The Predictive Success of Theories of Strategic Thinking: A Non-parametric Evaluation*

David Rojo Arjona
Chapman University

Daniel E. Fragiadakis
Villanova University

Ada Kovaliukaite
NYU Abu Dhabi

August 27, 2019

Abstract

Theories of non-equilibrium strategic thinking (e.g. Level-k and Cognitive Hierarchy) intend to describe how individuals actually behave. But how much of their descriptive accuracy is driven by being more permissive theories? We modify Selten [1991] axiomatic measure of predictive success to ensure individual consistency. By applying restrictions over observables to the individual data (echoing the revealed preference literature), we test the necessary and sufficient conditions of these theories and quantify the *economic losses* for deviations from the theory. The non-parametric results are favourable for these theories and show that their predictive success is not mechanically due to their permissiveness.

JEL classification: C14, C72, C91,
Keywords: Level-k, Cognitive Hierarchy, Predictive Success, Selten, Strategic Thinking, Behavioural Game Theory

*We are especially grateful to Matt Polisson for invaluable discussion, Sebastian Cortes Corrales for his proficient research assistance and Karl Schlag for his publicly available code of the z-test of Suissa and Shuster. This research has made use of the ALICE High Performance Computing Facility at the University of Leicester. We have benefited from the comments of Aurelien Baillon, Subir Bose, Jan Heufer, Dan Kovenock, Karl Schlag, Andre Stenzel, participants of the 17th SAET conference, seminar participants at Chapman, Indiana, Lausanne, Leicester, Rotterdam, Utrecht and Vienna. Any remaining errors are our own.

1 Introduction

There is increasing interest in how individuals *actually* reason in strategic situations. As a result of a body of evidence showing that the traditional concept of equilibrium fail to explain strategic behaviour (Camerer [2003]), this interest has translated into theories of *(non-equilibrium) strategic thinking* (Crawford et al. [2013]), where the rationality of players is not common knowledge. In models of this kind, players are categorized into types according to their reasoning level. An exogenous non-strategic type 0 will be the starting point for beliefs and behaviour of higher types. For example, one commonly used model of strategic thinking, Level-k (Nagel [1995], Stahl and Wilson [1995], Costa-Gomes and Crawford [2006]), pictures a type k best-responding only to a type $k-1$ player. So, type 1 best responds to type 0 choices, type 2 best responds to type 1 choices and so on. The related Cognitive Hierarchy (Camerer et al. [2004]) imposes a normalized Poisson distribution of lower types. So, type 1 best responds to type 0 choices, type 2 best responds to a (normalized Poisson) probability mix of type 0 and type 1 choices and so on. More generally, in a model of strategic thinking, a type k best responds to a convex combination of the choices of lower types. The *only* source for variations in predictions between models of strategic thinking arises due to the additional assumptions about this convex combination of lower types.

In this paper, we exploit that difference and provide a method to test the necessary and sufficient conditions of *any given* model of strategic thinking non-parametrically. Furthermore, using Selten [1991] index, we also evaluate the predictive success of a model, and correct the resulting passing rates of our test by incorporating the permissiveness of the model (i.e. how easy passing the test is for that particular model). Using this method, we evaluate the predictive success of Level-k, Cognitive Hierarchy and compare them with the most general model of strategic thinking, where no additional assumptions are imposed. This general model provides an upper bound for the predictive success of the theory of strategic thinking as a whole. Like previous results, we show that current models of strategic thinking are predictive compared to equilibrium. Furthermore, we are able to show that this predictive success is not mechanically due to being more permissive theories. More importantly, the upper bound of the strategic thinking theory does not cover all behaviour but the behaviour explained by the general model appears not to be statistically different from Level-k and Cognitive Hierarchy, when allowing for a small amount of errors.

The existing literature does not provide an evaluation of the predictive success for two reasons. First, current tests of Level-k and Cognitive Hierarchy are, at best, a joint test of the theory being right and some ancillary assumption. For example, many papers estimate competing models via maximum likelihood (e.g. Quantal Responses Equilibrium vs. Level-k) and see whether strategic thinking theories fit better (see Crawford et al. [2013] for a review). In many of these applications, extraneous and particular assumptions about noise need to be introduced because an observed action not predicted by the theory is enough to make the value of the likelihood function equal to 0 (uninformative). This problem of presenting joint tests also affects experimental designs developing some ex-ante testable hypothesis. Hargreaves Heap et al. [2014] present a test in which either the theory is wrong or the non-strategic type 0 depends on the strategic components of the game. Ideally, we would like a method precise enough to test the *necessary and sufficient conditions* of strategic thinking models and, at the

same time, flexible enough to test *any* potential model of strategic thinking.

Second, even if one were to take the results of these tests as prima facie evidence, alternative models of strategic thinking vary in terms of their permissiveness and, consequently, it should not be surprising that more permissive models mechanically describe behaviour better. There is no control for permissiveness in previous tests but we could form intuitive opinions about this trade-off. In some existing experiments using 3 x 3 normal form games, every possible choice is consistent with *some* type while, in other experiments, few choices would be predicted by any type (in particular, Level-k tests by Costa-Gomes and Crawford [2006], Fragiadakis et al. [2016]). If predictive success is taken seriously, however, we need a method producing a systematic evaluation rather than appealing to intuitions. The closest paper in this direction is Wright and Leyton-Brown [In Press] who use machine learning techniques to evaluate out-of-sample predictions of some strategic thinking models and look at the number of parameters of these models as a measure of permissiveness. Our main concern is that this definition of permissiveness can mislead evaluations. For example, Level-k would estimate $k-1$ parameters while Cognitive Hierarchy requires to estimate the parameter governing the Poisson distribution. However, as a consequence of that parameter, there are infinitely many more admissible distribution of lower levels in Cognitive Hierarchy than in Level-k, which may result into a larger set of predicted outcomes, depending on the game. We argue that, in line with methods coming from social sciences (Selten [1991]), a more appropriate method to account for permissiveness should evaluate how many of outcomes are consistent with the model (and the respective combinations of parameters) out of all possible outcomes.

In order to address these two problems, the method in this paper essentially represents the utility implications of every actual choice according to a particular candidate model (and its admissible distributions of lower types) into a system of inequalities — in the same fashion as Afriat [1972]. More concretely, for every observation, the utility associated to the observed choice must have been at least as high as the utility associated to every other alternative, according to that candidate model. Notice that if this condition is satisfied for the best-responses set, it will be necessarily satisfied for every other alternative. Thus, we can show that the candidate model is consistent with the observed behaviour *if and only if* the resulting system of inequalities has a solution. As a result, we test the necessary and sufficient conditions for the candidate model. Furthermore, as different models of strategic thinking only differ in the probabilities associated to lower types (and their associated actions) and these probabilities enter naturally in the calculation of the expected utility of a particular type, the method is flexible enough to test any alternative model of strategic thinking.

To evaluate predictive success, we follow Beatty and Crawford [2011] and their implementation of the measure of predictive success, built on the desirable properties proposed by Selten [1991]. In their paper, the fraction of individuals whose choices pass the test of the model are corrected by how easy passing the test was for that particular model. This ease to pass the test is measured by the relative size of the datasets consistent with the model with respect to all possible datasets that could have emerged. As Bronars [1987] noticed, this relative measure is inversely related to the power of the test of the model under the alternative hypothesis of uniformly random behaviour. Beatty and Crawford [2011] apply this method on demand data

and there is, at least, one important reason to adjust it to our game theory application.[1] By requiring that every individual choice is consistent with *one particular* type – rather than, for example, every individual choice being consistent with *some* type, we address concerns that appear in other implementations of the Selten measure of predictive success.[2]

As currently stands, the virtue of testing the necessary and sufficient conditions is also a limitation of our test in the presence of errors. One single error of any size is enough for an individual to fail the test. Consequently, the test can be too *stringent*. Fortunately, we can modify the system of inequalities by introducing an error parameter that decreases the payoff associated to the best responses of the model. By doing so, we can recover the lowest error ensuring that the system of inequalities is satisfied for the observed choices given a a particular model – in line with the ideas of the *critical cost efficiency index* (Afriat [1973]), or the less restrictive alternative by Varian [1990]. Thus, we can attribute an economic interpretation to this error parameter as the efficiency loss for deviations from the model.

We apply our method to the experimental data from Fragiadakis et al. [2017], which satisfy three conditions. First, we require individuals to make more than one choice in order to be able to measure individual (in)consistency across choices. Second, due to the theories of strategic thinking characterising players with a *fixed* type and distribution over lower types, the data should not provide opportunities for learning or updating of beliefs about other types.[3] Finally, the restrictions imposed on the observables via the system of inequalities depend exclusively on the unobserved theories of strategic thinking and, consequently, other commonly accepted sources of variation related to utility function (e.g. risk preferences and social preferences) need to be minimized. In particular, Fragiadakis et al. [2017] satisfy these conditions by having subjects in their experiment play a series of games without feedback between choices and using the binary lottery incentive mechanism introduced by Roth and Malouf [1979], which is agnostic about the functional form of the utility as long as utility is monotonically increasing in probabilities and compound lotteries can be reduced. The games used in their experiment are a modified version of the 11-20 game (Arad and Rubinstein [2012], Alaoui and Penta [2016]) that allows separation of types not only for level-k and Cognitive Hierarchy but also for any general model of strategic thinking. Fragiadakis et al. [2017] design a between-subjects experiment with an *actions treatment* to incentivise an *elicited beliefs treatment*. Their tests are set to deal with the beliefs data but not with the action data.

The results show that, in the stringent version of the test where no errors are allowed, Level-

[1] See Polisson et al. [2017] for an application of this implementation on contingency claims.

[2] Imagine an experiment where subjects make two decisions. In each decision, subjects need to choose an integer between "1" and "10". The theory under evaluation predicts "6" and "3" in the first and second decision, respectively. In that experiment, half of the subjects (*Group A*) choose "6" in the first decision and numbers other than "3" in the second decision. The other half of the subjects (*Group B*) choose "3" in the second decision but numbers other than "6" in the first decision. Thus, in each decision separately, the theory scores a Selten index of 0.4 (0.5 due to the fraction of subjects choosing consistently with the theory in that decision minus 0.1 of the precision of the theory in each decision). In many implementations, the Selten index would be averaged across games. As a result, an index of 0.4 would indicate some predictive success of the theory yet no single subject in this thought experiment is consistent with the theory - i.e., no subject chooses both numbers, "6" and "3".

[3] In the presence of feedback, learning models are probably more suited to explain the data (see Salmon [2001] for an evaluation of the power of experimental tests).

k and Cognitive Hierarchy describe the data better than equilibrium but worse than the general model. This difference is not significant, however. As soon as we allow subjects to exhibit small losses of efficiency, separations between the different models start to be significant. In fact, when we allow an economic loss of, at most, 10% (an arbitrary but conventional threshold), Equilibrium, Level-k, Cognitive Hierarchy and the generalized version of theories of strategic thinking can explain 0%, 2.5%, 17.5% and 35% of the subjects' behaviour, respectively.

These results are robust even after controlling for how easy is for each model to pass our test. More concretely, the results show that, for efficiency levels close to 1, the likelihood for these models to pass the test is generally close to 0. The generalized model present a likelihood of 0.18% for an efficiency loss of 10%. This implies that the power of the test for these theories is extremely high and close to 1 under the alternative hypothesis of uniformly random choices.

These results and the corresponding inference are based on Afriat's restrictive index, which punishes subjects for their worst deviations. Varian's index, which weighs every mistake equally, presents qualitatively the same results. Most notably, the frequency of subjects that can be explained by the theories of strategic thinking increases more rapidly with Varian's index and the gap between Level-k and Cognitive Hierarchy, and the general model is smaller. Furthermore, like with Afriat's index, these results of the more permissive index cannot be completely explained simply because of how easy is for these models to pass the test. All in all, these are good news for existing theories of strategic thinking.

The rest of the paper is organized as follows. Section II describes the method and the necessary and sufficient conditions of the test for each of the theories. Section III introduces the application and the details of the implementation of our method. Section IV presents the results and Section V discusses the wider implications of our method and results.

2 How to measure Predictive Success?

Selten [1991] investigates the desirable properties that a measure of predictive success, p, should have. He argues that it should appear uncontroversial to calculate p as a function of two elements. The first element is the *hit rate*, $r \in [0,1]$, which is defined as the relative frequency of correct predictions. The second element is the *area*, $a \in [0,1]$, which is defined as the relative size of the predicted outcomes with respect to all possible outcomes. Then, he proposes a list of desirable properties that a candidate functional forms, $p(r,a)$ should satisfy. For a cardinal characterisation, he proposes the following axioms:[4]

Monotonicity: $p(r_1,a) > p(r_2,a)$ for $r_2 < r_1$ and $p(r,a_1) > p(r,a_2)$ for $a_2 > a_1$
Equivalence: $p(0,0) = p(1,1)$
Aggregability: $p(\beta r_1 + (1-\beta)r_2, \beta a_1 + (1-\beta)a_2) = \beta p(r_1,a_1) + (1-\beta)p(r_2,a_2)$

These axioms have appealing interpretations in the context of finding a measure of predictive success. The monotonicity axiom states that, ceteris paribus, either a larger r (more accurate theory) or a smaller a (more precise theory) should lead to an increase of the measure

[4]Selten [1991] also proposes an alternative ordinal characterization where aggregability is replaced by continuity and cost-benefit evaluation axioms.

of predictive success. The equivalence (of irrelevant theories) axiom establishes that a completely accurate theory with no precision and a completely precise theory with no accuracy are equally (un)informative. The aggregability axiom allows different experiments to be compared with the same measure. A weighted arithmetic mean, where $\beta = \frac{N_1}{N}$, $1 - \beta = \frac{N_2}{N}$ and $N_1 + N_2 = N$, controlling for the different relative sizes of experiments, $\frac{N_i}{N}$, appears a natural method to aggregate hit rates. If one accepts this, the aggregability axiom imposes the same method to be used for the areas and, consequently, for the predictive measures.

Many functional forms satisfy these axioms. Selten and Krischker [1983] advocate for a very simple functional form which subtracts the area from the hit rate, $p = r - a \in [-1, 1]$. With these axioms and this particular measure p, the following theorem follows.

Theorem (Selten's Theorem)**.** *The function $p = r - a \in [-1, 1]$ satisfies monotonicity, equivalence and aggregability. If the function $\tilde{p}(r, a)$ also satisfies these axioms, there exist $\gamma_1 \in \mathbb{R}, \gamma_2 > 0 \in \mathbb{R}$ such that $\tilde{p}(r, a) = \gamma_1 + \gamma_2 \times p$.*

Proof. See Selten [1991], page 163; and the Appendix in Beatty and Crawford [2011] for an alternative proof. □

If the axioms are accepted to be desirable properties, the *Selten index* p is a candidate measure of predictive success because p satisfies those desirable properties. Furthermore, the second part of the theorem states that every other function satisfying the above axioms can be expressed as a linear transformation of the simple function p. Arguably, Selten's Theorem seems to plead for the use of $p = r - a$ given its simplicity with respect to other candidate measures.

We are not alone in considering these properties desirable and the Selten index meritorious. In fact, this measure has been widely applied in the analysis of individual choice data (in consumer demand: Beatty and Crawford [2011]; in lottery choices: Harless and Camerer [1994], Hey [1998]; in portfolio choices: Polisson et al. [2017]) and, to a lesser extent, in the analysis of choices in games (e.g., Van Huyck et al. [1997], Willinger and Ziegelmeyer [2001], Keser and Willinger [2007], Wang et al. [2010]). The application of this measure differs between some of the analysis of individual choices and the analysis of choices in games. In particular, influenced by the revealed preference literature, the application in individual choices have demanded individual consistency across decisions to claim that the dataset is also consistent with the theory. In contrast, most applications in games implicitly assume that different choices correspond to different experiments (even if choices are coming from the same individuals) and, by aggregability, predictive success is measured as the average Selten index across games.[5,6]

By abstracting from individual consistency across decisions, the process of averaging Selten indexes across games may overstate the predictive success of a theory. This concern can be easily presented through the following example. Imagine an experiment where subjects make two decisions. In each decision, subjects need to choose an integer between "1" and "10". The

[5]Exceptions are applications on datasets where subjects only make *one* decision. Applications on these datasets are uncommon, though (see Gächter and Riedl [2006] for an example).

[6]There is also a literature applying revealed preference methods to games in order to investigate individual consistency across decisions according to altruistic preferences (see, e.g., Andreoni et al. [2003],Andreoni and Miller [2002]) but they do not apply the Selten measure.

theory under evaluation predicts "6" and "3" in the first and second decision, respectively. In that experiment, half of the subjects (*Group A*) choose "6" in the first decision and numbers other than "3" in the second decision. The other half of the subjects (*Group B*) choose "3" in the second decision but numbers other than "6" in the first decision. Thus, in each decision, the theory scores a Selten index of 0.4 (0.5 due to the fraction of subjects choosing consistently with the theory in that decision minus 0.1 of the precision of the theory in each decision). As a result, the average Selten index across games associated to that theory is 0.4. This value supports the predictive success of the theory yet no single subject in this thought experiment is completely consistent with the theory - i.e., no subject chooses both numbers, "6" and "3".

The method proposed below addresses the problem illustrated in this simple example. Correcting overestimation is, however, not our only reason to look at individual consistency across choices. Not all deviations from the model have the same consequences; and two distinctive sets of individual choices may present different levels of efficiency (according to the evaluated model). The tests in revealed preferences are expressed in terms of the loss of efficiency generated by deviations with respect to the model. Incorporating this feature in calculating the Selten index provides a more *economic* interpretation to both the meaning of predicting accurately, $r = 1$, and to the distance from accurate predictions. The distance from correct predictions would indicate the economic losses of not choosing completely in line with the model. This measure of efficiency in our paper is expressed in actual payoffs rather than utility. Following previous expositions, we characterise theories of strategic thinking assuming players to be risk neutral, whose utility is linear in payoffs. The subsequent application, thus, requires to implement risk neutrality and the selected experimental data does so.

In the following subsections, we formally define the elements needed to test whether a *dataset* involving subjects playing a series of *games* satisfies the *consistency conditions* according to a *strategic thinking theory*; and how these elements are used in the construction of the *hit rate* and *area* to evaluate the predictive success of the theory.

2.1 Games and Dataset

We formally define the games that subjects see and how choices in a series of these games lead to a dataset.

Definition 1. *A normal form game, G, is a tuple $G =< N, (S_i), (\pi_i) >$, where $N = \{i, j\}$ is the set of two players, S_i is a finite set of strategies and $\pi_i : S_i \times S_j \to \mathbb{R}$ is the mapping from strategies to payoffs.*

We focus on two player normal form games because they are routinely used in experiments. In these games, subject i chooses a strategy $s_i \in S_i$ and, given the choices of player j, receives payoff π_i. Theories of strategic thinking, as formally defined next, are confined to explain initial instances of strategic interaction, where the relevant elements to determine choices are players, strategies and payoffs. These elements are precisely the components of normal form games. Hence, our method is restricted to these games.

Definition 2. *A dataset, $\mathcal{O}$, is a collection of a finite number of T observations such that $\mathcal{O} = \{(s^t, G^t)\}_{t=1}^{T}$.*

We define a dataset generally but it can be referred to a dataset of individual choices. We denote this as $\mathcal{O}_i$. In such dataset, the researcher observes player i facing game G and choosing $s_i \in S_i$ in period t, collecting T observations. We test strategic thinking theories in individual datasets.

2.2 Strategic Thinking Theories

Now, before defining formally theories of strategic thinking, we need to introduce some important concepts.

Definition 3. *A finite G-based epistemic type space is a tuple* $< (K_i), (m^i), (\hat{s}_i^j) >$ *where* $K_i = \{0, 1, 2..., k_i\}$ *is the set of types of players,* $m^i : K_i \to \Delta(K_j)$ *is the mapping from the type to the distribution of other types, and* $\hat{s}_i^j : K_j \to \Delta(S_j)$ *is the (mixed) strategy that a type of player i, with a type* k_i*, associates to other types.*

The type of a player, k_i, represents the number of lower types she can compute, as usually interpreted as an index of cognitive sophistication. Each player i, whose type is $k_i \in K_i$, uses a *strategic thinking function*, m^i, mapping from her type to a discrete distribution of the other player's type. Furthermore, the function $\hat{s}_i^j$ assigns a strategy for each type k_j by mapping each type to a probability distribution over the strategy space. Thus, the strategy that a player of type k_i associates to the other player, $\hat{s}_j(k_i)$, is the compound effect of m^i and $\hat{s}_i^j$. For convenience, we also include the following definition.

Definition 4. $f_k(h) \in [0, 1]$ *is the proportion of type h that a player of type k believes exists in the population.*

Then, the corresponding expected utility of player i whose type is $k_i > 0$ is defined as:

$$\Pi_i(s_i, k_i) = \sum_{k_{-i}} f_k\left(k_{-i}\right) \Pi_i\left(s_i, \hat{s}_j\left(k_i\right)\right). \tag{1}$$

The payoff function depends on the probability of the event of finding other types given a player's own type and the payoff associated to those events. We need to define some way in which a type of player computes that information to reach a choice.

Definition 5. *A strategy* $\tilde{s}_i\left(k_i\right)$ *is a best response for type* $k_i > 0$ *if* $\Pi\left(\tilde{s}_i, k_i\right) \geq \Pi\left(s', k_i\right) \forall s' \in S_i$*. The set of best responses is named* $BR(m^i(k_i))$*.*

The only remaining issue is to express a theory of strategic thinking as a function m^i. Theories of strategic thinking are comprised of the following list of assumptions:

Assumption 1. *Any type* $k_i > 0$ *player best responds to their beliefs over types.*

Assumption 1 states that theories of strategic thinking, unlike other behavioural game theory models (e.g., Quantal Response Equilibrium), retain optimization and are deterministic. The results of empirical tests of this assumption are not conclusive. There is positive evidence (e.g., Rey-Biel [2009]) but it is not universal (e.g., Costa-Gomes and Weizsäcker [2008]). We discuss further implications of this assumption for the reported results in Section 5.

Assumption 2. *For every type i, the strategy associated to player j whose type is $k_j = 0$, $\hat{s}_j(k_i)$ is exogenously given and independent of $m^i(k_i)$.*

Theories of strategic thinking do not include theories about their non-strategic element. However, two main alternative distributions over strategies for type 0 has been presented – either a disperse prior represented by a uniform random distribution or a distribution depending on the particular application.[7] In fact, Hargreaves Heap et al. [2014] have argued that the possibility to select the characterisation of type 0 depending on the application may reduce the predictive power of these theories. Our method is general enough to accommodate any characterisation of type 0 but, more importantly, any concern about the predictive success of the theory depending on type 0 can be controlled in our method through the area of the theory. The lack of theoretical foundations about type 0 also complicates the comparison between theories. Nonetheless, if this type is thought to follow a non-strategic, intrinsic reaction towards the game alone, the second part of Assumption 2 seems natural.

Assumption 3. $\sum_{h=0}^{k-1} f_k(h) = 1$.

This assumption has three noticeable features. First, it is another way to restate that the mapping $m^i : K_i \to \Delta(K_{-i})$ produces a discrete probability distribution. Second, it also includes the restriction that a player always believes that other players are lower types. Otherwise, we face epistemic problems. Finally, the distribution of (lower) types is a convex combination that sum up to exactly one. Thus, a player believes that *everyone else* is using the same mode of reasoning, with a lower type. Strategic thinking theories do not allow the sum to be lower than 1 – when only *a fraction* of the population would be explained by the strategic thinking theory – possibly because this leaves undefined what the mode of reasoning of the remaining fraction of the population is.

As a result of Assumption 1 – 3, the distribution of higher levels is anchored in a non-strategic type 0 via the best responses. Theories of strategic thinking differ in the particular functional form, m^i, which imposes restrictions on the proportions for each lower type, $f_k(h)$. We focus on the popular theories, Level-k, henceforth LK (Nagel [1995], Stahl and Wilson [1995], Costa-Gomes et al. [2001], Costa-Gomes and Crawford [2006]) and Cognitive Hierarchy, henceforth CH (Camerer et al. [2004]). However, in order to figure out the upper bound of the predictive success of strategic thinking theories, we also study a *generalized* version of these two theories (henceforth GLK).

Definition 6. *A function m^i is GLK if it satisfies Assumptions 1 - 3 and imposes no further restrictions.*

If one were to accept that theories of strategic thinking satisfy, at a minimum, the aforementioned assumptions, a GLK theory should provide indeed an upper bound to the predictive success of these theories because it satisfies these assumptions and nothing more. The popular theories, LK and CH, include the following additional restrictions with respect to GLK.

Assumption 4. $f_k(k-1) = 1$ *(LK assumption)*

[7] We come back to the particular characterisation of type 0 that we use in our application in Section 3.2

Definition 7. *A function m^i is LK if it satisfies Assumptions 1 - 4, and imposes no further restrictions.*

In LK, every player whose type is k_i assigns all the mass of the distribution of types to the immediate lower type. As a consequence, by using this restrictive model, it is very easy to infer the type from the action (Arad and Rubinstein [2012], Georganas et al. [2015]).

Assumption 5. $f_k(h) = \frac{g(h)}{\sum_{l=0}^{k-1} g(l)}$ *(CH assumption)*

where $g(h) = \frac{e^{-\tau} \times \tau^h}{h!}$ is the probability of observing realization h according to a Poisson distribution with the parameter τ, which is the mean and the variance of the distribution. Thus, the probability that a type k player believes that the opponent is h is provided by a *normalized* Poisson, ensuring that Assumption 3 is met.

Definition 8. *A function m^i is CH if it satisfies Assumptions 1 - 3 and Assumption 5, and imposes no further restrictions.*

Thus, CH depends exclusively on the parameter τ. A further assumption that we could impose, in line with others in the literature, is that τ is the same across subjects. This implies that two types k and $k+n$ share the same $g(h)$ but $f_k(h) \neq f_{k+n}(h)$. As GLK would not impose a similar restriction and LK would not accept such restriction, we refrain from imposing this assumption for comparative purpose. Consequently, the reported results in Section 4 will show an upper bound of the predictive success of CH.

It is straightforward to see that the set of predictions of LK and CH is included in the set of predictions of GLK. In other words, we should expect the predictions of LK and CH to be a less accurate description of the data. However, the higher number of assumptions and the type of assumptions place more restrictions in LK and CH than in GLK. Consequently, we should also expect that these theories are also more precise. This trade-off can be calculated using the hit rate and the area.

2.3 Hit Rates, Accuracy, Goodness of fit

The hit rate, $r \in [0,1]$, is the relative frequency of correct predictions. In order to account for individual consistency and address concerns of over-estimation, we redefine the relative frequency of correct predictions as the frequency of a subjects whose dataset is rationalizable by a particular theory of strategic thinking, m^i.

Definition 9. *The dataset $\mathcal{O}_i = \{(s_i^t, G^t)\}_{t=1}^T$ is rationalizable by a particular strategic theory if there exist a k_i and a m^i such that, a) the assumptions of the strategic theory are satisfied and, b) $\Pi_i(s_i^t, k_i) \geq \Pi_i(\tilde{s}_i^t, k_i)$ holds for every alternative strategy $\tilde{s}_i^t$ in every observation t.*

Checking for rationalizability is equivalent to checking whether there is a solution to a finite set of linear inequalities in which what is observed is as good as what the model would predict. We should notice that we do not need to check every strategy in the set of strategies, $s_i' \in S_i$, but only every strategy in the set of best-responses, $s_i' \in BR(m^i(k_i)) \subseteq S_i$. If the inequality above is satisfied for the latter, it is directly implied that it is satisfied for the former. This

follows directly from the definition of best responses (Definition 5). In fact, if there is a solution to the system, the definition of best responses implies that the second part of the definition is true with strict equality. However, when the system is relaxed to allow for errors, the weaker statement allowing also inequality becomes more meaningful as we will see below.

This idea of checking rationalizability via a set of linear inequalities has a long standing in economics and is at the core of the revealed preference literature (Samuelson [1938], Afriat [1967]). In general, we are not placing enough restrictions to *identify* an individual's type. Players with different types and different distributions over types could exhibit the same choices. Our method is focused on testing whether there exists, at least, one type and one distribution over types in line with the theory that solves the system of inequalities.[8] Furthermore, this method is computationally feasible and can be solved in a finite number of steps.

One open question is how many types are necessary to verify that the dataset is rationalizable by a particular theory. The number of types may be limited by the games under study. For some games, a higher type might not provide a better fit than a lower type after a certain threshold. For example, in our application, a better fit cannot be generally obtained with a type higher than 4 because the experimental games are designed to study the empirically most common types, 0–3 (see Crawford et al. [2013]). In other games where there is no such restriction, the only drawback of studying an additional type is imposed by the additional computational time to find whether a solution exists to the different systems of inequalities.[9]

In spite of this qualification, one of the central advantages of using this framework is that we obtain a test of the *necessary and sufficient conditions* of the aforementioned theories. The truth of this statement can be easily verified. If there is a k_i and a m^i that solves the system of inequalities, part a) ensures that this solution is consistent with the theory by definition. If there is a k_i and a m^i that it is not a solution, that implies that $s_i^t \notin BR(m^i(k_i)) \subseteq S_i$ for at least one observation, which violates Assumption 1 and, therefore, does not qualify as a theory of strategic thinking as defined above.

Following the example above, imagine that "6" and "3" were the predictions of a theory of strategic thinking for a particular k_i (i.e., the set of best responses for a particular k_i). According to the method above, nobody is consistent with the theory because every subject violated at least one inequality. Group A violated the inequality associated to the second decision while Group B violated the one associated to the first decision. So, we obtain $r = 0$ rather than $r = 0.5$. In this way, we address the overestimation concern. Now, no violation of any size is allowed. One minor mistake will make an individual to fail the test of the theory even if all other choices would have been rationalizable by the theory. As a result, we have produced a *stringent test*. Consequently, we may want to modify the method further to account for and, more importantly, measure deviations from the theory's predictions.

The revealed preference literature already offers some well-established methods to measure these deviations (Afriat [1972], Afriat [1973], Varian [1990]). If no k_i and a m^i solve the system, we can relax the system and, substituting the expression in b) above for the following expression $\Pi_i\left(s_i^t, k_i\right) \geq e^t \Pi_i\left(\tilde{s}_i^t, k_i\right)$ where $e^t \in [0,1]$ is the *efficiency index*. In particular, e^t is the relative percentage by which the payoff associated to the best-response needs to be reduced

[8] See Section 5 for a discussion of Kneeland [2015] and her method to identify orders of rationality.

[9] Further details of the algorithm applied to our application are presented in section 3.2.

for the inequality to be satisfied. There may be many reasons behind this loss of efficiency. One can think of the loss of efficiency as the result of errors committed in implementing the strategy recommended by the model, individual inability to figure out the correct decision (according to the model) or just genuine misspecification. For our tests, we need to look for the highest efficiency index satisfying the inequality, implicitly assuming that subjects deviate from the model in the less costly manner. Notice that the system is trivially satisfied when $e^t = 0$; and we are back to the stringent test when $e^t = 1$.

For each dataset, we generate a vector of the implied efficiency indexes $\overrightarrow{e} = \left(e^1, ...e^T\right)$. The stringent test – i.e., the necessary and sufficient condition for the observed dataset to be consistent with the model – requires that $\overrightarrow{e} = (1, 1, ..., 1)$. The closer a vector $\overrightarrow{e}$ is to the unit vector, the lower the distance between the model and the observed dataset. There are two main methods to aggregate this vector into a single measure.[10] Afriat [1972] and Afriat [1973] propose a conservative measure for the index, $e_a = min\left\{e^t\right\}$, which ensures that the whole system of inequalities is satisfied by choosing the minimum e^t. As an alternative, Varian [1990] proposes to focus on the square of the Euclidian distance between $\overrightarrow{e}$ and the unit vector, $e_v = \sum_{t=1}^{T}\left(1 - e^t\right)^2 \in [0, T]$. In the application, we focus on Afriat's index, $e_a \in [0, 1]$, because it is more widely used in the empirical literature of revealed preference and, as the hit rate, the domain of this index is between 0 and 1. The results using Varian's less conservative $e_v \notin [0, 1]$ are also presented as a robustness check in Section 4.1.

By relaxing the stringent test in this way, we can see the relationship between the level of efficiency demanded of individuals and the hit rate. If we require $e_a = 0$, the hit rate is going to be one. No efficiency is required and every individual dataset is going to be rationalized trivially. If we increase the demands, fewer individuals are rationalized. In the extreme, we demand individuals to be completely efficient and we are back to the stringent test, where $e_a = 1$.

2.4 Area, Precision, Power

Wright and Leyton-Brown [In Press] offer machine learning techniques that account for the lack of precision of a theory by looking at the number of parameters of the theory. It may be useful to show why we think that the number of parameters is a narrow definition of precision. In any application, LK needs $k - 1$ free parameters because the last one is determined by the sum of proportion of the others and the number of potential distributions of lower types is one per type. However, in CH, the single parameter τ governs the Poisson distribution and, consequently, the relative weights of the different types. Then, it is not necessarily the case that models with a lower number of parameters are more parsimonious than models with more parameters.

Take the popular p-beauty contest in which players have to choose a number between 0 and 100 and the winner is the closest number to 2/3 of the average number. With an uniform random type 0, higher LK types will choose numbers for this game in the series given by: $50 * (\frac{2}{3})^{k_i}$ where $k_i = \{1, 2, 3, ...\}$. For example, type 1 and 2 play exactly 33 and 22, respectively. CH predictions depend on τ and the set of predicted numbers in this game is

[10] There are other suggestions in the revealed preference literature (see Echenique et al. [2011] and Halevy et al. [In press]).

larger than LK's predictions. For example, it can be easily verified that the range of predicted numbers of a CH type 2 includes many numbers between 33 and 22, depending on the particular values of τ. This counter-example suggests that less parameters does not necessarily imply more parsimony but what would a good measure of parsimony be? [11]

The area of a theory, $a \in [0, 1]$, is defined as the relative size of the predicted outcomes with respect to all possible outcomes. So, the importance of this index is not the number of parameters but the consequence of this parameters in increasing or reducing the relative size of outcomes that are consistent with the theory of all possible outcomes. Following the same changes as in the hit rate, a possible outcome should be now redefined as a possible individual dataset. Then, we should expect $\prod_{t=1}^{T} \mid S^t \mid$ possible datasets as the result of an experiment collecting one strategy in each decision t over a set of strategies whose cardinality is $\mid S^t \mid$. Following the example above with two choices, the number of all possible datasets is $100 = 10 \times 10$. Out of these 100 possible datasets, only one dataset is predicted by the theory (i.e. the one in which 6 and 3 are chosen in the first and second question, respectively).

Effectively, calculating the area implies that we need to verify that the theory can rationalize a particular dataset (as described above) and that we need to do this for every possible dataset. Then, the fraction of rationalizable datasets over all possible datasets provides the area of the theory. Given that the number of possible individual datasets increases exponentially with each new game, calculating the area in this manner becomes computationally demanding.[12]

We could, instead, estimate the area as follows. First, we independently drawn datasets with equal probability and replacement from the universe of all possible datasets. Then, for each dataset from the sample, we verify that the theory can rationalize that particular dataset. The estimated area is, then, calculated as the relative size of the predicted datasets with respect to the drawn datasets. The strong law of large numbers implies that the area calculated using this method should be equal to the actual area with probability 1 - as the number of datasets tend to infinity. Any number of datasets that we decide to sample is arbitrary. On one hand, we want the size of the sample to be big enough so the deviation between estimated and true area is minimized and, on the other hand, we want the size to be small enough so we can compute the estimated area within a reasonable amount of time. In our application, we sample 100,000 individual datasets.[13]

In practice, sampling a dataset is formally equivalent to generating an *artificial* individual dataset where the individual makes uniformly random choices in each S^t. Thus, a different interpretation for this process is commonly used in the revealed preference literature - first described by Bronars [1987]. We should notice that $1 - a$ provides the probability of rejecting the theory given the alternative hypothesis of uniform random choices. Thus, we relate the area and the power of our test under the alternative hypothesis of uniform random behaviour. This particular hypothesis coincides with one of the most common specifications of type 0

[11] Thus, when we are shown evidence that variations of the CH model appear systematically among the most parsimonious models for a given dataset when looking at the number of parameters (e.g. Figure 6 in Wright and Leyton-Brown [In Press]), we would like to know that this effect is not driven by the mere fact that the underlying set of predicted outcomes is larger for CH as in the p-beauty contest example. The method of counting the number of parameters cannot address this concern.

[12] In our application, the total number of possible individual datasets is 4.56E+14.

[13] We compared these results to a sample of 10,000 and check that the difference is within 0.001, reinforcing the idea that, after increasing the size of the sample 10 times, the estimated area does not change.

but other alternative hypothesis are reasonable (e.g. most general theories as an alternative hypothesis). The power of the test with these alternative hypotheses, however, would not map into the inverse of the Selten's area as described above.

3 Application

We apply this method to the experimental data from Fragiadakis et al. [2017], which satisfy the three aforementioned conditions: multiple choices, no feedback and controlling for other sources of unobserved heterogeneity.

3.1 Experimental Games

Fragiadakis et al. [2017] modify the 11-20 game (Arad and Rubinstein [2012]) as follows. In a generic experimental game, g, a player i and her opponent simultaneously select integers, s_i and s_{-i}, respectively from a common strategy set $S_g = \{1, 2, \ldots, UB_g\}$, where UB_g is the game's upper bound. Player i earns s_i points automatically for selecting s_i. If s_i is exactly D_g less than s_{-i}, where D_g is g's commonly known undercutting distance, then i earns $B_g > UB_g \times D_g$ *additional* points. If $s_i = s_{-i}$, then player i earns $b_g \in (UB_g - 1, B_g - D_g)$ *additional* points. In the actual experiment, forty subjects face 11 of these games (see Table 1).

TABLE 1.—Parameters of the Experimental Games

Game g	1	2	3	4	5	6	7	8	9	10	11
UB_g	14	17	20	23	26	29	32	18	19	22	23
D_g	3							4			

The points that subjects earn in each of these games are converted into money at the end of the experiment. For that, separate and independent binary lotteries are used (Roth and Malouf [1979]). If a subject earns l points in a game, the corresponding lottery pays \$5 with probability $l/150$ and \$1 with probability $1 - l/150$. This mechanism has been commonly used in experiments to induce risk neutrality as subjects are forced to maximize the number of tickets of the lottery.[14] Consequently, an induced risk neutral agent will maximize the expectation of the following payoff function:

$$\pi_i^g(s_i, s_{-i}) = s_i + B_g \times \alpha_B + b_g \times \alpha_b \tag{2}$$

Thus, in this experiment, the expected payoff for an individual will depend on the probability α_B that she undercuts the other player exactly by D_g, $s_i = s_{-i} - D_g$ and the probability α_b that she chooses the same number as the other other player, $s_i = s_{-i}$. Both cases cannot be true at the same time but it can be the case that none of them is. Further details of the experimental procedures can be read in the original experiment by Fragiadakis et al. [2017].

[14] Although the mechanism has a strong theoretical basis and is moderately popular, it has not been universally celebrated (see Selten et al. [1999]).

3.2 Further Theoretical Details and Method Implementation

We next provide some details as to how the theories of Strategic Thinking (LK, CH and GLK) are applied to these games. These models are distinguished by the particular function m^i affecting the distribution of lower types. This function will impose restrictions on the possible α_B and α_b that a subject can expect given her type.

We need to define, first, the strategy profile of type 0. According to Assumption 2, this strategy is exogenously given. Many applications of non-equilibrium strategic thinking opt for either an uniform random type 0 or a better suited description given the application (see Crawford et al. [2013] for a review). Previous work in 11-20 games (Arad and Rubinstein [2012] and Alaoui and Penta [2016]) have chosen UB_g. In these games, UB_g is the number providing the highest payoff for a naive player who does not form beliefs about other players. The alternative of a type 0 choosing uniformly random in this game implies not only that individuals are unable to forming beliefs about others but also that they ignore the consequences of choosing one action rather than another. This alternative seems too strong given the results of Arad and Rubinstein [2012] showing that a substantial fraction of the population are type 0, who choose UB_g. In these games, the concern of using the correct type 0 specification has been muted since the best-response of a type 1 in these games is $UB_g - D_g$, independently of whether type 0 is UB_g or uniformly random choices.[15] For our analysis, this is, nevertheless, important because the utility derived from responding to one type 0 or another is different. With a uniformly random type 0, choices outside the set of best-responses receive a higher utility. In contrast, by using a UB_g, we are punishing deviations more harshly and, hence, we have a more restrictive test.

Once we define the type 0, one of the advantages of Fragiadakis et al. [2017] is that the set of $BR(m^i(k_i))$ for any strategic theory is straightforward.

Observation 1. *In game g, a GLK player i of type $k_i > 0$ who believes the chance of a type h action is $f_k(h)$. Given assumption 3, $\sum_{h=0}^{k-1} f_k(h) = 1$, her best response (given an arbitrary k_i and $\Delta(K_{-i})$), is an element of the following form $UB_g - \beta \times D_g$ where $\beta \in \{1, 2, ..., k\}$.*

Proof. *Suppose a GLK player i of type k selects some action, s_i, outside the corresponding sequence in game g. Her payoff will be s_i since, given her beliefs over types, neither the b_g bonus is realized, nor the B_g bonus. The aforementioned sequence always contains an element with a profitable deviation paying $s_i + B_g \times f_k(h) + b_g \times f_k(h') > s_i$, where $f_k(h)$ or $f_k(h')$ are always strictly greater than 0, given her beliefs.*

The cardinality of the set of best responses is given by the ceiling of the total number of strategies divided by the undercutting distance, $\lceil \frac{UB_g}{D_g} \rceil$.[16] The actual action for a given type k will depend on the particular distribution of types. Take the experimental game $g = 1$, where $UB_1 = 14$ and $D_1 = 3$. Type 0 is assumed to choose $s_i = UB_1 = 14$. Given this type 0, the sequence for type $k = 1$ only have an element: $s_i = UB_1 - D_1 = 11$. Type 2 chooses an element in the set $s_i = \{11, UB_1 - 2 \times D_1 = 8\}$, depending on whether she believes that the fraction of type 0 players is above or below $\frac{62}{165}$. Type 3 will choose an element in the set

[15] The restriction that $B_g > UB_g \times D_g$ ensures that this is the case.

[16] The ceiling function is defined as $\lceil x \rceil = min\{m \in \mathbb{Z} \mid m \geq x\}$.

$s_i = \{11, 8, UB_1 - 3 \times D_1 = 5\}$, depending on the particular distribution over lower types. Finally, type 4 will choose $n_i = \{11, 8, 5, UB_1 - 4 \times D_1 = 2\}$, depending on the particular distribution over lower types.

Notice that, in this example, type 5 or any higher type is forced to distribute the mass over lower types in the same domain as type 4 because there is no additional room to undercut. If a type 5 were to distribute the mass over a larger domain in another game when that is possible, that type would fail our stringent test (in particular, Assumption 3). This is true not only for the example but more generally for the Experimental Games. In general, the highest type allowed in a given Experimental Game is designed to corresponds with the cardinality of the set of best responses. This feature of the design addresses the question above about how many types to choose for our analysis. The Experimental Games were designed so that the lowest of the highest types is 4, and, consequently, the most common types in previous experiments (type 0 – 3) can be potentially observed in every game.

The payoff function is given by Equation 2 but subjects evaluate this expression depending on their type and distribution over lower levels as in Equation 1. Thus, theories of strategic thinking restrict α_B and α_b to be equated to the frequency of types that can be undercut $\tilde{s}_{-i} = s_i + D_g$ and choose the same strategy $\tilde{s}_{-i} = s_i$, respectively. When a player chooses a strategy outside the set of best responses, $s_i \notin BR(m^i(k_i))$, the corresponding payoff does not include any bonus, $\pi_i^g(s_i, m^i(k_i)) = s_i$ (see proof above). For best responses, $s_i \in BR(m^i(k_i))$, the associated utility depends on the particular theory of strategic thinking.

Suppose that an individual chose $s_i = 11$ in $g = 1$. If we evaluate the associated utility according to LK of type 2, $\pi_i^g(11, LK^i(2)) = s_i + B_g \times \alpha_B + b_g \times \alpha_b = 11 + 100 \times 0 + 35 \times 1$. A LK player of type 2 believes (according to the definition) that $f_2(1) = 1$ and, consequently, everyone chooses 11. Thus, she can only earn the b_g in addition to s_i. Now, if we evaluate the associated utility according to CH of type 2, $\pi_i^g(11, CH^i(2)) = 11 + 100 \times \frac{1}{1+\tau} + 35 \times \frac{\tau}{1+\tau}$. Finally, if we evaluate the associated utility according to GLK of type 2, $\pi_i^g(11, GLK^i(2)) = 11 + 100 \times z + 35 \times (1 - z)$ where z is the probability of finding a type 0 and the complement probability, $1 - z$ is the probability of finding a type 1. These examples make clear how the utilities in the inequalities are calculated.

The implementation of the test for the different theories is as follows. For GLK, we operationalize our test with an arbitrarily fine grid search on $f_k(h) \in [0, 1]$ for each k and $h \in K_i$ and check if the system of inequalities is satisfied at any node in the grid.[17] For LK, setting up the system of inequalities is straightforward because k is enough to derive the expected utility for the corresponding system. For CH, we also proceed with a grid search on τ over a sensible range. We select a range between $[0, 25]$. Camerer et al. [2004] present estimates of τ in the range $[0, 15.9]$. We performed some sensitivity analysis over the selection of the upper bound of the range and results do not change. A reason for this is that, in the Experimental Games, a CH player of type k is identical to a LK player of type 1 player when $\tau = 0$ and, when $\tau \gg k$, then, a CH player of type k tends to a LK player of type k because the mass of beliefs placed in type $k - 1$ increases with τ. Consequently, LK utilities are at the bounds of

[17] Grid search method is standard in the revealed preference literature (e.g., Crawford [2010]). We are not concerned about identification of the parameters but verification of rationalizability. So, it is enough to find a node for which the system of inequalities is satisfied. In case of not finding one, we should rely on the assumption of monotonicity of the measure p – in between nodes – or choose an arbitrarily finer grid.

CH results, governed by τ.

3.3 Notes on Equilibrium

CH and LK predictions tends towards equilibrium in dominance solvable games, such as the Experimental Games, when $\tau \to \infty$ and $k \to \infty$. But the set of equilibria can be larger. For example, the Experimental Games include D_g equilibria in pure strategies: $\{(1,1),(2,2),...,(D_g,D_g)\}$. This result is straightforward. When subjects cannot undercut, their highest payoff is matching the other players' choice. As this is true for every pair of strategies (s_i, s_i) where $1 \leq s_i \leq D_g$, there is a problem of equilibrium selection. Thus, equilibrium players will choose the strategy associated to the equilibrium that they believe the other player chooses.

The strategies in the equilibrium consistent with these theories need to be in the set of best responses of these theories (defined above): $(UB_g - \lceil \frac{UB_g}{D_g} \rceil \times D_g, UB_g - \lceil \frac{UB_g}{D_g} \rceil \times D_g)$ (*Level-k equilibrium*). In addition to this Level-k equilibrium, we can distinguish two additional criteria for selecting pure equilibria. One is the *payoff-dominant equilibrium* (Harsanyi and Selten [1988]), (D_g, D_g), which is Pareto dominant to any other equilibrium, and the other one is the *lower bound equilibrium*, $(1,1)$, which is Pareto dominanted by any other equilibrium.[18]

In short, we create a behavioural model EQ and assume that an EQ player will select strategies according to one of the three selection criteria above and will believe that the other player uses the same criterion. We will use the results of EQ as a baseline to compare the results of the theories of strategic thinking.

4 Results

Figure 1 depicts the histogram of actions in each game. Bars in red are choices consistent with equilibrium play, bars in green are consistent with actions consistent with strategic thinking theories (LK, CH and GLK) and the remaining bars appear in black.

There is apparent support for strategic thinking theories in each game. Not surprisingly, when we look at the average Selten measure across games, we find that the predictive success of strategic thinking theories and equilibrium is 0.58 and -0.052, respectively.[19] However, if strategic thinking theories are intended to explain individual behaviour in games, we should not ignore individual consistency when testing these theories. The method described in Section 2 addresses this.

Table 2 disaggregates the two components of this predictive measure, the hit rate (left panel) and the area (right panel). The corresponding results are further disaggregated by the efficiency levels, e_a, 0.9, 0.95 and 1. An efficiency level of 1 indicates that subjects pass the stringent test. The efficiency levels of 0.9 and 0.95 are the conventionally used thresholds and indicate a tolerance to efficiency losses of, at most, 10% and 5%, respectively.

[18] In a subset of the Experimental Games, there are other equilibria. However, there is no clear selection criterion that will lead to choose these equilibria in that subset of Experimental Games and choose the aforementioned equilibria in the complementary subset.

[19] Notice that, in our application, all strategic thinking theories are observationally equivalent if we look each game separately - as they predict the same actions in each game. Results by game and further details appear in Appendix A.

FIGURE 1.—Histograms by Game

TABLE 2.—Hit Rates and Area by Model and Efficiency Level

	Hit rate			Area (1- Power)		
	0.9	0.95	1	0.9	0.95	1
*G*LK	30%	27.50%	10%	0.18%	0.18%	0.003%
CH	12.50%	12.50%	2.50%	0	0	0
LK	2.50%	2.50%	2.50%	0	0	0
EQ	0	0	0	0	0	0

There are several features to notice in the table. As expected, demanding a higher efficiency reduces the hit rate of the theory as well as the area. The percentages in the cells decrease as we move from an efficiency level of 0.9 to one of 1. As expected, for any given efficiency level, the percentages in CH and LK are smaller than in *G*LK for the hit rates but higher for the areas. This result shows that, as expected, these theories are less accurate theories but more precise than *G*LK.

With the stringent test, *G*LK, CH, LK and EQ present a hit rate of 10%, 2.5%, 2.5% and 0, respectively. To some readers, this percentages may appear low but remember that e_a produces a conservative measure. What is more, $e_a = 1$ is the most stringent of these conservative measures. If we relax the assumption that subjects cannot make mistakes; and we look at the other (lower) thresholds, LK and EQ do not improve their hit rate but CH and *G*LK improve to 17.5% and 35%, respectively. So, allowing a small 10% of efficiency loss improves the hit rate of *G*LK (the most general form of strategic thinking theories) by factor 3.

Result 1. *In the data, the hit rate of the theories of strategic thinking is strictly positive for every efficiency level. Allowing a 10% efficiency loss, a substantial fraction (30%) of the*

individual choices can be explained with these theories. This fraction decreases as we restrict the theory further (e.g., CH and LK).

Given Result 1, we would like to know if the variation across theories of the hit rate is *significant*. To address this question, we build the 95% confidence intervals for each estimate in Table 2. The hit rates indicate the number of subjects out of the total number of subjects who pass the test according to the theory. This frequency follows a binomial distribution where the estimates are the proportion of successful events. We use this distribution to calculate the confidence intervals.[20] Formally, the performance of two theories is not statistically different if the estimate of a theory is within the confidence interval of the other. The results are presented in Figure 2.

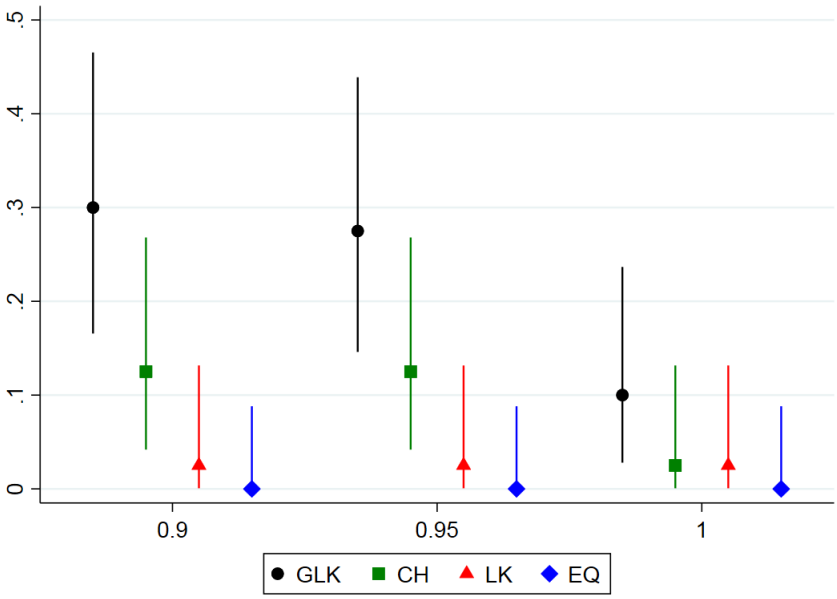

FIGURE 2.—Hit Rates with 95% Confidence Intervals by Model and Efficiency Index

When we look at the stringent tests, $e_a = 1$, there is no significant difference between theories. However, as we relax the stringent test, the differences start to appear quite rapidly. When $e_a = 0.95$, LK and EQ are not different from each other but they are different from CH and GLK, which, in turn, are not significantly different from each other. When $e_a = 0.9$, GLK presents a hit rate significantly higher than CH, which, still presents a hit rate significantly higher than LK and EQ.[21]

Result 2. *There are significant differences in the hit rates across theories. These differences appear related to the restrictions of the theory.*

[20] As EQ presents a proportion at the bound 0, we can only construct the one tail 97.5% confidence interval.

[21] When we use the more powerful z-test of Suissa and Shuster (1984), we find that the only no significance difference is when there is strictly less than 3 persons difference between theories.

Now, if we look at the area in the right panel of Table 2, we realize that the results of these tests are extremely powerful. Theoretically, the area is strictly greater than 0. However, most cells indicate an estimated area of 0 (i.e., 0 subjects out of 100,000 pass the test), indicating that the actual area must be a really small number even if strictly greater than 0. In turn, this indicates that the statistical power of the test with respect to the alternative hypothesis of random behaviour is 1. The only exception is *G*LK with an area of 0.18%, 0.18% and 0.003% at the efficiency levels 0.9, 0.95 and 1, respectively. Consequently, the statistical power of the tests for *G*LK at 0.9, 0.95 and 1 are 99.82 and 99.82 and 99.997, respectively.

Result 3. *In the data, theories of strategic thinking are extremely precise, generally, with areas of 0, leading to extremely powerful tests with respect to the alternative hypothesis of random behaviour.*

Except for the stringent test, the other thresholds are arbitrary. Thus, we investigate the robusteness of these results for other values of $e_a \in [0, 1]$. Figure 3 presents, for a given e_a, the frequency of (actual or artificial) subjects with a higher or equal e_a, according to each model. For each model, the hit rates are presented with solid lines and the areas with dashed lines. For $e_a = \{0.9, 0.95, 1\}$, the values in Figure 3 and in Table 2 coincide.

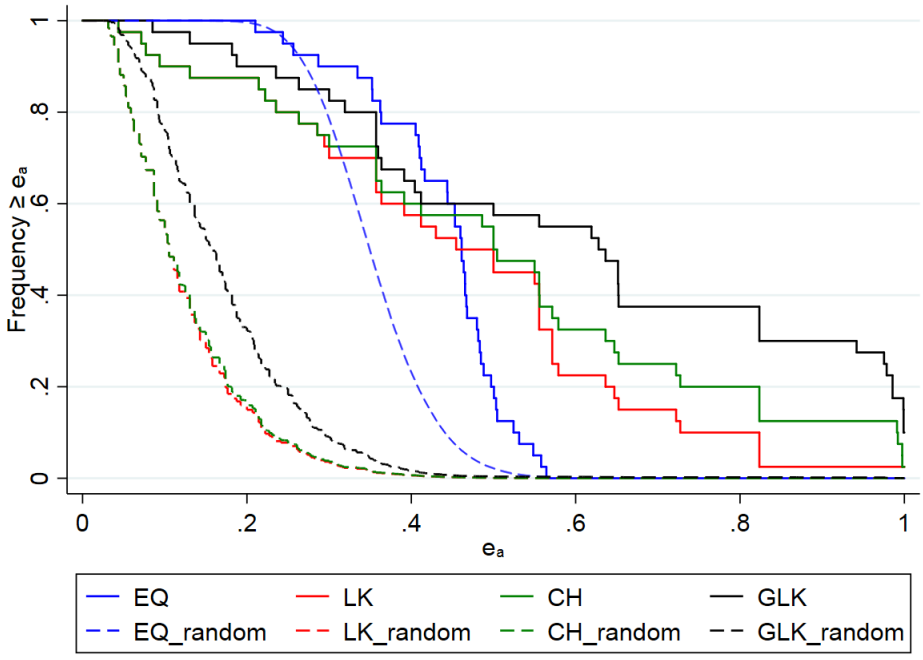

FIGURE 3.—Distributions of Efficiency Indices (Afriat) by Model

The tradeoff between the hit rate and the area is straightforward in Figure 3. When a lower efficiency index e_a is demanded, all the theories explain a higher frequency of subjects. At the same time, the area of the theory also increases significantly and, consequently, the power of test decreases to 0 for these low values of e_a. Furthermore, *G*LK stochastically dominate both

LK and CH for every value of e_a. This result ratifies that GLK have these two models nested within. Also, EQ crosses with the lines of the other models at some point. This effect is a direct consequence of EQ not being nested into any of these theories.

The Selten index, $p = r - a \in [-1, 1]$, aggregates the hit rate and the area based on the aforementioned axioms. The results for each theory and for the threshold levels is presented in Table 3.

TABLE 3.—Predictive Success by Model and Efficiency Index

Selten index	0.9	0.95	1
GLK	0.2982	0.2732	0.097
CH	0.125	0.125	0.025
LK	0.025	0.025	0.025
EQ	0	0	0

Result 3 establishes that our test is extremely powerful (area close to 0) and, therefore, the Selten axiomatic index is going to be dominated by the hit rate. Thus, the numbers in Table 3 are very similar to those in the left panel of Table 2.

Result 4. *The predictive success of the theories of strategic thinking is not mechanically due to the lower precision of these theories (as established in Result 3). The predictive success is the product of the accuracy of these theories (as established in Results 1 and 2).*

4.1 Robustness checks

One concern with the analysis presented so far is that, as Afriat's efficiency index is very conservative (i.e., subjects are measured by their worst deviation in the system of inequalities, $e_a = min\{e^t\}$), our results only provide a lower bound to the actual predictive success of models of strategic thinking.

An alternative efficiency index is proposed by Varian and measures the square of the Euclidian distance from rationalization. However, this index can take values higher than 1 and, therefore, the mapping into the hit rate and the area is not possible.[22] Still, we calculate the index according to each model for every actual subjects as well as for the artificial subjects. Figure 4 presents, for a given e_v, the frequency of (actual or artificial) subjects with a higher or equal e_v, according to each model. The actual subjects are represented with a solid line while the artificial subjects are represented with a dashed line. Now, $e_v = 0$ means a distance of zero with respect to the model. A higher e_v indicates a higher distance between the data and the model.

[22] We could have constructed several normalizations. One normalization divides e_v by T, which is the maximum distance. As we increase the number of observations, predictive theories will present smaller and smaller indices, which makes the comparison harder. Another normalization divides any e_v by the lowest value needed to rationalize every subject. This will only work as long as the lowest value is not 0. Furthermore, we would be comparing theories with different lowest values. As a decision about the normalization is not clear, we report the absolute values.

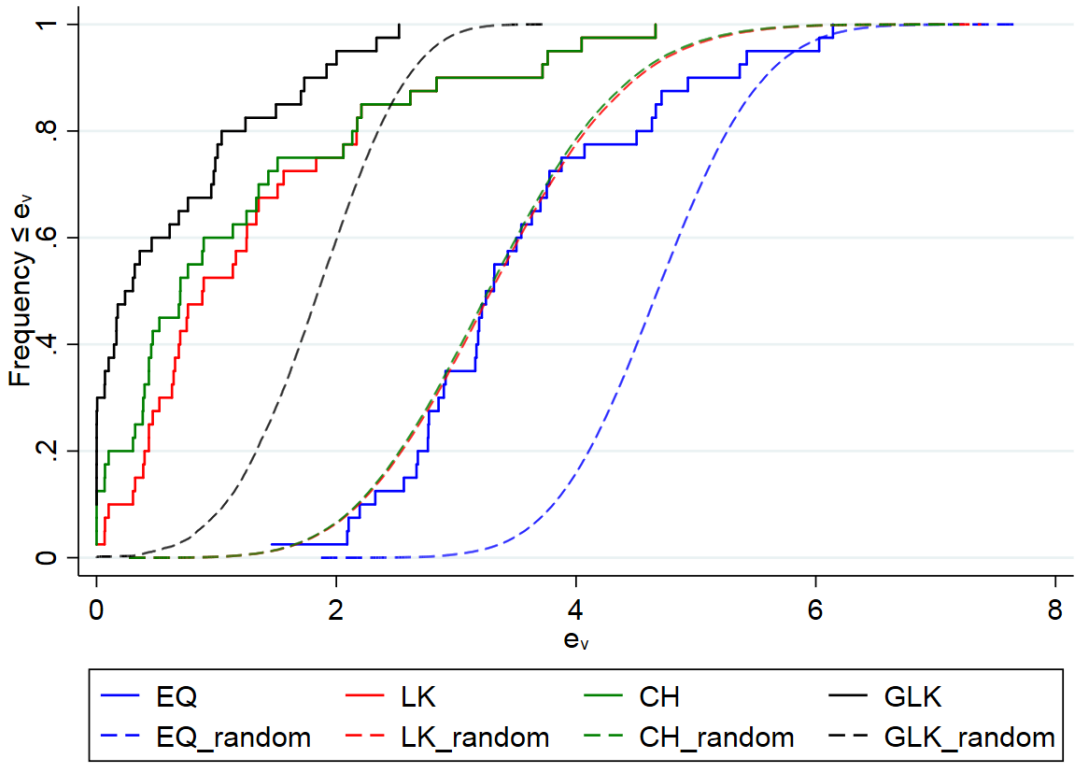

FIGURE 4.—Distributions of Efficiency Indices (Varian) by Model

The results show a similar pattern to that described in Result 4. For each theory of strategic thinking, the solid line (weakly) stochastically dominates the dashed line for every value of e_v. This suggest that the predictive power of these models is not mechanically due to the fact of being more permissive theories. We can also see that the ranking of the different theories remains the same. Yet, the difference between *G*LK and both CH and LK is shorter; and all theories of strategic thinking substantially outperform EQ.

When $e_v = 0$, the frequency of subjects should be the same when $e_a = 1$. However, as soon as we allow some loss of efficiency, the results may vary. For example, when we allow an average efficiency loss of 10%, which corresponds with a squared distance of $e_v = 0.11$, 37.5% of the subjects are classified according to *G*LK (in comparison with 30% when $e_a = 0.9$ and the highest efficiency lost was 10%). However, when we allow an average efficiency loss of 30%, which corresponds with a squared distance of $e_v = 0.99$, 75% of the subjects are classified according to *G*LK (in comparison with 30% when $e_a = 0.7$ and the highest efficiency lost was 30%).

Result 5. *The Result 4 of the conservative test are robust with more permissive tests. In fact, when some efficiency loss is allowed, the frequency of subjects consistent with the theories increase more rapidly in the permissive test than in the conservative test and these frequencies*

stochastically dominate those of random play.

5 Discussion

In this paper, we propose a method to test the necessary and sufficient conditions of models of strategic thinking, non-parametrically. Furthermore, we also control for how easy is for any given theory to pass our test. Thus, we can address questions about the predictive success of current models of strategic thinking as well as the upper bound for the theory of strategic thinking as a whole. Our results show positive evidence for current models and the theory of strategic thinking as a whole. These results are not mechanically due to these theories being permissive.

The inference from these results should be qualified in two dimensions. First, the inferences should be limited to the dominance solvable games in which we applied our method. However, our method could be used to study other games as long as the datasets satisfy three conditions: more than one choice is observed, no feedback and the underlying utility is monotonic in probabilities. Second, we study deterministic versions of strategic thinking, which includes Assumptions 1 - 3. These deterministic versions are the most popular versions and our Assumptions are met in those models. Relaxations of these Assumptions would obviously lead to a better description of the data. For example, one could relax Assumption 1 and allow noisy optimization and find a better description of behaviour (see Goeree and Holt [2004] and Goeree et al. [2017]). But, at the same time, passing the test would be easier for this more permissive version. Alternatively, we could use a different type 0 as Assumption 2 is silent about what type 0 is. As discussed in Section 3.2, the alternative uniform random type 0 would lead to a less severe punishment for deviations, which, in turn, would translate into a higher chance of passing the tests compared with those presented in the paper. Other particular hypothesis about type 0 play could lead to circular explanations (Hargreaves Heap et al. [2014]) so there are limits to relaxing Assumption 2. Assumption 3 is central to the theory of strategic thinking and it is not clear what a relaxation of this Assumption should look like. Instead, one could propose alternative distributions of lower levels and check with our method the predictive success of those alternative distributions. If an alternative model produces higher predictive success, controlling for permissiveness, we will be closer to bridge the gap between theory and evidence (Camerer [2003]). Our results show that, for our games, there is some room for this when errors are allowed.

Our method and results should be put in a wider context. The applicability of our method comes at a cost. Kneeland [2015] uses more complex *Ring* games (and datasets) in order to estimate an individual's type (order of rationality). In our two player normal form games, different types with different distributions of lower types can lead to the same choices. So, our method does not generally *point* identify the actual type but it allows statements about whether subjects are consistent with a specific model to be made in more commonly used and simple games.

More importantly, there has been a recent interest in the reliability of experimental economics, and replicability as a source for credibility in those experimental tests (Maniadis et al. [2014]; Camerer et al. [2016]). Two main reasons why replications are scarce in the profession

are the lack of appropriate incentives (i.e., small chance of impact in a top journal, which is especially important for junior scholars) and the substantial amount of resources required. Before incurring such a investment, one could look at the power of the test to be informed about its reliability. In behavioural game theory experiments, however, the null hypothesis usually involves equilibrium and the behavioural theory acts as the alternative hypothesis. Thus, it is difficult to calculate power because of the lack of precision of the behavioural theory. Using the Selten index, we can always circumvent this problem. We can estimate how easy passing a test is for a given model and, therefore, how confident we should be about an inference. Furthermore, the computational methods that we employ for this estimation are calculated within a reasonable amount of time. Thus, this is an affordable method to figure out the reliability of a test. One recommendation for journals is to only consider papers reporting a small permissiveness.

References

S. N. Afriat. The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77, 1967. 11

S. N. Afriat. Efficiency estimation of production functions. *International Economic Review*, 13(3):568–598, 1972. 3, 11, 12

S. N. Afriat. On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 14(2):460–472, 1973. 4, 11, 12

L. Alaoui and A. Penta. Endogenous depth of reasoning. *The Review of Economic Studies*, 83 (4):1297–1333, 2016. 4, 15

J. Andreoni and J. Miller. Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753, 2002. 6

J. Andreoni, M. Castillo, and R. Petrie. What do bargainers' preferences look like? experiments with a convex ultimatum game. *American Economic Review*, 93(3):672–685, 2003. 6

A. Arad and A. Rubinstein. The 11–20 money request game: A level-k reasoning study. *The American Economic Review*, 102(7):3561–3573, 2012. 4, 10, 14, 15

T. K. M. Beatty and I. A. Crawford. How demanding is the revealed preference approach to demand? *American Economic Review*, 101(6):2782–95, 2011. 3, 6

S. G. Bronars. The power of nonparametric tests of preference maximization. *Econometrica*, 55(3):693–698, 1987. 3, 13

C. F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003. 2, 23

C. F. Camerer, T-H. Ho, and J-K. Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004. 2, 9, 16

C. F. Camerer, A. Dreber, E. Forsell, T-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 2016. 23

M. Costa-Gomes, V. P. Crawford, and B. Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001. 9

M. A. Costa-Gomes and V. P. Crawford. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768, 2006. 2, 3, 9

M. A. Costa-Gomes and G. Weizsäcker. Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762, 2008. 8

I. Crawford. Habits revealed. *The Review of Economic Studies*, 77(4):1382–1402, 2010. 16

V. P. Crawford, M. A. Costa-Gomes, and N. Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51 (1):5–62, 2013. 2, 11, 15

F. Echenique, S. Lee, and M. Shum. The money pump as a measure of revealed preference violations. *Journal of Political Economy*, 119(6):1201–1223, 2011. 12

D. Fragiadakis, D. Knoepfle, and M. Niederle. Who is strategic? Working paper, 2016. 3

D. Fragiadakis, A. Kovaliukaite, and D. Rojo Arjona. Testing cognitive hierarchy assumptions. Working paper, 2017. 4, 14, 15

S. Gächter and A. Riedl. Dividing Justly in Bargaining Problems with Claims. *Social Choice and Welfare*, 27(3):571–594, 2006. 6

S. Georganas, P. J. Healy, and R. A. Weber. On the persistence of strategic sophistication. *Journal of Economic Theory*, 159:369–400, 2015. 10

J. K. Goeree and C. A. Holt. A model of noisy introspection. *Games and Economic Behavior*, 46(2):365 – 382, 2004. 23

J. K. Goeree, P. Louis, and J. Zhang. Noisy introspection in the 11–20 game. *The Economic Journal*, page In press, 2017. doi: 10.1111/ecoj.12479. 23

Y. Halevy, D. Persitz, and L. Zrill. Parametric recoverability of preferences. *Journal Political Economy*, In press. 12

S. Hargreaves Heap, D. Rojo Arjona, and R. Sugden. How portable is level-0 behavior? a test of level-k theory in games with non-neutral frames. *Econometrica*, 82(3):1133–1151, 2014. 2, 9, 23

D. Harless and C. F. Camerer. The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–89, 1994. 6

J. C. Harsanyi and R. Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, 1988. 17

J. D. Hey. An application of selten's measure of predictive success. *Mathematical Social Sciences*, 35(1):1 – 15, 1998. 6

C. Keser and M. Willinger. Theories of behavior in principal–agent relationships with hidden action. *European Economic Review*, 51(6):1514 – 1533, 2007. 6

T. Kneeland. Identifying higher-order rationality. *Econometrica*, 83(5):2065–2079, 2015. 11, 23

Z. Maniadis, F. Tufano, and J. A. List. One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104(1):277–90, 2014. 23

R. Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):1313–1326, 1995. 2, 9

M. Polisson, J. K.-H. Quah, and L. Renou. Revealed preferences over risk and uncertainty. Discussion Paper Series, Department of Economics 201706, Department of Economics, University of St. Andrews, 2017. 4, 6

P. Rey-Biel. Equilibrium play and best response to (stated) beliefs in normal form games. *Games and Economic Behavior*, 65(2):572–585, 2009. 8

A. E. Roth and M. W. Malouf. Game-theoretic models and the role of information in bargaining. *Psychological review*, 86(6):574, 1979. 4, 14

T. C. Salmon. An evaluation of econometric models of adaptive learning. *Econometrica*, 69 (6):1597–1628, 2001. 4

P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938. 11

R. Selten. Properties of a measure of predictive success. *Mathematical Social Sciences*, 21(2): 153–167, 1991. 1, 2, 3, 5, 6

R. Selten and W. Krischker. *Comparison of Two Theories for Characteristic Function Experiments*. Springer, Berlin, 1983. 6

R. Selten, A. Sadrieh, and K. Abbink. Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory and Decision*, 46(3):213–252, 1999. 14

D. O. Stahl and P. W. Wilson. On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995. 2, 9

J. B. Van Huyck, J. P. Cook, and R. C. Battalio. Adaptive behavior and coordination failure. *Journal of Economic Behavior and Organization*, 32(4):483 – 503, 1997. 6

H. R. Varian. Goodness-of-fit in optimizing models. *Journal of Econometrics*, 46(1):125 – 140, 1990. 4, 11, 12

J. T-Y. Wang, M. Spezio, and C. F. Camerer. Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3):984–1007, 2010. 6

M. Willinger and A. Ziegelmeyer. Strength of the social dilemma in a public goods experiment: An exploration of the error hypothesis. *Experimental Economics*, 4(2):131–144, 2001. 6

J. R. Wright and K. Leyton-Brown. Predicting human behavior in unrepeated, simultaneous-move games. *Games and Economic Behavior*, In Press. 3, 12, 13

6 Appendix A

TABLE 4.—Selten Index by Game

Game	Selten Index	
	Equilibrium	LK-actions
1	0	0.625
2	-0.075	0.625
3	-0.05	0.575
4	-0.05	0.575
5	-0.05	0.65
6	-0.075	0.575
7	-0.05	0.55
8	-0.05	0.525
9	-0.05	0.55
10	-0.075	0.6
11	-0.05	0.525
Mean	-0.052	0.58

TABLE 5.—Hit Rates and Area by Model and Efficiency Level for the last 6 games

	Hit rate			Area (1- Power)		
	0.9	0.95	1	0.9	0.95	1
*G*LK	55%	55%	25%	6.65%	6.6%	3.46%
CH	15%	15%	7.50%	0	0	0
LK	5%	5%	5%	0	0	0
EQ	0	0	0	0	0	0

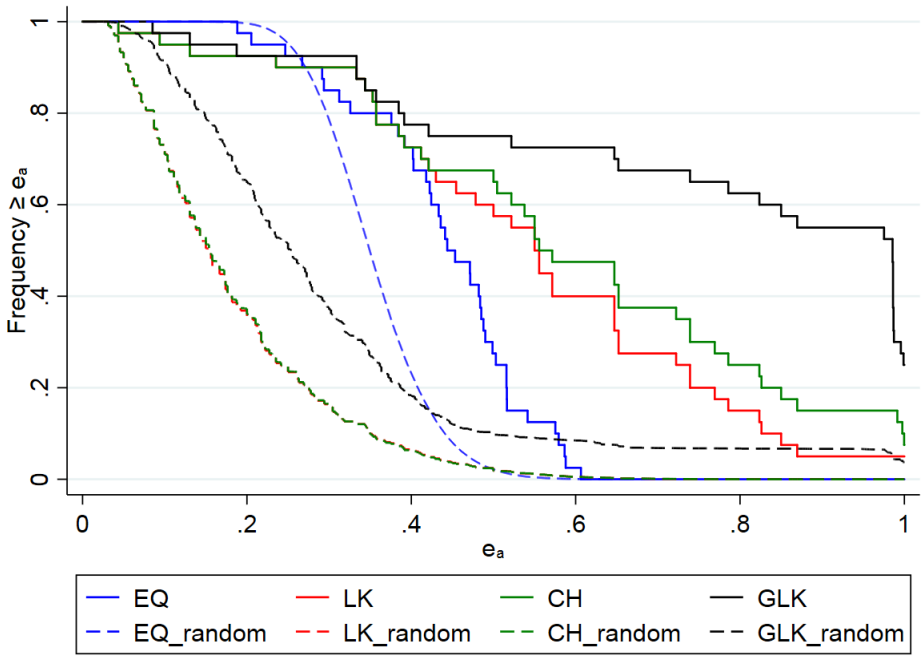

FIGURE 5.—Distributions of Efficiency Indices (Afriat) by Model for the last 6 games

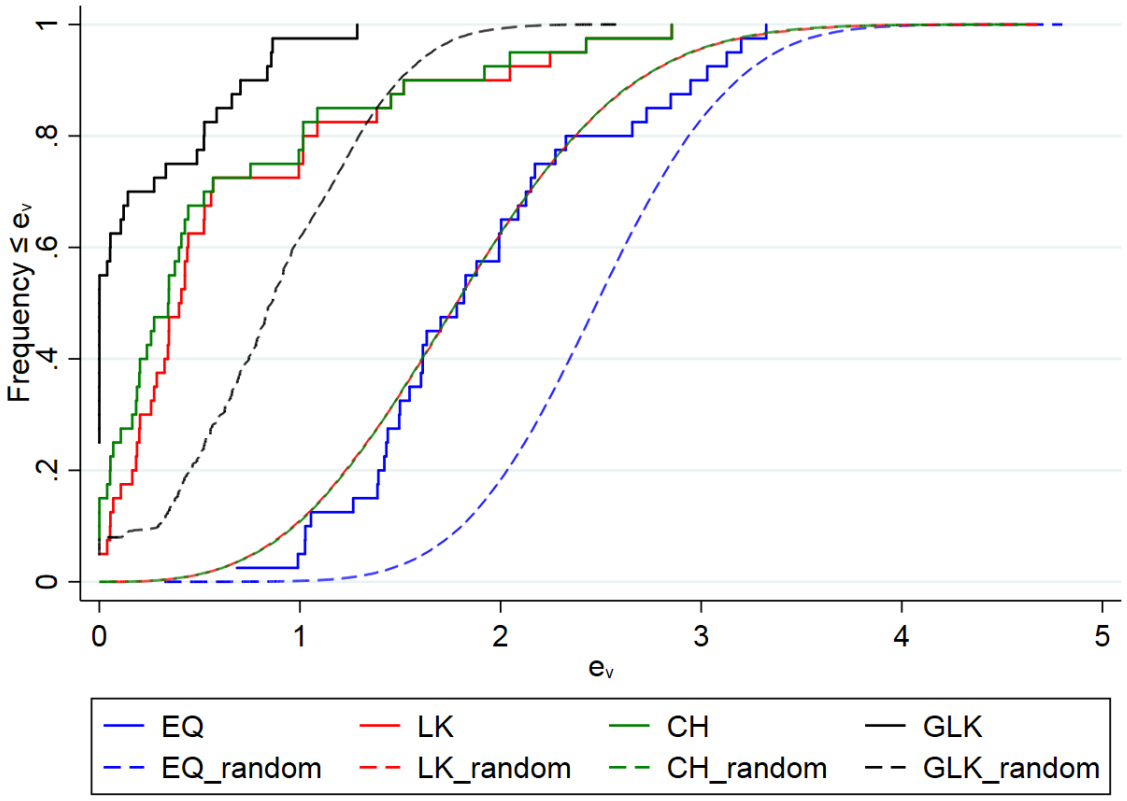

FIGURE 6.—Distributions of Efficiency Indices (Varian) by Model for the last 6 games