MODELING AND TESTING FOR HETEROGENEITY IN OBSERVED STRATEGIC BEHAVIOR

Ernan Haruvy, Dale O. Stahl, and Paul W. Wilson*

Abstract—Experimental data have consistently shown diversity in beliefs as well as in actions among experimental subjects. This paper presents and compares alternative behavioral econometric models for the characterization of player heterogeneity, both between and within subpopulations of players. In particular, two econometric models of diversity within subpopulations of players are investigated, one using a model of computational errors and the other allowing for diversity in prior beliefs around a modal prior for the subpopulation.

I. Introduction

ECONOMIC theories often prescribe a single mode of behavior for "economic" agents. Real and experimental data, on the other hand, often suggest that humans exhibit several distinct modes of behavior.[1] This gap between theory and data calls for a general model of behavioral heterogeneity that allows for multimodal heterogeneity, is empirically testable, and yields robust predictions of behavior. With the goal of representing behavioral heterogeneity in a rich but parsimonious manner, it is natural to begin with a mixture model consisting of a small number of archetypal behaviors that together span the empirically pertinent range of human behaviors. Because it is unlikely that any of these archetypes will exactly capture the behavior of any real human, a general model of heterogeneity must also allow for within-type diversity.

For example, in Stahl and Wilson (1994), a hierarchical theory of strategic thinking with four boundedly rational archetypes (level-0, level-1, level-2, and Nash) was presented with supporting experimental evidence. A fifth "worldly" type was added in Stahl and Wilson (1995) (henceforth, SW). Diversity among players of the same type was explained by errors in the computation of expected payoffs.[2]

Are computational errors the best explanation of within-type diversity? Alternatively, could within-type behavioral diversity be better explained as variations in the priors (beliefs) around a type-specific mean? The level-1 archetype in SW believes that everyone else is equally likely to choose any action (the uniform prior). Instead, there may be a subpopulation of players with priors clustered around the uniform prior, thereby generating within-type diversity in the observed behavior for a level-1 subpopulation.

In this paper, we develop a model of diverse priors as well as an encompassing model with both computational errors and diverse priors. For the purpose of testing these models, we carefully design an experiment and add a computer interface that allows each player to enter hypotheses about the distribution of other players' choices and have the computer calculate the expected payoff of each strategy against a given hypothesis. This on-screen calculator has the potential of substantially reducing computational errors, thereby leaving diversity in priors as the main source of within-type diversity. We estimate an econometric model with diverse priors and one in which within-type diversity is explained solely by computational errors. We compare the results and test each specification against an encompassing model that nests the two models of within-type diversity.

The paper is organized as follows. In section II, we present a stochastic model incorporating the SW level-*n* theory, first with computational errors and then with diverse priors as the source of within-type diversity. Section III contains the details of our experimental design, and section IV presents our statistical methodology and the encompassing model. In section V, we describe the raw data, estimation results, and the tests of robustness of the models across the games. In section VI, we present the comparison tests and results. Conclusions are discussed in section VII.

II. The Formal Theory

In the SW level-*n* theory, different subpopulations of players are allowed different degrees of iteration in a self-referential process of modeling other players' strategies. Of particular interest in the hierarchy are level-1 players, who hold a uniform prior over other players' actions, and level-2 players, who hold the belief that the population consists mostly of level-1 players. The games presented in this analysis are 3×3; hence, the ability to distinguish higher levels given this data is limited. However, additional levels in the hierarchy may be identifiable in more-complex games. In addition, level-*n* theory considers a level-0 (random) behavior, a Nash behavioral mode, and a hybrid ("worldly") behavioral mode.

To estimate the parameters that characterize the model, we must have a model of errors for each subpopulation, so that the likelihood of observed choices within each subpopulation is strictly positive. We describe two econometric models that differ in terms of a player's deviation from the expected action of his subpopulation (that is, within-type diversity).

Received for publication February 12, 1998. Revision accepted for publication April 17, 2000.

* University of Texas.

Partial funding of this research was provided by grant SBR-9410501 and SBR-9631389 from the National Science Foundation (NSF), but nothing herein necessarily reflects views of the NSF. Ray Battalio assisted in the design of the computer interface, and Eric Battalio created the computer code. We are grateful to anonymous referees for very constructive comments. Naturally, we retain full responsibility for any errors and omissions.

[1] For example, see Stahl and Wilson (1994, 1995), Nagel (1995), Stahl (1995), El-Gamal and Grether (1995), and McKelvey and Palfrey (1992).

[2] This approach was also adopted by McKelvy and Palfrey (1995).

The Review of Economics and Statistics, February 2001, 83(1): 146–157

A. Within-Type Diversity Modeled by Computational Errors

Let t denote a player's type, that is, an index of the subpopulation to which a player belongs. Let U_{gjk} denote the payoff in game g to strategy j given the other player chooses strategy k; $j, k \in \{1, \ldots, J\}$. Let $q(g, t)$ be the prior belief of player type t (belonging to subpopulation t) for game g; $q_j(g, t)$ is the probability corresponding to strategy $j \in \{1, \ldots, J\}$. Let P_g^{NE} denote the $J \times 1$ Nash equilibrium (NE) mixed strategy for game g, and let P^0 denote uniform probabilities over the J strategies.

The prior belief of a participant of type t for game g is given by a two-parameter (ϵ_t, μ_t) logit-like specification:

$$q_j(g, t) = \epsilon_t \frac{\exp(\mu_t U_{gj} P_0)}{\sum_{l=1}^{J} \exp(\mu_t U_{gl} P_0)} + (1 - \epsilon_t) P_{gj}^{NE}, \tag{1}$$

$$j = 1, \ldots, J$$

where U_{gj} is the $1 \times J$ row vector of the payoffs to action j in game g. The rationale for the formulation of equation (1) is that it allows for an encompassing prior that can, through appropriate parameter restrictions, capture level-1, level-2, and Nash behaviors, as well as a hybrid type that borrows from all three. In the calculations error model, we allow players to have calculation errors so that a type-t player's perceived expected payoff in game g from action $k \in \{1, \ldots, J\}$ is

$$y_{gk}^*(t) = y_{gk}(t) + e_{gk}, \tag{2}$$

where

$$y_{gk}(t) = U_{gk} q(g, t)$$

and e_{gk} denotes the error term for action k in game g. If the e_{gk} are independent and identically Gumbel (type-I extreme value) distributed, we can use the conditional logit specification (McFadden, 1974) of the probability that a player chooses strategy k given his type t:

$$P(k|g, t) = \frac{\exp(\gamma_t y_{gk}(t))}{\sum_{l=1}^{J} \exp(\gamma_t y_{gl}(t))}, \tag{3}$$

where the scale parameter γ_t can be interpreted as the precision parameter of the player's calculations.[3] Finite values of γ_t will allow positive probability on all strategies, thereby allowing for players' deviations from their prescribed path of action as well as for choices of dominated strategies. We can interpret this as within-type behavioral diversity relative to the perfectly precise case ($\gamma_t = \infty$). We can model the SW level-0 ("random") type by setting $\gamma_t = 0$ in equation (3), which gives uniform probabilities over the J choices forced by a level-0 player, we denote $P(\cdot|g, 0)$ by P^0.

SW defined five archetypes, corresponding to the following regions of the parameter space:

Type 0: $\gamma_0 = 0$ (the random (uniform) player[4])
Type 1: $\gamma_1 > 0.1$ $\epsilon_1 = 1$ $\mu_1 = 0$ (best response to a uniform prior)
Type 2: $\gamma_2 > 0.1$ $\epsilon_2 = 1$ $\mu_2 > 0.1$ (best response to a type-1 opponent)
Type 3: $\gamma_3 > 0.1$ $\epsilon_3 = 0$ (Nash player)
Type 4: $\gamma_4 > 0.1$ $\epsilon_4 \in (0.1, 0.9)$ $\mu_4 > 0.1$ (sophisticated (worldly) player)

In addition, Haruvy, Stahl, and Wilson (1999) (hereafter, HSW) examined evidence for a Maximax archetype and found that it significantly improves the fit. The behavior of this Maximax type is characterized by the probabilistic choice function equation (3) with $y_{gk}(t) = \max_j U_{gkj}$. An alternative prior-based specification that postulates a prior with probability 1 on the column containing the largest payoff was also considered, but statistical tests in HSW rejected the prior-based specification in favor of the Maximax specification.

Let i index the player and $a(i, g) \in \{1, 2, 3\}$ denote the action of player i in game g. Then, assuming a given player uses the same decision process over all games, the probability of player i's choices over G games conditional on belonging to subpopulation t is

$$P_t^i = \prod_{g \in G} P(a(i, g)|g, t). \tag{4}$$

Hence, the likelihood of player i's observed joint choices is

$$L_i = \sum_{t=0}^{5} \alpha_t P_t^i, \tag{5}$$

where α_t is the ex ante probability of being a type t player. The log-likelihood to be maximized over the entire sample of N players is then

$$LLF = \sum_{i=1}^{N} \log L_i. \tag{6}$$

[3] In other applications of the conditional logit model, the scale parameter in the Gumbel distribution is typically normalized to unity to avoid the identification problem. The normalization is not necessary here because γ_t is the only free parameter appearing in equation (3).

[4] Level-0 players are analogous to the zero-intelligence traders of Gode and Sunder (1993). The lower bounds of 0.1 on γ_t ($t = 1, \ldots, 4$) and μ_2 are imposed to ensure identification of the types.

Maximizing equation (6) with respect to the unknown parameters $\{\alpha_t, \gamma_t, \mu_t, \epsilon_t\}$, yields the desired parameter estimates.[5]

B. *Within-Type Diversity Modeled by Diverse Priors*

A second potential source of within-type behavioral diversity arises from players' priors on other players' strategies; that is, individuals in subpopulation t may have some variation among their priors. To allow for diverse priors, we assume that all priors for a particular type are identically distributed with the same mean and standard deviation in a way that will distinguish one type from other types.

In the diverse-priors model, the precision parameter γ is assumed to be large, resulting in the participant playing the best response action to his prior with probability close to 1. Hence, if a player's type t and prior q were known in advance, the probability of choosing strategy k in game g would be

$$B(k|g, t, q) = \lim_{\gamma \to \infty} \frac{\exp(\gamma\, U_{gk} q(g, t))}{\sum_{l=1}^{J} \exp(\gamma\, U_{gl} q(g, t))}, \tag{7}$$

with $q(g, t)$ given by equation (1). The absence of calculation error requires that we find an alternative way to model a player's deviation from the modal behavior of his behavioral type. We therefore allow the participant to deviate from the expected prior of his type and assume that this deviation follows some distribution. Let $f(\,\cdot\,|q(g, t), \sigma^t)$ denote the probability density function for the prior of type t, with mean $q(g, t)$ as prescribed in equation (1), and standard deviation σ^t, where σ^t is a parameter to be estimated. This specification allows us to distinguish within-type diversity due to computational errors from within-type diversity due to diverse priors. The ex-ante probability of a type-t participant choosing strategy k in game g is

$$P(k|g, t) = \int B(k|g, \hat{q}) f(\hat{q}|q(g, t), \sigma_t) d\hat{q}. \tag{8}$$

In particular, we take $f(\,\cdot\,|q(g, t), \sigma^t)$ to be the normal distribution with mean $q(g, t)$ and standard deviation σ^t, truncated to the probability simplex. Computing $P(k|g, t)$ via equation (8) involves a nontrivial computational problem. (Our numerical integration algorithm is described in appendix B.)

Assuming a player uses the same decision process over all games, the probability of player i's joint choices if he is type t is calculated as in equation (4), and the unconditional likelihood of player i's joint choices is calculated as in equation (5). The log-likelihood over the entire sample of N players is computed as in equation (6). Maximizing the log-likelihood with respect to the unknown parameters $\{\alpha_t, \sigma_t, \mu_t, \epsilon_t\}$ yields the desired parameter estimates.

It is conceivable that, in modeling heterogeneity in various settings, behaviors that are ad hoc (that is, irrational in the Savage-Anscombe sense in that they are not based on a prior) could be present. The diverse-prior framework of capturing within-type diversity introduced above cannot account for within-type diversity in these behavioral subpopulations. Within-type diversity in such subpopulations must be modeled using a precision parameter (as in the computational-error approach), even when the rest of the behavioral types are modeled with the diverse-priors specification.

One such ad hoc type is the Maximax type discussed in subsection IIA. Recall that the Maximax type was found by HSW to be superior to the prior-based specification of optimistic behavior in the computational-error framework. This result holds in the diverse-priors framework as well.[6] We therefore proceed with only the Maximax specification for optimistic behavior. We model within-type diversity of this type as arising from computational errors, even within the framework of the diverse-priors model. An additional rationale for incorporating an optimistic type that is not prior based into the diverse-priors model is that it enables us to nest the computational-error model and diverse-priors model within an encompassing model that incorporates both.

C. *Different Behavioral Paths for Different Models of Within-Type Diversity*

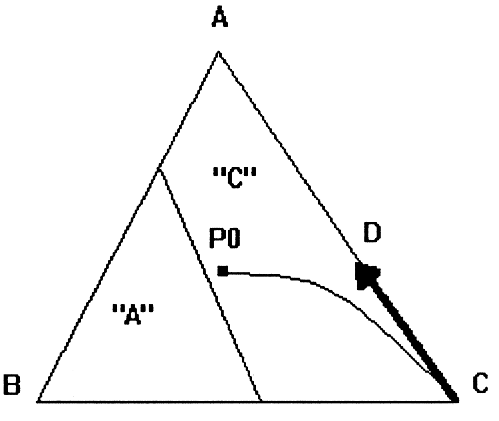
We have developed two models for within-type diversity: computational errors and diverse priors. These models are not merely different functional representations of similar concepts of diversity among types. On the contrary, they specify fundamentally different behaviors and often differ in their classification of a given behavior into a particular type.

It is important to note that, if a subpopulation is characterized by zero within-type diversity, both models of within-type diversity make identical predictions and yield equivalent likelihood functions. Precision (γ_t) in the computational-error approach is driven to infinity and the standard deviation (σ_t) in the diverse priors approach is driven to zero—yielding precisely the same predictions. However, if within-type diversity is present, as the precision in the computational-error model decreases and standard deviation in the diverse prior model increases, different

[5] The six α parameters correspond to the population proportions of the six behavioral types. Five of the types (all but level-0) have precision parameters, denoted by γ. Two of the types (2 and 4) have a free μ parameter. Only level-4 has a free ϵ parameter. Because the α's must sum to 1, the total number of free parameters is thirteen.

[6] Specifically, considering (a) a pure diverse-priors model in which the optimistic type is specified as prior based, and (b) a diverse-priors specification for level-1, level-2, naïve Nash, and worldly types (plus a computational-errors specification for the Maximax type) the latter fits the experimental data better (in terms of likelihood) than does the former. Moreover, using Monte Carlo methods for nonnested hypothesis testing (as in HSW), this difference is statistically significant.

FIGURE 1.—A DEMONSTRATION OF THE DIFFERENT BEHAVIORAL PATHS OF THE COMPUTATIONAL-ERROR MODEL AND THE DIVERSE-PRIORS MODEL



patterns emerge. As a behavioral type's precision falls, it drives the prediction of action probabilities for that type toward P^0. As a behavioral type's standard deviation increases, it drives the prediction of action probabilities for that type towards a behavior characterized by a probability of choosing action j equal to the proportion of the belief space to which j is the best response.

We demonstrate the distinction between the models with an example. Consider a three-strategy game with strategies A, B, and C. Figure 1 presents the three-strategy simplex. Each point in the simplex denotes a unique prior. Area C is the set of priors for which C is the best response. For example, given the pure level-1 prior $P^0 = (1/3, 1/3, 1/3)$, the best response is C. Likewise, area A denotes the set of priors for which A is the best response. Strategy B, in this particular example, is a dominated strategy; we therefore do not have an area B corresponding to priors for which the best response is B.

In the computational-error model, vertex C would denote the behavior of a level-1 player with infinite precision. As the precision falls, the player's behavior approaches the center of the simplex. (That is, the player engages in random behavior.)

In the diverse priors model, vertex C would denote the behavior of a player with a prior drawn from a distribution centered at P^0 with a standard deviation approaching zero. As the standard deviation rises, the radius of the circle of likely priors around P^0 increase, yielding an increasing probability that the player chooses strategy A. On the simplex, this behavior is represented by a movement from point C to point D as the standard deviation rises. Notice that, in contrast to the computational-error model, the dominated strategy B is never chosen, even as the standard deviation rises.

Although the presence of a dominated strategy was convenient for this example, it is not at all necessary for the divergence in prediction paths. For example, consider game 5 in table 1. In that game, no strategy is dominated; choice A is a best response to roughly 85% of all possible hypotheses in the hypothesis space. Hence, if a particular subpopulation contains much within-type diversity, the choice probabilities for that subpopulation predicted by the diverse-priors model, as standard deviation increases, moves toward $P^S \equiv (0.85, 0.04, 0.11)$, eventually converging to P^S as σ_t approaches infinity. In contrast, the computational-error model, as the precision decreases, the choice probabilities approach (1/3, 1/3, 1/3). Thus, our examples show that the two models predict fundamentally different behavioral paths. Results produced by one model cannot be replicated by the other in the presence of within-type diversity.

III. The Experimental Design

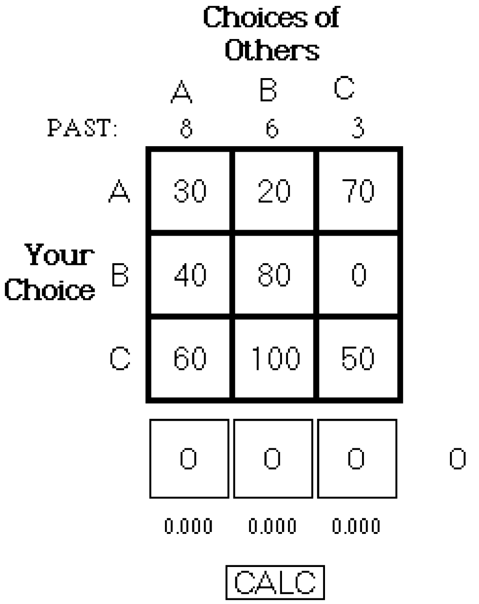
We selected fifteen symmetric 3 × 3 games which were simple to understand, yet rich enough to permit identification of each player type. Some games were dominant-

TABLE 1.—DECISION MATRICES USED IN THE EXPERIMENT

	A	B	C		A	B	C		A	B	C
1)	25*	33	0	2)	15	31	12	3)	15	28	15
A	25	30	100	A	75	40	45	A	10	100	0
B	40	45	65	B	70	15	100	B	5	60	70
C	31	0	40	C	70	60	0	C	80	30	10
4)	29	10	19	5)	45	10	3	6)	28	4	25
A	70	90	38	A	30	50	100	A	10	100	40
B	100	0	40	B	40	45	10	B	0	70	50
C	88	48	43	C	35	60	0	C	20	50	60
7)	31	8	19	8)	27	9	22	9)	3	14	41
A	25	30	100	A	80	60	0	A	75	0	45
B	60	31	51	B	40	10	50	B	80	35	45
C	95	30	0	C	100	5	20	C	100	35	41
10)	32	12	14	11)	15	35	8	12)	35	3	20
A	30	100	50	A	0	100	50	A	40	100	65
B	40	0	90	B	90	63	50	B	33	25	65
C	50	75	29	C	46	82	52	C	80	0	65
13)	13	10	35	14)	33	2	23	15)	17	2	39
A	45	50	21	A	30	100	22	A	40	15	70
B	41	0	40	B	35	0	45	B	22	80	0
C	40	100	0	C	51	50	20	C	30	100	55

* Notes: Underlined numbers at the top of each 3 × 3 matrix are the aggregate choice data.

FIGURE 2.—THE COMPUTER SCREEN



Choices of Others

		A	B	C	
PAST:		8	6	3	
Your Choice	A	30	20	70	
	B	40	80	0	
	C	60	100	50	
		0	0	0	0
		0.000	0.000	0.000	

solvable, and all had unique, symmetric Nash equilibria; five games had unique mixed NE (3, 5, 8, 10, 14). No two archetypes generated the same profile for all fifteen games. The payoff matrices for the row player are presented in table 1; the transposes of these matrices give the payoffs for the column player.

Each participant played each game once on a computer terminal with no feedback until all fifteen games were played; each decision matrix was presented on the computer screen. (See figure 2.) The amount of time allocated for the players to make choices for all fifteen games was fifteen minutes. Within that time, a player could revisit any game and revise the choice for that game. This feature increases the likelihood that a participant's observed behavior comes from a single model of other players.

Looking at the decision matrix as a row player, each participant had to choose a single pure strategy for each game by clicking with a pointing device on any row of the matrix, which highlighted it. In contrast to previous experiments of this type, players were provided an on-screen calculator to reduce computational errors. The participant could enter a hypothesis about the choices of the other players, and the computer would calculate and display hypothetical payoffs for each pure strategy.

To determine payoffs, after all fifteen games were played, for each participant i we computed "token earnings" for each game as $U_{ga(i,g)}P_g^{-i}$, where P_g^{-i} denotes the empirical distribution of all other participants in game g. Token earnings were then translated game by game into the percentage chance of winning \$2.00 for a given game via the roll of three, color-coded, ten-sided dice, which generated a random number uniformly distributed on [0, 99.9] for each game.

Extensive instructions were given before the start of the experiment to ensure that all participants understood the computer interface and how their payoffs would be determined. Following a training period,[7] each participant was given a screening test designed to eliminate potential entrants who were not fully clear on how the game works. More importantly, the screening test was designed to ensure common knowledge among all players that all other players understood the basics of the game.

Three sessions of the experiment were run with three groups of 22, 15, and 21 participants, respectively, for a total of 58 participants. The participants were upper-division business, engineering, social science, and natural science students at the University of Texas. The average payment per participants was \$27.64 for a two-and-a-half hour session.[8]

Prior to running any experiments, we selected a partition of the fifteen games into two subsets—subset I with ten games and subset II with five games—which we would use for testing the out-of-sample predictive power of the models. Each subset is a fair temporal sample and includes the full variety of games of the whole set. Based on these objectives, subset II consists of games 2, 5, 8, 11, and 15.

IV. Statistical Methodology

All parameters were estimated via the maximum-likelihood method; we used a parametric bootstrap to estimate confidence intervals, as described in appendix A, algorithm 2. In each case, the log-likelihood takes the form of equation (6). Maximization was accomplished using the simplex method of Nelder and Mead (1965), together with a coarse grid of initial values to ensure that a global maximum was achieved. The simplex method requires only function evaluations; although it is perhaps not computationally efficient relative to other methods such as the EM algorithm, it is easier to implement for our purposes.

We are faced with two competing nonnested hypotheses on the source of within-type diversity. We nest the two hypotheses within an encompassing model and test each restricted model against the unrestricted encompassing alternative, using likelihood ratios. The encompassing model is constructed by introducing computational error into the diverse-priors model by modifying $B_{ij}(q^t)$ in equation (7). For each type t, instead of taking the limit as $\gamma \to \infty$, we evaluate the expression in equation (7) following the limit operator at γ_t, a parameter to be estimated. Adding four γ_t parameters to the diverse-priors model yields a seventeen-parameter encompassing model.

The test of the diverse-priors model against the encompassing model is of the form H_0: $\gamma_t = 5$ versus H_1: $\gamma_t < 5$, $t = 1, \ldots 4$. For the game payoffs in our experiment, a

[7] For complete instructions, see Stahl (1999).

[8] A session comprised two periods each with the same fifteen games (thus the 2.5-hour duration). For the investigation in this paper, we used data from only the first period.

value for $\gamma_t > 3$ is essentially equivalent to infinity, in the sense that the probabilistic choice function, $P(k|g, t)$, puts virtually all mass on just one action choice. Hence, there is no loss from imposing an upper bound of 5 on γ_t, as we do for computational convenience. Furthermore, the test with the gammas constrained not to exceed 5 is equivalent to testing with no upper bound, so the regularity conditions for the consistency of the likelihood-ratio test are met.

The test of the diverse-priors model against the encompassing model is of the form H_0: $\sigma_t = 0$ versus H_1: $\sigma_t > 0$, $t = 1, \ldots 4$. Under the null hypothesis, the unconstrained model has some of the true parameters on the boundary of the parameter space. Although the classical regularity conditions (as typically stated in advanced econometrics textbooks) are not met, the maximum-likelihood estimators remain consistent.[9] Self and Liang (1987) suggest that the true asymptotic distribution of the likelihood-ratio statistic under the null hypothesis is a mixture of *chi*-squares with 0, 1, . . . , 4 degrees of freedom. Because the right tail of the density of any such mixture lies to the left of a *chi*-square (4) density, the conventional *chi*-square (4) test would be too conservative, increasing the *p*-value of the observed statistic and lowering the probability of rejection; hence, our rejections of the null hypotheses using the conventional *chi*-square tests would hold under the true asymptotic distribution. We also test the joint restrictions in the null hypotheses using a parametric bootstrap procedure for the likelihood-ratio statistic. (See appendix A, algorithm 1.)

V. The Experimental Data and Results for Each Model

A. *Descriptive Characteristics of the Raw Data*

The aggregate choice data are given in table 1. The total numbers of participants making a particular choice are the underlined numbers above each matrix.

One of the questions we address in this paper is whether participants indeed fall into the pure types of level-*n* theory. Previous experiments indicate that they do, and so does maximum-likelihood analysis in this experiment. Moreover, from the raw data, it is clear that eight participants out of the 58 in our experiment behaved in a manner strictly consistent with one of the pure types we discussed above. We have identified four players who had chosen the level-1 strategy in every game, one player who can be identified as a level-2 player with one deviation, and three participants who always chose the row with the highest payoff in it (the pure Maximax types).

Recall that the computer interface allowed individual players to enter personal hypotheses about the choices of the other players. The average number of personal hypotheses entered per participant was 41.55 (with a variance of 31.94), clearly indicating that players, on average, used the interface. Dividing the average by the number of games yields 2.81, the mean number of hypotheses per participant per game.

A study of this type causes some concern that players may change their frame of mind during the experiment, altering the way they make decisions after an initial learning period in which they learn by noting their own mental processes. If this were the case, we would not be able to treat all games as equal in identifying a player's type. For that reason, we allowed players to revisit any game as many times as they wished within the fifteen allotted minutes. That way, if a player learned a technique or developed a better sense of the experiment during the course of play, he was able to go back and apply that new ability to games he had already frequented. Looking at the data, we investigate whether participants took advantage of this feature. The data reveal that each game was revisited on average 1.074 times per player. Out of these revisitations, participants revised approximately one-third of their choices. These findings support our single-frame-of-mind assumption.

Another concern was that the on-screen calculator would not be utilized to the extent we hoped. The percentage of best responses to calculator hypotheses was 82.2% of all choices for which an hypothesis was entered, which indicates that most players made use of the calculator in a way that helped them make the best choices corresponding to their prior. However, 16.9% of decisions were made without using the on-screen calculator at all. Thus, we cannot assume that calculation errors were eliminated by the availability of the on-screen calculator.

B. *The Computational-Error Model*

The thirteen-parameter computational error yields a log-likelihood of −676.992. (See parameter estimates in table 2.[10]) To guage the estimates for the precision coefficients (γ_t), first note from the game payoff matrices in table 1 that a typical payoff difference is 10; second, from equation (3), with $J = 3$ and $y_g = (10, 0, 0)$, the logit probability for the high-payoff action would be $\exp(10\gamma_t)/[2 + \exp(10\gamma_t)]$. For values of $\gamma_t = \{0.05, 0.1, 0.3, 1.0\}$, these probabilities are $\{0.452, 0.576, 0.909, 0.9999\}$ respectively.

To test the robustness of the model to the games used, we conducted two tests using subsets I and II. (See the end of section III.) First, we estimated the model on each subset separately and obtained log-likelihood values of −454.035 and −234.909, respectively. We then compared these numbers to the calculated log-likelihood for each subset using the parameter estimates obtained for the full set: −457.47 and −238.39. Respective *p*-values, based on the asymptotic

[9] See Redner (1981), Self and Liang (1987), and Feng and McCulloch (1996).

[10] This model incorporates the non-prior-based specification of Maximax behavior (subsection IIA). In contrast, using a prior-based specification yields −689.177 and fails the bootstrap comparison in Haruvy, Stahl, and Wilson (1999).

TABLE 2.—PARAMETER ESTIMATES AND CONFIDENCE INTERVALS FOR THE MAIN MODELS ON THE ENTIRE SAMPLE

A) The Computational-Error Model

	Parameter Estimates	Bootstrapped 95%-Confidence Interval	
γ_1	0.479	0.325	1.283
γ_2	0.150	0.100	0.272
μ_2	0.095	0.069	0.163
γ_3	0.428	0.150	5.000
γ_4	0.165	0.122	0.217
μ_4	0.009	0.000	0.022
ϵ_4	0.646	0.495	0.829
γ_5	0.100	0.100	0.136[a]
α_0	0.063	0.000	0.145
α_1	0.139	0.051	0.241
α_2	0.127	0.047	0.224
α_3	0.074	0.002	0.171
α_4	0.415	0.269	0.556
α_5	0.183	0.086	0.277
Log-likelihood	−676.992		

B) The Diverse-Priors Model

	Parameter Estimates	Bootstrapped 95%-Confidence Interval	
σ_1	0.126	0.093	0.160
σ_2	0.619	0.375	0.989
μ_2	0.159	0.102	0.636
σ_3	0.300	0.180	0.549
σ_4	0.102	0.050	0.142
μ_4	0.156	0.080	0.453
ϵ_4	0.474	0.296	0.709
γ_5	0.145	0.103	0.249
α_0	0.162	0.071	0.278
α_1	0.258	0.127	0.393
α_2	0.184	0.037	0.315
α_3	0.156	0.044	0.317
α_4	0.126	0.049	0.229
α_5	0.114	0.035	0.212
Log-likelihood	−662.669		

C) The Encompassing Model

	Parameter Estimates	Bootstrapped 95%-Confidence Interval	
σ_1	0.0631	0.0067	0.1020
σ_2	0.1991	0.0697	0.3106
μ_2	0.1700	0.0816	0.8607
σ_3	0.0934	0.0149	0.2151
σ_4	0.2121	0.1414	0.4666
μ_4	0.0117	0.0000	0.0289
ϵ_4	0.7944	0.5080	0.9000
γ_1	1.4421	0.5045	5.000
γ_2	1.2615	0.2440	5.000
γ_3	0.2106	0.1036	0.4973
γ_4	2.3346	0.5413	5.000
γ_5	0.1408	0.1011	0.2301
α_0	0.0891	0.0007	0.1795
α_1	0.1384	0.0488	0.2491
α_2	0.0910	0.0181	0.1763
α_3	0.1788	0.0773	0.3163
α_4	0.3848	0.2414	0.5232
α_5	0.1180	0.0343	0.2127
Log-likelihood	−655.755		

[a] Because the estimate for γ_5 lies on the lower boundary, this confidence interval may not be reliable.

chi-square distribution for the likelihood-ratio test, are 0.908 and 0.905. Thus, we cannot reject the null hypothesis that the parameter estimates from the full set of games are valid for each subset of games.

Second, we compared the unrestricted maximum for subset II (−234.91) with the calculated log-likelihood value of subset II using the parameter estimates obtained from subset I (−242.59). The asymptotic *chi*-square distribution with thirteen degrees of freedom yields a p-value of 0.285. Thus, we cannot reject the null hypothesis that the parameter estimates from subset I are valid for subset II. In other words, the parameter estimates are stable across these subsets of games, demonstrating the out-of-sample predictive power of the computational-error model.

C. The Diverse-Priors Model

The thirteen-parameter diverse-priors model yields a log-likelihood of −662.669. (See parameter estimates in table 2.) As noted in section IIC, there are substantive behavioral differences between the types in the computational-error model and the diverse-priors model, and this is manifested in the substantial differences in the estimates of the proportions of types (αs) in table 2. For example, the estimated proportion of level-0 types in the diverse-priors model is roughly 10% larger than that of the computational-error model, which is to be expected because the only way that the latter model can explain the choice of a dominated strategy is as a level-0 choice.

To test the robustness of the model with respect to the games used, we again conducted two tests using subsets I and II. The likelihood-ratio statistic for subset I versus the full set of games is 11.08, and 6.77 for subset II versus the full set of games. The asymptotic *chi*-square distribution with thirteen degrees of freedom yields p-values of 0.60 and 0.91, respectively. Thus, we cannot reject the null hypothesis that the parameter estimates from the full set of games are valid for subsets I and II. Furthermore, the likelihood-ratio statistic from predicting subset II from subset I is 8.78, which is asymptotically distributed *chi*-square with thirteen degrees of freedom and yields a p-value of 0.79. Thus, we cannot reject the null hypothesis that the parameter estimates from subset I are valid for subset II. In other words, the parameter estimates are stable across these subsets of games, demonstrating the out-of-sample predictive power of the diverse-priors model.

D. The Encompassing Model

The seventeen-parameter encompassing model yields a log-likelihood value of −655.755. (See table 3 for the parameter estimates.) To investigate the robustness of this encompassing model, we use subsets I and II in the manner used for the restricted models. The likelihood-ratio statistic of subset I relative to the full set of games is 10.14, and the likelihood-ratio statistic of subset II relative to the full set is 4.93. The corresponding p-values from the asymptotic *chi*-square distribution with seventeen degrees of freedom are 0.90 and 0.998, respectively. The likelihood-ratio statistic

TABLE 3.—PARAMETER ESTIMATES UNDER DIFFERENT WITHIN-TYPE DIVERSITY SPECIFICATIONS

A) The Computational-Error Model:

Entire Sample		Nonuser Group of 32		User Group of 17	
γ_1	0.4795	γ_1	0.4876	γ_1	5.0000
γ_2	0.1500	γ_2	0.1437	γ_2	3.8961
μ_2	0.0952	μ_2	0.0844	μ_2	0.3759
γ_3	0.4286	γ_3	0.3933	γ_3	NI
γ_4	0.1652	γ_4	0.2270	γ_4	0.1672
μ_4	0.0099	μ_4	0.0000	μ_4	0.0224
ϵ_4	0.6462	ϵ_4	0.4122	ϵ_4	0.8600
γ_5	0.1000	γ_5	0.1000	γ_5	0.1000
α_0	0.0630	α_0	0.1303	α_0	0.0000
α_1	0.1390	α_1	0.1207	α_1	0.1764
α_2	0.1266	α_2	0.2069	α_2	0.0588
α_3	0.0743	α_3	0.1044	α_3	0.0000
α_4	0.4145	α_4	0.3120	α_4	0.4977
α_5	0.1826	α_5	0.1257	α_5	0.2671
Log-likelihood	−676.992		−395.495		−157.701

B) The Diverse-Prior Model:

Entire Sample		Nonuser Group of 32		User Group of 17	
σ_1	0.1262	σ_1	0.0724	σ_1	0.0000
σ_2	0.6186	σ_2	0.5308	σ_2	0.0216
μ_2	0.1591	μ_2	0.1898	μ_2	0.3367
σ_3	0.3002	σ_3	0.2713	σ_3	NI
σ_4	0.1016	σ_4	0.1071	σ_4	0.2809
μ_4	0.1559	μ_4	0.1542	μ_4	0.0210
ϵ_4	0.4739	ϵ_4	0.4758	ϵ_4	0.9000
γ_5	0.1445	γ_5	0.1793	γ_5	0.1133
α_0	0.1619	α_0	0.3035	α_0	0.0000
α_1	0.2583	α_1	0.1275	α_1	0.1761
α_2	0.1840	α_2	0.1107	α_2	0.0588
α_3	0.1559	α_3	0.1951	α_3	0.0000
α_4	0.1256	α_4	0.1704	α_4	0.5588
α_5	0.1143	α_5	0.0928	α_5	0.2063
Log-likelihood	−662.669		−401.390		−141.582

C) The Encompassing Model:

Entire Sample		Nonuser Group of 32		User Group of 17	
σ_1	0.0631	σ_1	0.0653	σ_1	0.0114
σ_2	0.1991	σ_2	0.1945	σ_2	0.0180
μ_2	0.1700	μ_2	0.1331	μ_2	0.3439
σ_3	0.0934	σ_3	0.1324	σ_3	NI
σ_4	0.2121	σ_4	0.0572	σ_4	0.2725
μ_4	0.0117	μ_4	0.0105	μ_4	0.0208
ϵ_4	0.7944	ϵ_4	0.5681	ϵ_4	0.9000
γ_1	1.4421	γ_1	1.3268	γ_1	5.0000
γ_2	1.2615	γ_2	0.2619	γ_2	5.0000
γ_3	0.2106	γ_3	0.1751	γ_3	NI
γ_4	2.3346	γ_4	4.9978	γ_4	2.9892
γ_5	0.1408	γ_5	0.1765	γ_5	0.1134
α_0	0.0891	α_0	0.1396	α_0	0.0000
α_1	0.1384	α_1	0.1246	α_1	0.1761
α_2	0.0910	α_2	0.1995	α_2	0.0588
α_3	0.1788	α_3	0.3181	α_3	0.0000
α_4	0.3848	α_4	0.1241	α_4	0.5590
α_5	0.1180	α_5	0.0941	α_5	0.2061
Log-likelihood	−655.755		−389.717		−141.566

NI: Not identifiable. When the mixture parameter alpha for any type equals 0, the sigma and/or gamma parameters corresponding to that type are not identifiable.

for subset II relative to subset I is 13.25 with a p-value of 0.72. Hence, the parameter estimates are stable across these subsets of games, demonstrating out-of-sample predictive power of the encompassing approach.

As a measure of goodness of fit, we computed the statistic

$$\sum_{g \in G} \sum_{j \in J} \frac{(n_j^g - n\hat{P}_j(\alpha, g))^2}{n\hat{P}_j(\alpha, g)},$$

where

n is the total number of individuals,

n_j^g is the number of players choosing action j in game g, and

$\hat{P}_j(\alpha, g)$ is the predicted probability of choice j in game g by a randomly selected participant.

Under the null, this statistic is distributed *chi*-square with thirty degrees of freedom (the number of strategies minus 1 times the number of games). The computed values are 29.41 (p-value = 0.496) for the encompassing model, 27.809 (p-value = 0.581) for the computational-errors model, and 35.966 (p-value = 0.209) for the diverse-priors model. Hence, we cannot reject, at any commonly accepted level of significance, that either fitted model generated the data. Graphical illustrations of the fit of the empirical versus predicted choice distribution are provided in figure 3.

VI. Model Comparison Tests

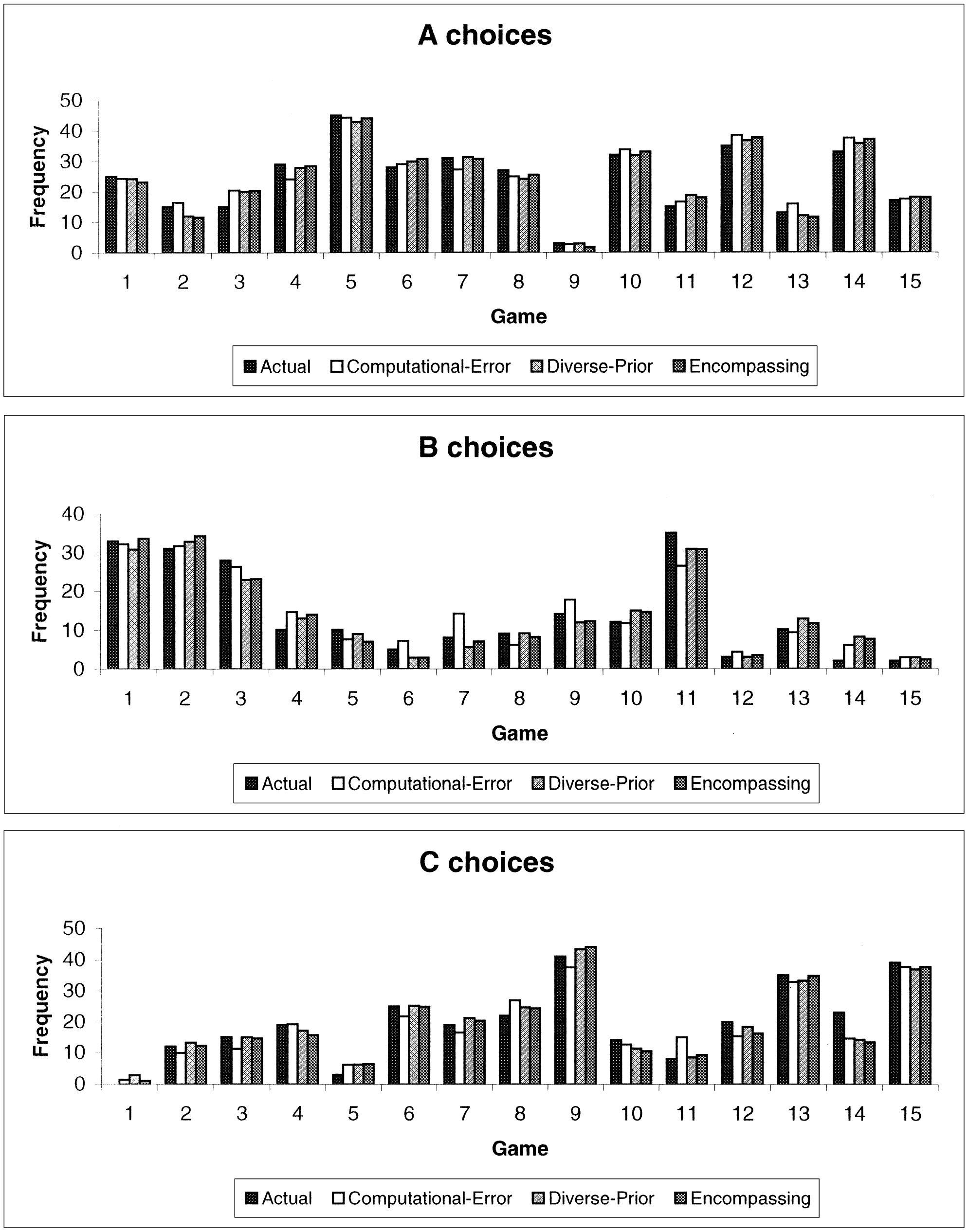
We now turn to the question of which approach to modeling within-type diversity is most appropriate. One possibility is that only one of these approaches is appropriate. In that case, only one approach fails to be rejected in a nested comparison to the encompassing model. A second possibility is that both of these approaches play a significant role and must therefore be captured jointly in a comprehensive—albeit computationally cumbersome—model. A third possibility is that each approach is appropriate in a limited segment of the player population. To entertain these possibilities, we perform nested comparisons of each model to an encompassing model that incorporates both approaches.

A. *The Entire Sample*

There is no reason to assume that observed heterogeneity in human behavior is due solely to diverse priors or solely to computational errors. We nest both approaches within an encompassing model that includes the Maximax type, as did both the computational-error model and the diverse-priors model in the preceding sections. The log-likelihood for the encompassing model is −655.755. In comparison, the computational-error and diverse-priors models yield log-likelihoods of −676.992 and −662.669, respectively.

We want to compare the diverse-priors model against the encompassing model, and the computational-error model against the encompassing model. The likelihood-ratio sta-

FIGURE 3.—PREDICTED VERSUS ACTUAL CHOICE FREQUENCIES



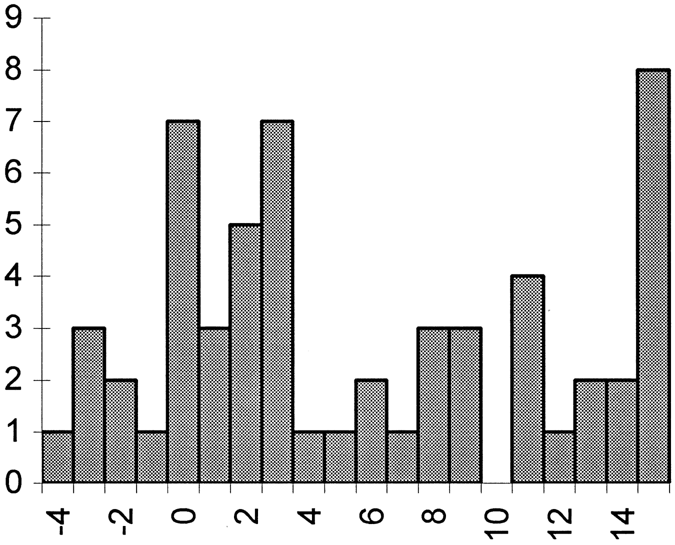
tistic for the first test is 13.09; the asymptotic *chi*-square distribution with four degrees of freedom gives a p-value of 0.011, and the bootstrap yields a p-value of 0.032. The likelihood-ratio statistic for the second test is 42.47; the p-values for both the asymptotic *chi*-square distribution and the bootstrap procedure are less than 0.001, respectively. Hence, we reject both restricted models in favor of the encompassing model.

B. Subpopulations Sorted by Calculator Usage

Estimating on the entire sample of players in the previous section, we rejected each restricted specification of within-type diversity in favor of an encompassing specification that accounts for both. This result may be due to the fact that the participants differed widely in the extent of use of the on-screen calculator, so some could be exposed to significant computational errors while others could have insignificant computational error. Hence, the computational error model and the diverse-priors model may each be best for some participants, but not for all participants.

We could use computed posterior probabilities that each participant was a computational-error type or a diverse-priors type and sort the participants accordingly. However, this would be a pointless data-mining exercise yielding no

Figure 4.—A Histogram of the Difference Between the Number of Best Responses and the Number of Nonbest Responses



predictive insight. In accordance with the scientific method, we suggest an a priori criterion that could be useful in predicting which model might better fit which participant, and then test the predictive performance of this criterion. One such independent criterion predicts that the diverse-priors model would fit better the players who utilized the on-screen calculator because these players made few or no calculation errors. Having recorded the players' hypotheses prior to making a choice, we are able to sort players according to intensity of calculator use. We sort players into two groups: an interface user group (for which the diverse-priors model would be expected to fit better) and a nonuser group.

If the final choice made in a game was the best response to the last hypothesis entered just prior to making that choice, we say that this was a "correct" choice; otherwise, we say that the final choice was "incorrect." It could be argued that a good candidate criterion for sorting players by intensity of calculator use is the number of correct choices. However, this criterion is inadequate because it does not penalize incorrect responses. A natural choice for a sorting criterion is the difference between the number of best responses and the number of incorrect responses to the last hypothesis entered. We will henceforth refer to this criterion as the *difference criterion.* Looking at a histogram on the difference criterion (figure 4), one can observe natural cutoff points between 9 and 10 and between 5 and 6. Using both cutoff points, we obtain three groups: a user group (a difference of 10 or greater) consisting of 17 participants, an undetermined group (a difference of 6 to 9) consisting of 9 participants, and a nonuser group (a difference less than 6) consisting of 32 participants.

For the user group of 17 players, the log-likelihood for the diverse-priors model is -141.58 and -157.70 for the computational-error model. In comparison, the encompassing model yields a log-likelihood of -141.566 when estimated on the user group. Thus, clearly the computational-error model is rejected at any reasonable significance level (both *chi*-square and bootstrap p-values < 0.001), whereas the diverse-priors model cannot be rejected at any reasonable significance level (*chi*-square p-value of 0.999 and a bootstrap p-value $= 0.72$).

For the nonuser group of 32 players, the log-likelihood for the diverse-priors model is -401.39 and -395.50 for the computational-error model. Compared to the encompassing model's log-likelihood of -389.72 with the nonuser group, the likelihood-ratio statistic for the computational error model against the encompassing is 11.556, with an asymptotic *chi*-square p-value of 0.021 and a bootstrap p-value of 0.009. We can therefore reject the computational-error model at the 5% significance level. The likelihood-ratio statistic for the diverse-priors model against the encompassing model is 23.346, with both asymptotic *chi*-square and bootstrap p-values of less than 0.001. We can therefore reject the diverse-priors model at the 5% level or better.

Although a subpopulation that would best fit the computational-error model may exist, neither of the above criteria yield a subgroup for which the computational-error model is statistically superior. The problem may be that these criteria do not reflect well on a player's propensity towards calculation errors. For example, a participant could have entered his true belief as a hypothesis, mentally noted the best response, but then entered irrelevant hypotheses before finally choosing the original best response. This participant would be classified as a nonuser though he clearly made a good use of the calculator and should be better fit by the diverse-priors model. Likewise, a type-2 player who used the calculator to find type-1's best response and then chose the best response to type-1 without using the calculator again would be classified as a nonuser. Thus, our nonuser group could be contaminated with unidentified users, thereby affecting the nonnested test. On the other hand, the user group is a group in which players have a very high number of best responses and therefore the problems just described do not apply in that group, so we can be confident in our bootstrap results.

VII. Conclusions

Theoretical and empirical work are often difficult to merge. Whereas theory prescribes precise behavioral paradigms, real decision-makers, in social settings and in the laboratory, rarely behave in the definite manner prescribed by theory. Nonetheless, theoretical models can be tested with an appropriate model of errors imposed on the data. Heretofore, the accepted model of errors followed the logit formulation, implying calculation errors or imprecision in the calculation of expected payoffs. The most important of this article's contributions lies in the investigation of the hitherto unexplored alternative diverse-priors specification to modeling within-type diversity, as compared to the more traditional computational-error approach. This investigation was done in the context of the multitype "level-*n*" model of boundedly rational strategic thinking. We designed and conducted an experiment to determine which, if any, specification is superior in the selected framework.

To compare these two alternatives of modeling within-type diversity, we nested both of them within an encompassing model and bootstrapped the likelihood-ratio statistic to complement the conventional *chi*-square test. We subsequently concluded that both types of within-type diversity are present in the entire sample of players. However, separating players into subgroups by intensity of calculator use, the diverse-priors model better fits the group of participants who made few or no calculation errors.

Our theory and empirical findings may serve as a complementary theory of initial conditions for dynamic learning theories, potentially serving as a building block in better models of equilibrium selection. Further research (Haruvy, 1999) should incorporate the characterization of initial-period (one-shot) play provided here into leading dynamic theories to investigate how players' models are updated over time when the players are given aggregate information about recent population choices.

REFERENCES

El-Gamal, Mahmoud A., and David M. Grether, "Are People Bayesian? Uncovering Behavioral Strategies," *Journal of the American Statistical Association* 90 (432) (1995), 1137–1145.

Feng, Ziding D., and Charles E. McCulloch, "Using Bootstrap Likelihood Ratios in Finite Mixture Models," *Journal of the Royal Statistical Society, ser. B.* 58 (1996), 609–617.

Gode, Dhananjay K., and Shyam Sunder, "Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality," *Journal of Political Economy* 101 (1) (1993), 119–137.

Haruvy, Ernan, *Initial Conditions and Adaptive Dynamics—An Approach to Equilibrium Selection* unpublished PhD dissertation, University of Texas (1999).

Haruvy, Ernan, Dale O. Stahl, and Paul Wilson, "Evidence for Optimistic and Pessimistic Behavior in Normal-Form Games," *Economics Letters* 63 (3) (1999), 255–259.

McFadden, Daniel L., "Conditional Logit Analysis of Qualitative Choice Behavior" (pp. 105–142), in Paul Zarembka (Ed.), *Frontiers in Econometrica* (New York: Academic Press, 1974).

McKelvey, Richard, and Thomas R. Palfrey, "An Experimental Study of the Centipede Game," *Econometrica* 60 (4) (1992), 803–836.

——— , "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior* 10 (1) (1995), 6–38.

Nagel, Rosemary, "Unraveling in Guessing Games: An Experimental Study," *American Economic Review* 85 (5) (1995), 1313–1326.

Nelder, John A., and Ronald Mead, "A Simplex Method for Function Minimization," *Computer Journal* 7 (1965), 308–313.

Redner, Richard A., "Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions," *Annals of Statistics* 9 (1) (1981), 225–228.

Self, Steven G., and Kung-Yee Liang, "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association* 82 (398) (1987), 605–610.

Stahl, Dale O., "Boundedly Rational Rule Learning in a Guessing Game," *Games and Economic Behavior* 16 (2) (1996), 303–330.

——— , "Evidence Based Rule Learning in Symmetric Normal Form Games," *International Journal of Game Theory* 28 (1) (1999), 111–130.

Stahl, Dale O., and Paul Wilson, "Experimental Evidence of Players' Models of Other Players," *Journal of Economic Behavior and Organization* 25 (3) (1994), 309–327.

——— , "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior* 10 (1) (1995), 213–254.

APPENDIX A

Implementation of the Bootstrap

To implement the bootstrap likelihood-ratio tests of the null hypotheses, we must generate pseudodata for the bootstrap under the null. To accomplish this, we employ the following two algorithms.

Algorithm 1

1. Estimate both the constrained and unconstrained models, yielding parameter estimates $\hat{\theta}^o$ and $\hat{\theta}$, respectively, as well as the corresponding log-likelihood values $\hat{l}^o(\hat{\theta}^o|X)$ and $\hat{l}(\hat{\theta}|X)$.
2. Compute the likelihood-ratio statistic as $\hat{\lambda}(\hat{\theta}^o, \hat{\theta}|X) = -2(\hat{l}^o - \hat{l})$.
3. Draw pseudodata X^* using the constrained parameter estimates.[11]
4. Reestimate both the constrained and unconstrained models using the pseudodata to obtain parameter estimates $\hat{\theta}^{o*}$ and $\hat{\theta}^*$, respectively, as well as the corresponding log-likelihood values $\hat{l}^{o*}(\hat{\theta}^{o*}|X)$ and $\hat{l}^*(\hat{\theta}^*|X)$.
5. Compute the bootstrap value of the likelihood-ratio statistic as $\hat{\lambda}^*(\hat{\theta}^{o*}, \hat{\theta}^*|X^*) = -2(\hat{l}^{o*} - \hat{l}^*)$.
6. Repeat step 3 through 5 M times.

After step 6, we have a set of M bootstrap values $\{\hat{\lambda}^*_m\}_{m=1}^M$. The p-value for our likelihood-ratio test is then computed as $M^{-1} \sum_{m=1}^M I(\hat{\lambda}^*_m > \hat{\lambda})$, where $I(\cdot)$ denotes the indicator function. In all cases, we set $M = 1000$.

The bootstrap employed for estimating confidence intervals for individual parameters in the unconstrained model differs slightly from the above algorithm. In particular, pseudodata sets are generated using the unconstrained parameter estimates.

Algorithm 2

1. Estimate the unconstrained model, yielding parameter estimates $\hat{\theta}$.
2. Draw pseudodata X^* using the unconstrained parameter estimates.
3. Reestimate the unconstrained model using the pseudodata to obtain parameter estimates $\hat{\theta}^*$.
4. Repeat step 2 and 3 M times.

After step 4, we have a set of M (M is set to 1000) bootstrap values $\{\hat{\theta}^*_m\}_{m=1}^M$. To obtain an estimated 95%-confidence interval for a particular element of θ, we sort the M corresponding elements of $\{\hat{\theta}^*_m\}_{m=1}^M$ and take the 2.5th and 97.5th percentiles as bounds of the estimated confidence interval.

[11] To generate the pseudodata, we first generate a pseudorandom uniform deviate v on the interval [0, 1] for each player. We then use the (constrained) estimates of the mixture parameters to partition the [0, 1] interval into J (J being the number of behavioral types) partitions, with partition j of a length equal to type j's estimated proportion in the population. To determine a pseudoplayer type, we assign a player the type j if v falls in the jth interval. Next, we simulate choices in each game, conditional on the pseudoplayer type. For each type and each game, we have estimated probabilities $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$ corresponding to each of the three possible choices. To simulate choices in a particular game by a particular player conditional on his type, we generate a new pseudorandom uniform deviate u on [0, 1], and record the player's choice as 1 if $u \leq \hat{p}_1$, as 2 if $\hat{p}_1 < u \leq \hat{p}_1 + \hat{p}_2$, and as 3 if $\hat{p}_1 + \hat{p}_2 < u$. For each pseudoplayer of a particular type, this process is repeated for each game.

APPENDIX B

Integration Algorithm

The integration required by the diverse-priors model, equation (8), calls for a numerical integration algorithm. Because we needed to use bootstrap techniques to estimate confidence intervals, the integration algorithm had to be computationally efficient. For this purpose, we generated a finite grid on the three-dimensional simplex as follows. The [0, 1] interval was divided into 21 equal-sized subintervals; the collection of endpoints of these subintervals (22 in all) forms the grid for each dimension. Projecting these grids onto the three-dimensional simplex generates (22 × 23/2 =) 253 grid points on that simplex.

The integral in equation (8) is then approximated by

$$\int B(k|g, \hat{q})f(\hat{q}|q(g, t), \sigma^t)d\hat{q} \approx \sum_h B(k|g, q^h)\hat{f}(q^h|q(g, t), \sigma^t)w(q^h), \tag{A1}$$

where

the summation is over the grid points,

$w(\,)$ is a positive weight function on the grid points such that $\Sigma_h w(q^h) = 1$,

$$\hat{f}(q^h|q(g, t), \sigma^t) = f(q^h|q(g, t), \sigma^t/\Sigma_i f(q^i|q(g, t), \sigma^t)w(q^i),$$

is the normalized density function so $\Sigma_h\hat{f}(q^h|q(g, t), \sigma^t)w(q^h) = 1$, and

the weight $w(q^h)$ can be interpreted as the area of the simplex associated with grid point q^h.

There are three distinct kinds of grid points: the three vertices, the strictly interior points, and the interior edge points. When one draws lines parallel to the boundaries of the simplex that bifurcate the imaginary lines connecting the grid points, a regular rhombus surrounds each strictly interior grid point. Relative to these rhombi, the interior edge points are surrounded by an area half as large as the area of one of these rhombi. Depending on the orientation of the rhombi, the area surrounding a vertex is 1/8 or 1/4 of the area of the rhombi, and, taking each orientation as equally valid, the average is 1/6. Accordingly, we assigned weights for the grid points as $W/6$, W, and $W/2$, respectively, where $W \equiv 1/220.5$ ensures that the weights sum to unity.

In the absense of computational constraints, one should choose the size of the grid intervals to be small relative to the standard deviation of f, that is, σ^t. However, because we want to estimate the computational-errors model as a nested version of the encompassing model, we need to allow the σ^t's to go to 0, but that would require infinitely fine grids, which is computationally infeasible. Consider what goes wrong with equation (A1) when a fixed finite grid is used while σ^t goes to 0. If the mean of f, $q(g, t)$, does not correspond exactly to one of the grid points, then the bulk of the mass of f will not be approximated by any grid point, so the numerical summation will be a poor approximation of the true integral. On the other hand, if $q(g, t)$ corresponds exactly to one of the grid points, then the bulk of the mass of f will be concentrated on that grid point, so the numerical summation will be a reasonable approximation of the true integral.

Because the centroid of the simplex, (1/3, 1/3, 1/3), is a grid point[12] and because this point represents the prior of a level-1 archetype, we use equation (A1) with no further modification to compute the choice probabilities for level-1 types. However, for level-2, Nash and Worldly types, the prior $q(g, t)$ generally does not correspond exactly to a grid point. Therefore, equation (A1) may be unacceptably inaccurate for small σ^t.

To circumvent this problem, we added an additional point to the grid exactly corresponding to $q(g, t)$, and we adjusted the weight functions appropriately. Let w_0 be the weight assigned to the new grid point [we chose $w_0 = W/4$]; then the sum of the weights assigned to the original grid points must be reduced by w_0. We confine this adjustment to the three grid points closest to the new point. The new grid point, $q(g, t)$, can be expressed as a unique convex combination of these three nearest grid points: let this convex combination be denoted by $(\beta_1, \beta_2, \beta_3)$. Then, we reduce the weight assigned to these three points by $\beta_i w_0$ respectively; thus, the grid point closest (in this sense) to the new point has its weight reduced the most. This also preserves the property that $\Sigma_h w(q^h) = 1$, where the summation now includes the new point. The normalized density $\hat{f}$ is similarly adjusted.

Before implementing this variable-point algorithm, we often failed to get convergence of the maximization subroutine, and the results were not robust to changes in the grid interval size. After implementing this algorithm, we easily obtained convergence, and the results were robust to the grid interval size, and continuous as σ^t goes to 0; indeed, we found no difference in the encompassing model with arbitrarily small values of σ^t, and the computational-error model with σ^t exactly 0.

[12] By choosing the number of intervals to be a multiple of 3, the centroid of the simplex is one of the grid points.