LEARNING AND EXPERIMENTS: THE BOOTSTRAP TO THE RESCUE

Andreas Blume*
(ablume@pitt.edu)
412-648-7088

Douglas V. DeJong**
(douglas-dejong@uiowa.edu)
319-335-0919

Aaron Lowen**
(aaron-lowen@uiowa.edu)
319-335-3797

George R. Neumann**
(george-neumann@uiowa.edu)
319-335-0850

N.E. Savin**,+
(gene-savin@uiowa.edu)
319-335-0855

*Department of Economics
University of Pittsburgh
Pittsburgh, PA 15260
USA

**Tippie College of Business
University of Iowa
Iowa City, IA 52242
USA
Fax: 319-335-1956

September 5, 2002

+ Corresponding author

ACKNOWLEDGEMENTS: We gratefully acknowledge the comments of Joel Horowitz and of seminar participants at Tilburg University. The authors thank the NSF for its support.

General field: econometrics, game theory and experiments

JEL Classification Numbers: C72, C91, C92

LEARNING AND EXPERIMENTS: THE BOOTSTRAP TO THE RESCUE

ABSTRACT

An important issue in experimental economics is the performance of tests with asymptotic critical values when using sample sizes typically available in practice. Using asymptotic critical values, Blume et al. (2002) tested the parameters of stimulus-response (SR) and belief-based learning (BBL) learning models with experimental data from sender-receiver games. With these same models, we carry out a Monte Carlo investigation of the true levels of the tests with asymptotic critical values. The results show that there are substantial differences between the empirical and nominal levels of the tests. The bootstrap often reduces the distortions in the levels of tests that occur when asymptotic critical values are used. Our Monte Carlo investigation shows that the bootstrap essentially eliminates the differences between the empirical and nominal levels of the tests for sample sizes typically found in practice. Because increasing the number of subjects in laboratory experiments is often an impractical method of increasing the sample size, the bootstrap provides a practical method for controlling the level of tests in experimental economics.

1. INTRODUCTION

Blume et al. (2002) tested the parameters of stimulus-response (SR) and belief-based learning (BBL) learning models using data from experiments with sender-receiver games. The t-tests for testing hypothesis about the parameters were based on asymptotic critical values. The models considered by Blume et al (1998, 2002) completely specify the data generating process (DGP) except for the true values of the parameters. Hence, these models can be simulated for the available sample sizes. Using these models, we carried out a Monte Carlo investigation of the true levels of the tests when asymptotic critical values are used. The Monte Carlo results show that the first-order asymptotic theory often gives a poor approximation to the finite-sample distributions of the test statistics with the samples available in experiments with sender-receiver games. As a consequence, the nominal level of the tests based on asymptotic critical values can be very different from the true levels.

One approach to this problem is to increase the sample size by generating more data in the experiments. However, this is not a solution that can be readily implemented in practice. As Fisher (1935, p. 61) noted in the context of agriculture experiments: "The practical limit to plot subdivision is set, in agricultural experiments, by the necessity of discarding a strip at the edge of each plot…as smaller plots are used, a larger proportion of the experimental area has to be discarded." Analogous considerations apply to experimental design in economics. For example, experimental subjects require compensation; the coordination and control necessary to conduct an experiment becomes

more unwieldy as the number of subjects increases; and the size of the laboratory constrains the number of subjects.

In the case of sender-receiver games, the sample is an N by T panel where N is the number of pairs of players and T is the number of periods of the game. The size of the laboratory and the experimental budget constrain the number of players. The number of periods of play can be increased, but beyond a certain point the additional observations produced by the additional periods are not informative about learning. This point is the number of periods required to reach equilibrium behavior. In other words, after equilibrium is achieved, there is no further learning and hence no further information about learning. Blume et al. (2002) discuss the issues that arise in experiments with large T under the heading of convergence bias.

Hence, the practical problem is how to improve upon first-order asymptotic approximations without increasing the sample size. A potential solution is provided by the bootstrap, which is a method for estimating the distribution of a statistic or a feature of the distribution, such as a moment or quantile. The bootstrap can be implemented for the SR and BBL learning models by generating bootstrap samples from the models using estimated parameters instead of the true parameters. We carried out a Monte Carlo investigation of the ability of the bootstrap to reduce the distortions in the level of a symmetrical, two-tailed t-test that occur when asymptotic critical values are used. The results of the investigation show that the bootstrap essentially eliminates the differences between the empirical and nominal levels of the tests that occur with asymptotic critical values and that it does so with the sample sizes available in laboratory experiments.

If the objective is to obtain confidence intervals for the parameters, then the bootstrap can be used to reduce the error in the coverage probability. The Monte Carlo results show that bootstrap critical values can reduce the distortions in the coverage probabilities that occur with asymptotic critical values. However, in the case of the confidence intervals, the bootstrap does not completely remove the distortions with sample sizes used in the experiments.

The organization of the paper is the following. Section 2 describes the sender-receiver game, Section 3 describes the SR model, and Section 4 describes the BBL model. The bootstrap is presented in Section 5, and the design of the Monte Carlo experiments in Section 6. The results of the Monte Carlo investigation are reported in Section 7. Concluding comments are contained in Section 8.

2. THE GAME

In experiments using sender-receivers games, players are randomly matched from two populations: N senders and N receivers. Each period all players are matched, each sender with one receiver, and all pairings are equally likely. The game played by each pair in each period is between an informed sender and an uninformed receiver. The sender is privately informed about his type, θ_1 or θ_2, and types are equally likely. The sender sends a message, "1" or "2", to the receiver, who responds with an action, a_1 or a_2. Payoffs depend on the sender's private information, his type, and the receiver's action, but not on the sender's message. We focus on a common interest game in which the incentives of the players are fully aligned. If the sender's type is θ_1 (θ_2) and the receiver takes action a_2 (a_1), the payoffs to the sender and receiver are 700,700, respectively; otherwise, the payoffs are zero for both the sender and receiver.

A strategy for a sender maps types into messages; for a receiver, a strategy maps messages to actions. A strategy pair is a Nash equilibrium if the strategies are mutual best replies. More formally let Θ be a finite set of types and $\pi(\theta)$ the prior distribution of types. The sender's set of pure strategies is the set of mappings, s: $\Theta \rightarrow M$, from the type set to the finite set of messages, M. The receiver's set of pure strategies is the set of mappings, r: $M \rightarrow A$, from M to the finite set of actions A. Given type $\theta \in \Theta$, message $m \in M$, and action $a \in A$, the sender's payoff is $v_S(\theta, a)$ and the receiver's payoff is $v_R(\theta, a)$. Messages do not directly affect payoffs. For any finite set X, let $\Psi(X)$ denote the set of probability distributions over X. The payoff from a mixed action $\alpha \in \Psi(A)$ is $v_i(\theta, \alpha) = \sum_{a \in A} v_i(\theta, a)\alpha(a),\ i = S, R.$

We denote mixed behavior strategies for the sender and the receiver by σ and ρ, respectively. Let $\sigma(m, \theta)$ denote the probability of type θ sending message m and let $\rho(a,m)$ stand for the probability that the receiver will choose action a in response to message m. The pair (σ, ρ) is a Nash equilibrium if σ and ρ are mutual best replies:

$$if\ \sigma(m,\theta) > 0, then\ m\ solves \max_{m' \in M} \sum_{a \in A} v_s(\theta, a)\rho(a, m')$$

and

$$if\ \rho(a,m) > 0, then\ a\ solves \max_{a' \in A} \sum_{\theta \in \Theta} v_R(\theta, a')\sigma(m, \theta)\pi(\theta).$$

The equilibrium is called separating if each sender type is identified through his message. In a pooling equilibrium, the equilibrium action does not depend on the sender's type; such equilibria exist for all sender-receiver games. For example, a separating equilibrium is one where the sender sends "2" if she is θ_1 and "1" otherwise and the receiver takes action a_2 after message "2" and a_1 otherwise. An example of a

pooling equilibrium is one in which the sender, regardless of type, sends "2" and the receiver always takes action a_2.

In each period, players play a two-stage game. Prior to the first stage each sender is informed about her type. In the first stage, a sender sends a message to her paired receiver. In the second stage, a receiver takes an action after receiving a message from his paired sender. Each sender and receiver pair then learns the sender type, message sent, action taken and payoff received. All players next receive information about all sender types and all messages sent by the respective sender types. This information is provided for the current and all previous periods of the game played by the particular cohort of players.

In this setting we observe players' actions, not their strategies. Also, players receive information about actions, not strategies. They do not observe which message (action) would have been sent (taken) by a sender (receiver) had the sender's type (message received) been different. This is important for how the learning rules are formulated.

3. STIMULUS-RESPONSE MODEL

SR and belief-based models both use propensities to determine choice probabilities. We use the index i to refer to one of the N senders. Each sender is randomly assigned a type, θ, each period. We use the index j to refer to receivers. This section describes the SR model.

By SR we mean that the individual play of each sender is affected only by rewards obtained from own past play. Following Alvin E. Roth and Ido Erev (1995),

define the propensity, $Q_{i\theta m}(t)$, of sender *i* to send message *m* at time t when her type is θ as:

$$Q_{i\theta m}(t) = \varphi_0 Q_{i\theta m}(t-1) + \varphi_1 X_{i\theta m}(t-1) \qquad (1)$$

where $X_{i\theta m}(t-1)$ is the reward sender *i* receives from sending message *m* at time *t*-1(with $X_{i\theta m}(t-1) = 0$ if sender *i* sent the other message in period *t*-1 or her type was not θ). There is a propensity for each possible message. We refer to the parameter of $Q_{i\theta m}(t-1)$ as the memory parameter and the parameter of $X_{i\theta m}(t-1)$ as the learning parameter. Given this specification of propensities, the probability that sender i sends message m is a logit-like function

$$P_{i\theta m}(t) = \Pr(\text{Sender i of type } \theta \text{ sends m at time t}) = \frac{\exp(Q_{i\theta m}(t))}{\sum_{m'} \exp(Q_{i\theta m'}(t))}. \qquad (2)$$

To complete the specification of the SR model we require an initial condition for the propensities, the values of $Q_{i\theta m}(1)$. Values chosen for $Q_{i\theta m}(1)$ affect $P_{i\theta m}(1)$ and the speed with which rewards change probabilities of making a particular choice. In the spirit of Roth and Erev (1995) we set $Q_{i\theta 1}(1) = Q_{i\theta 2}(1) = 350$, which is on the scale of rewards received by participants in the experiments analyzed by Blume, et al. (2002).

The senders, who can be of two types, can send message "1" or "2". Let $y = I\{\text{message} = \text{"2"}\}$, where $I\{A\}$ is the indicator function that takes the value 1 if event *A* occurs and 0 otherwise. Let $P_{im}(t) = (2-\theta)P_{i1m}(t) + (\theta-1)P_{i2m}(t)$ be the probability that sender *i* sends message *m* in period t. The log likelihood function for the sender data is

$$\ln L(\varphi_0, \varphi_1) = \sum_{i=1}^{N} \sum_{t=1}^{T} [y_i(t) \ln(P_{i2}(t)) + (1-y_i(t)) \ln(1-P_{i2}(t))] \qquad (3)$$

where $P_{i2}(t)$ is the probability of sending message "2".

To show how the $P_{i\theta 2}$ and hence the likelihood function (3) depends on the parameters, it is convenient to rewrite the propensity (2) as a partial sum:

$$Q_{i\theta m}(t) = \varphi_0^t Q_{i\theta m}(1) + \varphi_1 \sum_{j=1}^{t-1} \varphi_0^{t-1-j} X_{i\theta m}(j). \quad (4)$$

Using (4), the probability $P_{i\theta 2}(t) = 1/[1+\exp(\Delta Q_{i\theta}(t))]$ where

$$\Delta Q_{i\theta}(t) = \varphi_0^t (Q_{i\theta 1}(1) - Q_{i\theta 2}(1)) + \varphi_1 \sum_{k=1}^{t-1} \varphi^{t-1-k} (X_{i\theta 1}(k) - X_{i\theta 2}(k)). \quad (5)$$

From (5) we see that the likelihood function depends on the initial difference in the propensities, $Q_{i\theta 1}(1) - Q_{i\theta 2}(1)$, and on the memory and learning parameters.

In principle, the differences $Q_{i\theta 1}(1) - Q_{i\theta 2}(1)$ can be estimated along with φ_0 and φ_1. The drawback of estimating the differences is that the estimates are typically imprecise. In the present context, however, the differences do not enter into the expression for the P_{i2}. This is because in the Blume et al. (1998) design, the experiments are designed (by privatizing the messages) to make $Q_{i\theta 1}(1) - Q_{i\theta 2}(1) = 0$, $\theta = 1,2$, and thus eliminate the effect of the initial conditions for the propensities. Consequently, (5) reduces to

$$\Delta Q_{i\theta}(t) = \varphi_1 \sum_{j=1}^{t-1} \varphi_0^{t-1-j} (X_{i\theta 1}(j) - X_{i\theta 2}(j)), \quad (6)$$

which implies that the probabilities P_{i2} depend only on the parameters φ_0 and φ_1 and similarly for the likelihood function (3).

In an analogous manner, we can derive the characteristics of the DGP for receivers. Recall receivers only observe the message sent. As with senders, the individual play of each receiver is affected only by rewards obtained from own past play

and consequently can be considered separately from sender play. We define the propensity, $Q_{jma}(t)$, of receiver *j* to take action *a* at time *t* having received message m as:

$$Q_{jma}(t) = \gamma_0 Q_{jma}(t\text{-}1) + \gamma_1 X_{jma}(t\text{-}1) \qquad (7)$$

where $X_{jma}(t\text{-}1)$ is the reward receiver *j* gets from action *a* at time *t*-1 having received message m (with $X_{jma}(t\text{-}1) = 0$ if receiver *j* took the other action in period *t*-1, or if the message was not m). There is a propensity for each action given the message sent. The probability of receiver *j* taking action *a* is a logit-like function defined analogously to equation 2; the log likelihood function for the receiver data is defined analogously to equation 3, with $y = I\{\text{action} = \text{“2”}\}$.

The likelihood function factors for senders and receivers, and, hence, the maximization of the likelihood can be carried out separately for senders and receivers. For this reason, we have focused only on the senders. This completes the description of the SR model. What remains to be discussed are identification issues.

Identification of the parameters φ_0 and φ_1 depends on the speed of learning. Consider the behavior of the difference in rewards, $X_{i\theta 1}(t) - X_{i\theta 2}(t)$, $\theta = 1, 2$. If play converges to equilibrium in the first round, then the difference in the rewards does not change over time. The consequence is that φ_0 and φ_1 are not identified. Of course, if the reward parameter is zero ($\varphi_1 = 0$), then φ_0 is not identified.

More generally, the speed of learning determines the amount of the available sample information that is relevant for estimating the parameters. Suppose $X_{i\theta 1}(t) - X_{i\theta 2}(t) = c$ for $t > T^*$ then increasing T beyond T*, $T > T^*$, will not increase the precision of the estimator. Rapid learning means that T* is small, and hence there is, relatively speaking, little information available to estimate the parameters. On the other hand,

increasing T beyond T* will appear to improve the fit of the model to the data when in fact there is no learning after T^*. We refer to this effect as convergence bias. Convergence bias is discussed in more detail in Blume et al. (2002).

4. **BELIEF-BASED LEARNING MODEL**

While the SR and BBL models both use propensities to determine choice probabilities, they differ in that senders and receivers in the SR model are affected only by rewards obtained from own past play. In the BBL model, we have homogeneous senders and homogeneous receivers, and senders and receivers update their beliefs using the same set of information, the previous period's history of sender types and messages sent.

In the BBL model, we define the propensity, $Q_{i\theta m}(t)$, of the sender to send message *m* at time *t* when type is θ as:

$$Q_{i\theta m}(t) = \beta_0 Q_{i\theta m}(t-1) + \beta_1 X^e_{i\theta m}(t-1) \qquad (8)$$

where $X^e_{i\theta m}(t-1)$ is the expected reward of the sender from sending message *m* based on data available at time t-1. The expected reward is calculated using the past frequencies of play. For senders, $\eta^{t-1}(\theta|m)$ is the frequency of type θ given message *m* in period t-1, and for receivers, $\rho^{t-1}(a|m)$ is the frequency of action *a* given message *m* in period t-1. The sender choice probabilities again are logit as in (2) with (8) replacing (1) as the definition of $Q_{i\theta m}$ in the likelihood function (3).

As in the Blume et al. (1998) laboratory experiments, senders cannot predict future behavior of receivers from population information about past behavior of receivers. Hence, senders calculate the *expected* frequency of action *a* given message *m*, $\hat{\rho}^{t-1}(a \mid m)$, from data on past sender play. The sender formula for the expected reward is

$$X^{e}_{i\theta m}(\text{t-1}) = \sum_{a} v_s(a,\theta)\hat{\rho}^{t-1}(a \mid m) \tag{9}$$

where

$$\hat{\rho}^{t-1}(a \mid m) = \left\{ \begin{array}{c} 1 \;\; if \;\; a = \arg\max_{a'} \sum_{\tau \le t-1} \sum_{\theta} v_s(a',\theta)\eta^{\tau}(\theta \mid m) \\ 0 \;\; if \;\; a \notin \arg\max_{a'} \sum_{\tau \le t-1} \sum_{\theta} v_s(a',\theta)\eta^{\tau}(\theta \mid m) \\ 0.5 \;\; otherwise \end{array} \right\} .$$

Senders have the information necessary to calculate the point estimate. Receivers, on the other hand, are assumed to predict future sender behavior directly. Specifically, let $Q_{\text{jma}}(\text{t})$ be the propensity of the receiver to take action *a* at time *t* having received message m:

$$Q_{\text{jma}}(\text{t}) = \delta_0 Q_{\text{jma}}(\text{t-1}) + \delta_1 X^{e}_{\text{jma}}(\text{t-1}) \tag{10}$$

where the receiver formula for the expected reward is

$$X^{e}_{\text{jma}}(\text{t-1}) = \sum_{\theta} v_R(a,\theta)\eta^{t-1}(\theta \mid m) . \tag{11}$$

To illustrate an individual propensity, consider the propensity to choose action a_1 given the message sent $m = 1$,

$$Q_{\text{j1a}_1}(\text{t}) = \delta_0 Q_{\text{j1a}_1}(\text{t-1}) + \delta_1 \left\{ p(\theta_1 \mid m=1)(0) + \left[1 - p(\theta_1 \mid m=1)(700) \right] \right\}$$

where $p(\theta_1|\text{m=1})$ is the probability that the sender is type θ_1 given message "1" was sent. This probability is calculated using period *t*-1's distribution of sender types and messages sent. The receiver's payoff from choosing action a_1 when the sender is type θ_1 (θ_2) is 0 (700). There is a propensity for each possible action given the message sent. Thus, receiver choice probabilities are again logit, as in (2), and the log likelihood function for the receiver is analogously defined, as in (3). This completes the description of the BBL model. As with the SR model, we focus on the senders in BBL.

5. THE BOOTSTRAP

In the Monte Carlo experiments with the SR and the BBL models, the null hypotheses tested are composite, meaning that they do not completely specify the data generation process (DGP). This section describes how the bootstrap can be used to test a composite hypothesis when the test statistic is asymptotically pivotal.

The problem involved in testing a composite hypothesis is illustrated using the memory parameter in the SR model. Consider testing the null hypothesis H_0: $\varphi_0 = 0.8$. In this example, the data generation process (DGP) is unknown when H_0 is true because the value of the updating parameter φ_1 is unknown. As a result, the exact, finite sample distribution of the test statistic for testing H_0 is unknown when the null hypothesis is true. The approach adopted here is to base the test of H_0 on an estimator of the *Type I critical value*. This is the critical value that would be obtained if the exact finite-sample distribution of the test statistic were known when the null is true; see Horowitz and Savin (2000).

Let T_{n0} be the test statistic for testing H_0. In the experiments, the null H_0 is tested using a symmetrical, two-tailed test. H_0 is rejected by such a test if $|T_{n0}|$ exceeds a suitable critical value and is accepted otherwise. The exact, α-level Type I critical value is $z_{n\alpha}$ where $z_{n\alpha}$ is defined by $P(|T_{n0}| > z_{n\alpha}) = \alpha$. A test based on this critical value rejects H_0 if $|T_{n0}| > z_{n\alpha}$. Such a test makes a Type I error with probability α.

However, $z_{n\alpha}$ cannot be calculated in this application except in a special case. The exception is when T_{n0} is *pivotal*. A test statistic is pivotal if its finite-sample distribution does not depend on any unknown parameters when the null is true. For example, the t-statistic for testing a hypothesis about the mean of a normal population or a slope

coefficient in a normal linear regression model is pivotal. However, pivotal test statistics are not available in most econometric applications unless strong distributional assumptions are made. In particular, pivotal test statistics are not available for the applications in experimental economics in this paper. For example, in the SR model, the finite-sample distribution of T_{n0} depends on the value of updating parameter φ_1 when H_0 is true.

When H_0 is composite and the test statistic is not pivotal, it is necessary to replace the true Type I critical value with an approximation. First-order asymptotic distribution theory provides one approximation. Most test statistics in econometrics are asymptotically pivotal. For example, suppose T_{n0} is asymptotically $N(0,1)$ when H_0 is true. Hence, the critical value from the standard normal distribution can be used to approximate the Type I critical value. In this case, H_0 is rejected if $| T_{n0} | > 1.96$. This test makes a Type I error with a probability that is approximately .05; in other words, the nominal level of this test is .05. The main disadvantage of this approach is that first-order asymptotic approximations can be very inaccurate with the sample sizes encountered in applications. As a result, the true and nominal probabilities that a test rejects a correct null hypothesis can be very different when the critical value is obtained from the asymptotic distribution of the test statistic. Indeed, this is the finding from our Monte Carlo investigation for tests about the parameters of the SR and BBL models.

The bootstrap provides a way to obtain approximations to the Type I critical value of a test and the probability of a Type I error that are more accurate than the approximations of first-order asymptotic theory. The bootstrap does this by using the information in the sample to estimate the parameters of the DGP and, thereby, the finite-

sample distribution of the test statistic. The bootstrap estimator of the Type I critical value is, in fact, the Type I critical value of the estimated finite-sample distribution of the test statistic. This estimated distribution is obtained by carrying out a Monte Carlo experiment in which random samples are obtained from the model with estimated parameter values instead of the true values. It turns out that if the test statistic is asymptotically pivotal and certain technical conditions are satisfied, the bootstrap approximations are more accurate than those of first-order asymptotic theory. For details, see Hall (1992) and Horowitz (1997, 1999)

In the present context bootstrap sampling can be carried out in three ways, and hence there are three ways to estimate the Type I critical value, that is, to calculate the bootstrap critical value. Suppose a sample of size n (= NxT panel) has been generated by random sampling from the SR model. Call this the estimation sample. Let $\hat{\varphi}_{n0}$ and $\hat{\varphi}_{n1}$ be the maximum likelihood (ML) estimators of φ_0 and φ_1. Suppose further that the objective is to test H_0. Then bootstrap sampling can be carried out in the following ways:

Boot1. Generate a bootstrap sample of size n by random sampling from the SR model but using the maximum likelihood estimates of the parameters from the estimation sample instead of the true values, that is, by setting $\varphi_0 = \hat{\varphi}_{n0}$ and $\varphi_1 = \hat{\varphi}_{n1}$. Using the bootstrap sample, re-estimate the parameters of the model and compute the test statistic for testing H_0. Call its value T_{n0}^1 Estimate the α-level Type I critical value of the test from the empirical distribution of T_{n0}^1 that is obtained by repeating this procedure many times. Let $z_{n0.\alpha}^1$ denote the estimated critical value.

Boot2. Generate a bootstrap sample of size n by random sampling from the SR model but using the hypothesized value of the memory parameter and the ML estimate of the updating parameter instead of the true value, that is, by setting $\varphi_0 = 0.8$ and $\varphi_1 = \hat{\varphi}_{n1}$. Using this sample re-estimate the parameters of the model and compute the test statistic for testing H_0. Call its value T_{n0}^2. Estimate the α-level Type I critical value of the test from the empirical distribution of T_{n0}^2 that is obtained by repeating this procedure many times. Let $z_{n0.\alpha}^2$ denote the estimated critical value. In this procedure, H_0 has been imposed in generating the bootstrap samples via the memory parameter.

Boot3. Generate a bootstrap sample of size n by random sampling from the SR model but using the hypothesized value of the memory parameter and the constrained ML estimate of the updating parameter instead of the true values, that is, by setting $\varphi_0 = 0.8$ and $\varphi_1 = \tilde{\varphi}_{n1}$ where $\tilde{\varphi}_{n1}$ is a constrained ML estimate of φ_1. The estimate $\tilde{\varphi}_{n1}$ is obtained by maximizing the likelihood subject to the constraint that $\varphi_0 = 0.8$. Using this sample re-estimate the parameters of the model and compute the test statistic for testing H_0. Call its value T_{n0}^3. Estimate the α-level Type I critical value of the test from the empirical distribution of T_{n0}^3 that is obtained by repeating this procedure many times. Let $z_{n0.\alpha}^3$ denote the estimated critical value. In this procedure, H_0 has been imposed directly in the selection of the value for φ_0 and indirectly in the choice of the value for φ_1.

In the Monte Carlo experiments, the null hypothesis H_0 is tested using the Wald t-statistic. The formula for calculating the Wald t-statistic is determined by the choice of the bootstrap sampling procedure. Given a bootstrap sample generated by *Boot1*, let φ_{n0}^1

and φ_{n1}^{1} be the ML estimators of φ_0 and φ_1, and let S_{n0}^{1} and S_{n1}^{1} be the square roots of the diagonal elements obtained from the inverse Hessian of the log likelihood (3) evaluated at φ_{n0}^{1} and φ_{n1}^{1}. The formula for Wald t-statistic is

$$T_{n0}^{1} = (\varphi_{n0}^{1} - \hat{\varphi}_{n0}) / S_{n0}^{1}$$

This t-statistic is centered on $\hat{\varphi}_{n0}$, not 0.8, because in the population being sampled by the bootstrap the true value of the memory parameter is $\varphi_0 = \hat{\varphi}_{n0}$.

Given a sample generated by *Boot2*, let φ_{n0}^{2} and φ_{n1}^{2} be the ML estimators of φ_0 and φ_1, and let S_{n0}^{2} and S_{n1}^{2} be the square roots of the diagonal elements obtained from the inverse Hessian of the log likelihood (3) evaluated at φ_{n0}^{2} and φ_{n1}^{2}. In this case, the formula for Wald t-statistic is

$$T_{n0}^{2} = (\varphi_{n0}^{2} - 0.8) / S_{n0}^{2}$$

This t-statistic is centered on 0.8 because in the population being sampled by the bootstrap the true value of the memory parameter is now $\varphi_0 = 0.8$. For a sample generated by *Boot3*, the analogous t-statistic is $T_{n0}^{3} = (\varphi_{n0}^{3} - 0.8) / S_{n0}^{3}$, which again is centered on 0.8 and for the same reason as in *Boot2*.

Horowitz (1997, 1999) notes that the results of Monte Carlo experiments have shown that the numerical accuracy of the bootstrap tends to be higher the more efficiently the DGP is estimated. In particular, gains in efficiency and performance can be obtained by imposing the constraints of the H_0 when obtaining the estimate of the DGP. Accordingly, we expect that *Boot3* will have the greatest numerical accuracy and that *Boot2* will usually have greater numerical accuracy than *Boot1*.

If the objective is to obtain a confidence interval for φ_0 rather than to test a hypothesis, the bootstrap can be used to reduce the error in the coverage probability. This is done using bootstrap critical values instead of asymptotic critical values in constructing the confidence interval. However, the bootstrap sampling procedures *Boot2* and *Boot3* are not available for confidence intervals.

6. DESIGN OF MONTE CARLO EXPERIMENTS

This section describes the design of the Monte Carlo experiments that investigate the ability of the bootstrap to provide improved finite-sample critical values for the Wald t-tests of the parameters of the SR and BBL models.

In the Monte Carlo experiment with the SR model the values of the parameters are $\varphi_0 = \varphi_1 = 0.8$, and, similarly, with BBL $\beta_0 = \beta_1 = 0.8$. The value of 0.8 represents a compromise between two difficulties. First, if the updating parameter φ_1 (β_1) is zero, then the memory parameter φ_0 (β_0) is not identified. Second, if, φ_1 (β_1) is "too high", the play rapidly converges to equilibrium, which means that the data provide little information about learning; in other words, there is little data for the experimentalist to learn about learning behavior.

The experiments for SR and BBL were played with a population of N senders and N receivers, where N= 6, 12, 24, 400. The first three values of N were chosen because of physical limitations of laboratory size for conducting experiments. Our experience and that of other experimenters has been that laboratory experiments with more than 30 participants are very difficult to run, and most university experimental laboratories are configured to handle a maximum of 30-40 players. The value N = 400 was selected to illustrate the speed with which the Central Limit Theorem works.

Each replication consisted of a game played for T = 20 periods. The value 20 was chosen because it is the value used in Blume et al. (1998). Another reason involves convergence bias issues. As noted in the introduction and section 3, additional observations produced by additional periods do not provide information about learning once equilibrium is reached. For the SR model, equilibrium is usually not reached by T = 20 when N = 6. However, equilibrium play occurs often before 20 periods in the case of BBL.

To simplify exposition, the Monte Carlo experiment is only described in detail for testing the memory parameter of the SR model at the nominal .05 level. There were 1,000 Monte Carlo replications in the experiment. Each replication consisted of the following steps:

1. Generate an estimation sample of size n (=NxT) by random sampling from the SR model. Estimate the unknown parameters by ML and compute the test statistic for testing H_0. Call its value T_{n0}.

2. Generate 200 bootstrap samples of size n using *Boot1*, *Boot2* and *Boot3*. Estimate the .05-level Type I critical values, denoted by $z^1_{n0.05}$, $z^2_{n0.05}$, and $z^2_{n0.05}$.

3. Reject H_0 at the nominal .05 level based on the *Boot1* critical value if $|T^1_{n0}| > z^1_{n0.05}$, on the Boot2 critical value if $|T^2_{n0}| > z^2_{n0.05}$, and on the *Boot3* critical value if $|T^3_{n0}| > z^2_{n0.05}$. Reject H_0 at the nominal .05 level based on the asymptotic critical value if $|T_{n0}| > 1.96$, which is the asymptotic .05-level critical value for the symmetrical, two-tailed t-test.

In Step 2, the 10%-level Type I critical values were also estimated using the *Boot1*, *Boot2* and *Boot3* procedures and the asymptotic distribution. These critical values were then used in Step 3 to test H_0. We also calculated nominal 95 and 90 percent confidence intervals for the memory parameter and the updating parameter using *Boot1* critical values and asymptotic critical values.

The program used to generate the estimation samples and also the bootstrap samples was written in GAUSS. For the SR model, it took approximately 17.55 seconds on a Gateway Opti Plex GX 4000 with an Intel Pentium 4 processor operating at 1.7 gigaherz running under Windows NT 4.0 to generate 1,100 estimation samples of an experiment with T = 20 and N = 6. The same program took 407.86 seconds with the same configuration to generate 1,100 replications with T = 20 and N = 400.

For each replication of the SR model, the likelihood function, for example, equation (3), was maximized using the OPTIMUM procedure in GAUSS for a variety of laboratory computers running Windows NT 4.0. The typical time in seconds to maximize all 1,100 likelihood functions was 182.51 when N = 6. OPTIMUM was called using a first-derivative quasi-Newton algorithm, BFGS (Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970)). For the SR bootstrap with N=6, *Boot1* took about 8.5 hours; *Boot1, Boot2* and *Boot3* combined took a total of about 42.5 hours.

The possibility of experimental failure is the reason that 1,100 estimation samples were generated to ensure that we had 1,000 samples for the Monte Carlo. We define an experimental failure as a situation where the ML estimates do not exist; in other words, the estimation sample provides no information about the parameter of interest. Experimental failure occurs frequently when examination is made of low frequency events. In the cases we consider, uninformative experiments do occur, but rarely. Table I shows the number of estimation samples where the ML estimates failed to exist. ML estimates do not exist in these cases because no rewards were generated in the sample. This occurs with low, but not negligible, probability in games with few participants. This is analogous to the situation in logit or probit models where the probability of success is

low and there are few observations. If experimental failure occurs, we follow the conventional procedure of dropping the sample and replacing it with a new sample, that is, dropping the estimation sample for which the ML estimates do not exist and generating a new estimation sample.

Given an estimation sample, the likelihood function was maximized starting from the true values of the parameters, namely, $\varphi_0 = \varphi_1 = 0.8$ for the SR model and $\beta_0 = \beta_1 = 0.8$ for the BBL model. The true values are used as the starting point for two reasons. One is that whatever the starting values, the likelihood function can be maximized in cases where the data is informative about the learning process. The other is that the likelihood function is not globally concave, but it is concave in expectation (Blume et al. (2002)). Experiments using simulated annealing as the optimization method have convinced us that starting at the true values led to convergence of the ML estimates in those cases where the data were informative about the learning process.

The guidelines used for the Monte Carlo estimation samples also apply to the bootstrap samples. If the ML estimates do not exist for the estimation sample, then this sample cannot be used as the basis for the bootstrap sampling. The converse is not true, however. If the ML estimates exist for the estimation sample, these ML estimates can be used to generate bootstrap samples from the SR model or the BBL model. However, the ML estimates need not exist for all bootstrap samples. If they do not exist for a bootstrap sample, that bootstrap sample is discarded and another one is generated in its place.

7. MONTE CARLO RESULTS

This section reports the results of a Monte Carlo investigation of the ability of the bootstrap to reduce the distortions in the level of a symmetrical, two-tailed t- test that occur when asymptotic critical values are used and to provide improved coverage probabilities for confidence intervals.

The test statistics employed in the experiments are asymptotically distributed as standard normal when the null hypotheses are true. Figure 1 presents the kernel estimates of the finite-sample distributions of the t-statistics for testing the parameters of the SR model. The estimated finite-sample distributions are noticeably nonnormal when N = 6 and N = 12. In these cases, the estimated distributions of the t-statistic for testing the memory parameter are skewed to the right and those for testing the updating parameter are skewed to the left. The distributions are nearly symmetric when N = 400.

The kernel estimates of the finite-sample distributions of the t-statistics for testing the parameters of the BBL model are presented Figure 2. Again, when N = 6 and N = 12, the estimated finite-sample distributions of the t-statistic for testing the memory parameter are skewed to the right and those for testing the updating parameter are skewed to the left. The main difference between the estimated distributions of the t-statistics in Figure 1 and Figure 2 when N = 6 and N = 12 is that those in Figure 2 are much more skewed. This means that the asymptotic distribution is a poorer approximation to the true, finite-sample distributions of the t-statistics in the BBL model than the SR model. Another consequence is that estimates of the Type I critical value based on the standard normal distribution will be less accurate in the case of BBL.

Table II reports the empirical rejection probabilities or levels of the nominal .05 and .10 level symmetrical, two-tailed tests when N = 6 and N = 12. The upper panel gives the empirical levels for the SR model. When asymptotic critical values are used, the empirical levels are too large, especially for the memory parameter when N = 6. Using the *Boot1* and *Boot2* critical values reduce the differences between the empirical and nominal levels for 30 of the 32 critical values, with the differences being smaller with *Boot2*. With *Boot3* critical values, the differences between the empirical and nominal levels are very small, even for N = 6. In these experiments, the *Boot3* version of the bootstrap essentially remove the distortions in the levels that occur with asymptotic critical values, and it does so at N = 6.

The lower panel of Table II presents the empirical levels for the BBL model. When asymptotic critical values are used, the empirical levels are again too large. Now the greatest differences between the empirical and nominal levels occur for the updating parameter when N = 6. The *Boot1* and *Boot2* critical values reduce the differences between the empirical and nominal levels for the updating parameter, but not for the memory parameter. In the case of the memory parameter, these differences are actually larger for *Boot2* critical values than for *Boot1* critical values. With *Boot3* critical values, the differences between the empirical and nominal levels are small when N = 6, except for the nominal .10 test of the memory parameter, and very small when N = 12. Thus, the *Boot3* critical values essentially eliminate the level distortions that occur with asymptotic critical values, in particular when N= 12.

Because the tests are symmetric, two-tailed tests, the percent of rejections corresponding to the lower- and upper-tail bootstrap critical values are of interest. These

percentages were calculated although they are not shown in Table II. It turns out that these percentages are roughly equal and of the right magnitude when *Boot3* critical values are used. For example, in the case of the nominal .05 tests for the SR model when N= 6, the empirical lower-and upper-percentages are 2.9 and 2.1 for the memory parameter and 2.8 and 2.8 for the updating parameter. For BBL, the empirical percentages are 2.1 and 2.7 for the memory parameter and 2.3 and 2.5 for the updating parameter when N = 12.

In these experiments, the numerical accuracy of the bootstrap is greater for the SR model than for the BBL model. The *Boot3* critical values essentially remove the level distortions when N = 6 for the SR model, but only when N = 12 for the BBL model. This is not surprising. There is typically more information on learning in estimation samples generated by the SR model than those generated by the BBL model. As noted earlier, this is because convergence to equilibrium play often takes place more rapidly in the BBL model; convergence typically occurs in less than 20 periods in the case of BBL, which is not the case in the SR model.

Table III reports the empirical coverage probabilities of nominal 95 and 90 percent confidence intervals for the SR and BBL models for N = 6, 12, 24 and 400. In these experiments, the empirical coverage probabilities for confidence intervals that use asymptotic critical values are too low when N = 6. Further, in this case, the differences between empirical and nominal coverage probabilities are large, especially for the memory parameter in the SR model and the updating parameter in the BBL model. The differences between the empirical and nominal coverage probabilities based on

asymptotic critical values are reduced when N = 12. There is little or no distortion in the coverage probabilities of the asymptotic confidence intervals when N = 24.

With *Boot1* critical values, the differences between the empirical and nominal coverage probabilities are reduced for the memory parameter in the SR model and the updating parameter in the BBL model when N = 6. While the differences between the coverage probabilities are reduced, *Boot1* critical values do not eliminate the distortions that occur with asymptotic critical values. What is disappointing is that when N = 12, the *Boot1* critical values do not remove the distortions in the coverage probabilities for the nominal 95 percent confidence intervals, except in the case of the updating parameter for the BBL model. The good news is that with *Boot1* critical values, the distortions are eliminated for the nominal 90 percent confidence intervals.

8. CONCLUDING COMMENTS

Experimental economics is rich in its ability to generate data under laboratory conditions. Like other experimental sciences it too is constrained in the experiments that can be performed because of physical limitations. Our Monte Carlo investigation shows that the bootstrap provides improved finite-sample critical values for the tests of the parameters of the SR and the BBL models using data from experiments with sender-receiver games. With bootstrap critical values, the differences between the empirical and nominal levels of the tests can be made very small with sample sizes available in laboratory experiments. Thus, the bootstrap provides a practical method for controlling the probability of making a Type I error in the experiments considered here.

The ability of the bootstrap to reduce the distortions in the level of tests that occur with asymptotic critical values can be investigated in other settings in experimental

economics. This paper has focused on only one of the sender-receiver games examined in Blume et al. (1998, 2002), Game1. The approach in this paper can be easily extended to investigate the performance of the bootstrap in the other games in Blume et al. (1998, 2002). Moreover, other dynamic games can be considered, for example, those proposed by Crawford (1995) and Feltovich (2000). In the game used in this paper, all players followed the same learning rule. The approach can also to be extended to situations where different players use different learning rules. In this situation, a test of the proportion of players playing a given rule is of interest. Finally, another potential application is the type of hybrid models that have been proposed by Cramer and Ho (1999).

REFERENCES

Blume, A., D.V. DeJong, Y.G. Kim and G.B. Sprinkle (1998): "Experimental evidence on the evolution of the meaning of messages in sender-receiver games," *American Economic Review,* 88, 1323-1340.

Blume, A., D.V. DeJong, G. R. Neumann and N.E. Savin (2002): "Learning and communication in sender-receiver games: an econometric investigation," *Journal of Applied Econometrics, 17*, 225-248.

Broyden, C.G. (1970): "The convergence of a class of double-rank minimization algorithms," *Journal of the Institute of Mathematics and Applications*, 6, 76-90.

Camerer, C. and T.H. Ho (1999): "Experience-weighted attraction learning in games: A unifying approach," *Econometrica*, 67, 827-874.

Crawford, V. P. (1995): "Adaptive dynamics in coordination games," *Econometrica*, 63, 103-143.

Feltovich, N. (2000): "Reinforcement-based learning models in experimental symmetric-information games," *Econometrica*, 68, 605-641.

Fisher, R. A. (1935): *The Design of Experiments. London*: Oliver and Boyd

Fletcher, R. (1970): "A new approach to variable metric algorithms," *Computer Journal*, 13, 317-322.

Goldfarb, D. (1970): "A family of variable metric updates derived by variational means," *Mathematics of Computing*, 24, 23-26.

Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Horowitz, J.L. (1997): "Bootstrap Methods in Econometrics," in *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress*, ed. by D.M. Kreps and K.R.Wallis. Cambridge: Cambridge University Press, pp. 188-222.

Horowitz, J.L. (1999): "The Bootstrap in Econometrics," in *Handbook of Econometrics*,*Vol 5*, ed. by J. J. Heckman and E. E. Leamer. Amsterdam: Elsevier, Forthcoming.

Horowitz, J.L. and N.E. Savin (2000): "Empirically relevant critical values for hypothesis tests: a bootstrap approach," *Journal of Econometrics*, 95, 375-389.

Roth, A.E. and I. Erev (1995): "Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, 8, 164-212.

Shanno, D.F. (1970): "Conditioning of quasi-newton methods of function minimization," *Mathematics of Computing*, 24, 647-656.

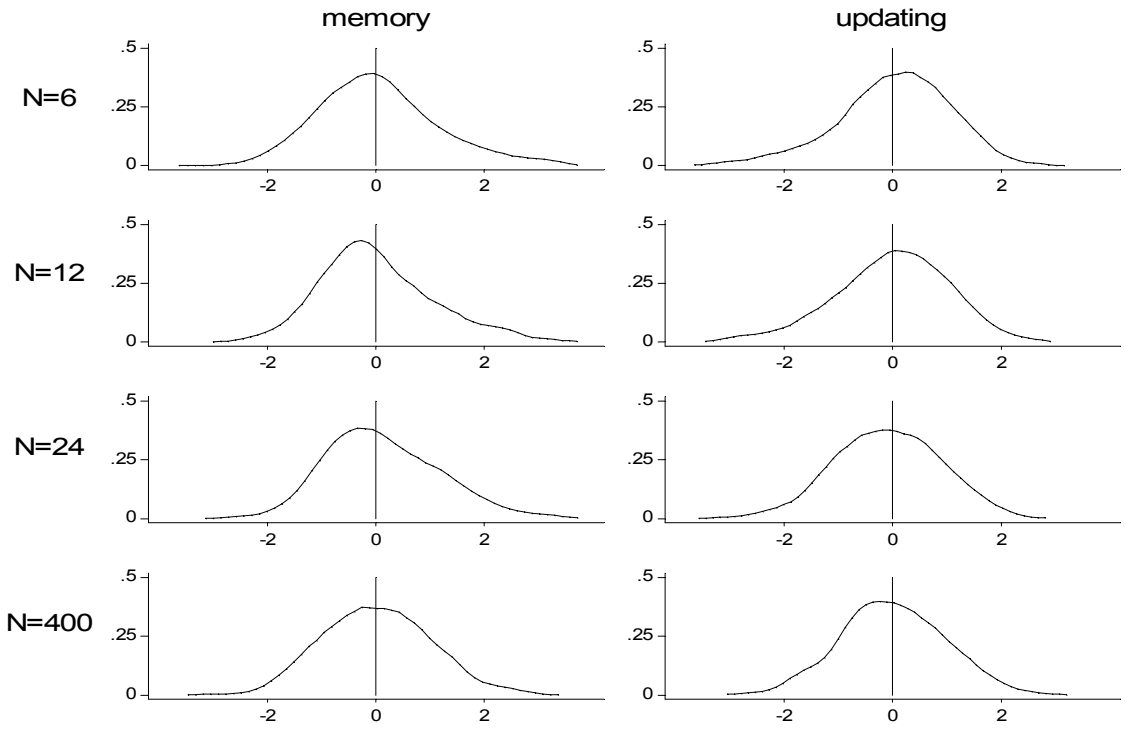

Figure 1.–Kernel estimates of the finite-sample distributions of the t-statistics for the SR model based on 1,000 Monte Carlo replications.

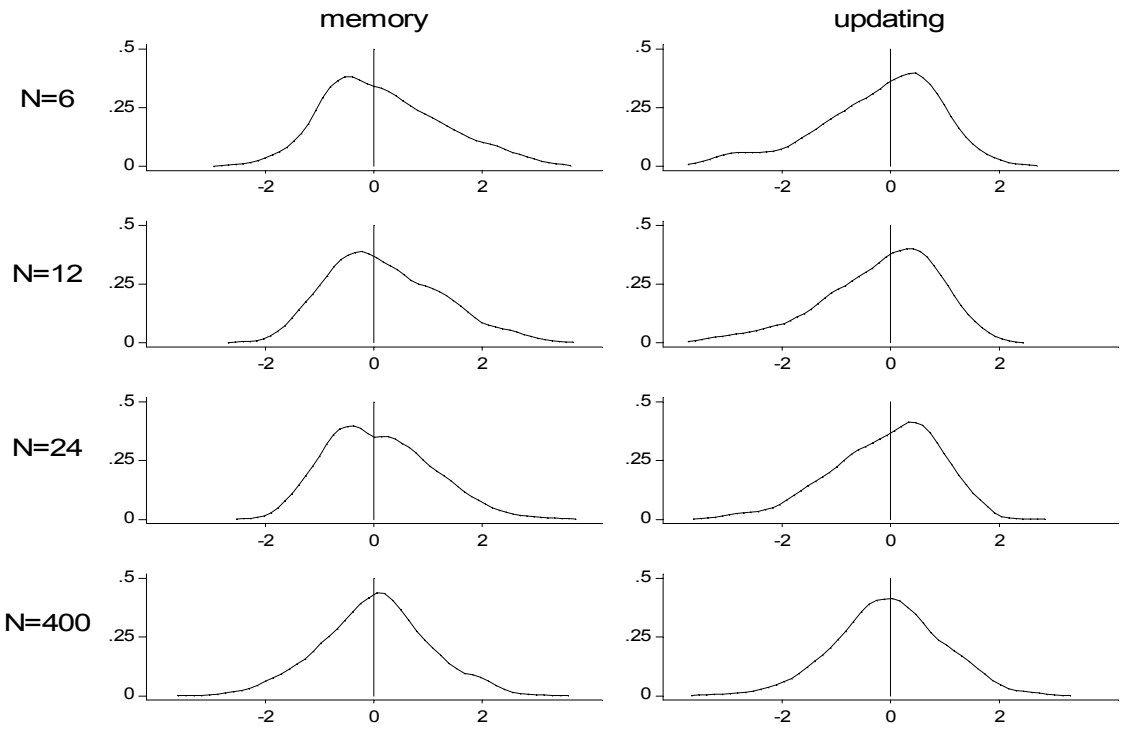

Figure 2.–Kernel estimates of the finite-sample distributions of the t-statistics for the BBL model based on 1,000 Monte Carlo replications.

TABLE I
Experimental Failures by Number of Players and Model

N	SR	BBL
6	48	41
12	14	12
24	0	1
400	0	0

Note: The table reports the number of estimation samples out of 1,100 for which the ML estimates do not exist.

TABLE II

Empirical Rejection Probabilities (Percent) of Nominal .05 and .10 Level Symmetrical, Two-tailed Tests for SR and BBL Models

	Nominal .05 Level Tests				Nominal .10 Level Tests			
	Critical Values							
N	Asymp	Boot1	Boot2	Boot3	Asymp	Boot1	Boot2	Boot3
	SR Model							
	Memory Parameter H_0: $\varphi_0 = 0.8$							
6	11.0	8.7	5.2	4.8	15.1	12.3	10.8	9.9
12	9.1	6.5	5.6	5.4	12.1	10.6	9.8	10.7
	Updating Parameter H_0: $\varphi_1 = 0.8$							
6	7.2	9.0	6.4	5.1	12.2	13.5	12.0	9.7
12	7.7	7.4	6.7	5.9	11.9	10.9	10.2	10.3
	BBL Model							
	Memory Parameter H_0: $\beta_0 = 0.8$							
6	8.8	11.9	13.6	6.4	14.3	16.4	18.4	12.2
12	6.7	6.9	8.2	4.6	10.6	10.9	13.3	8.7
	Updating Parameter H_0: $\beta_1 = 0.8$							
6	12.6	11.2	5.8	5.4	17.1	14.4	11.0	11.2
12	9.4	5.3	4.4	4.7	12.5	9.1	9.9	10.4

Notes: The empirical rejection probabilities are computed using 1,000 Monte Carlo estimation samples, each of which was used to generate 200 bootstrap samples. The 95 percent confidence intervals for the .05 and .10 levels are (3.6, 6.4) and (8.1, 11.9), respectively.

TABLE III

Empirical Coverage Probabilities (Percent) of Nominal 95 and 90 Percent Confidence Intervals for SR and BBL Models

	95 Percent Intervals		90 Percent Intervals	
	Critical Values			
N	Asymp	Boot1	Asymp	Boot1
	SR Model			
	Memory Parameter φ_0			
6	89.0	91.3	84.9	87.7
12	90.9	93.5	87.9	89.4
24	93.1		88.0	
400	95.6		91.1	
	Updating Parameter φ_1			
6	92.8	91.0	87.8	86.5
12	92.3	92.6	88.1	89.1
24	95.4		90.8	
400	96.4		90.7	
	BBL Model			
	Memory Parameter β_0			
6	91.2	88.1	85.7	83.6
12	93.3	93.1	89.4	89.1
24	95.6		91.6	
400	94.2		88.1	
	Updating Parameter β_1			
6	87.4	88.8	82.9	85.6
12	90.6	94.7	86.8	90.9
24	94.9		91.0	
400	94.8		89.8	

Notes: The empirical rejection probabilities are computed using 1,000 Monte Carlo estimation samples, each of which was used to generate 200 bootstrap samples. The 95 percent confidence intervals for the 95 and 90 percent coverage probabilities are (93.6, 96.4) and (88.1, 91.9), respectively.