W}{L}$.