# DEPARTMENT OF ECONOMICS

# DISCUSSION PAPER SERIES

**INFERENTIAL EXPECTATIONS**

**Gordon Douglas Menzies**

**Daniel John Zizzo**

# Inferential Expectations

Gordon Douglas Menzies[1]
University of Technology, Sydney
and Australian National University

Daniel John Zizzo[2]
BREB Unit, Department of Economics,
University of Oxford, Oxford OX1 3UQ, UK.

**March 2004**

**Abstract**

We propose that the formation of beliefs be treated as statistical hypothesis tests, and we label such beliefs *inferential expectations*. If a belief is overturned through the build-up of evidence, agents are assumed to switch to the rational expectation. Thus, rational expectations is a special case of inferential expectations if agents are unconcerned about mistakenly changing their beliefs (the test size $\alpha$ equals unity), or if there is so much information available about a parameter that it is known with certainty (the sampling distribution of the estimator collapses to a point). We present the results of an individual choice experiment showing preliminary support for inferential expectations in comparison to either rational expectations, or adaptive expectations with one degree of freedom. Depending on how the critical region is determined, either 27-35% or 57-65% of the agents display test size $\alpha < 0.9$. The impact of inferential expectations is illustrated by showing how it alters a simple model of the exchange rate and a Lucas supply function.

*Keywords*: expectations, macroeconomics, rationality.
*JEL Classification Codes*: C91, D84, E50, F31.

---

[1] Tel: +61-2-95147728; fax +61-2-95147711.  Email address: gordon.menzies@uts.edu.au
[2] Tel.: +44-1865-281481; fax +44-1865-286581. Email address: daniel.zizzo@economics.ox.ac.uk
Important Note: this version of the paper fully replaces the previously posted discussion paper draft, the analysis of the experimental data of which contains computational errors. Additional corrections have also been made.

# 1. Introduction

Rational Expectations (RE) applies the principle of rational behavior to the acquisition and processing of information and to the formation of expectations (Maddock and Carter, 1982). That is, economic modellers and policy-makers who use RE as a working hypothesis bestow upon their representative agents the statistical prowess necessary to behave as if they were able to calculate mathematical expectations, and, when information is limited, to calculate unbiased and efficient parameter estimators.[3]

This theory has had a central role in macroeconomics since the 1970s. It appears in Lucas' (1972) 'Islands' model, Dornbusch's (1976) 'Overshooting' model, Hall's RE Permanent Income Hypothesis (1978), Real Business Cycle theory (e.g., Kydland and Prescott, 1982), and a host of new Keynesian models (e.g., Woodford, 1991). It continues to appear in models of recent financial crises (eg. Agenor et al., 1999).

Despite its influence, the number of alleged empirical failures of RE has built up over the passage of time. In the context of consumption behavior, 'near rational' departures from RE were used to explain the excess sensitivity of consumption to anticipated changes in income (Cochrane, 1989, following Akerlof and Yellen, 1985). Similarly, evidence against Uncovered Interest rate Parity (Frankel and Rose, 1995) led some writers to suggest near rational departures from RE (Gruen and Menzies, 1995). More recently, Mankiw (2000) has argued that it is not possible to reconcile RE with inflation persistence and the observed output responses to monetary policy shocks. Moreover, RE implies a disinflationary boom in case of a fully credible disinflation announcement. His later work has tried to replace sticky prices with 'sticky information' adjustments (Mankiw and Reis, 2002, 2003).

Other negative evidence comes from experimental forecasting studies. Although RE predictions are not rejected as null hypotheses in some contexts (see Dwyer et al., 1993), the most common outcome is that individuals do not hold RE (e.g., Schmalensee, 1976; Blomqvist, 1989; Camerer, 1995; Beckman and Downs, 1997; Swenson, 1997). In addition, experimental research often finds either under-utilization or over-utilization of priors (Camerer, 1995). Some writers, perplexed by the rapid change of beliefs in the 1997 Asian financial crises, also wandered from the hypothesis of RE, speculating instead that the markets did

---

[3] The name rational *expectations* emphasises the use of mathematical expectations. But any realistic theory of 'rational' belief formation must take account of parameter estimation. Over the 1990s there were 137 papers in the *American Economic Review* that undertook estimation (Ziliak and McCloskey, in press).

not fully understand the mechanisms operating in their environment until unfolding events made them look more closely (Corbett and Vines, 1999).

In this paper we propose a simple alternative to RE for those environments where its validity is doubted. We suggest that belief formation be treated as a Neyman-Pearson hypothesis test, dubbed *Inferential Expectations* (IE). If a belief is overturned through the build-up of evidence, agents are assumed to switch to the RE of the variable. Thus, RE is a special case of IE if agents are unconcerned about mistakenly changing their beliefs (the test size $\alpha$ equals unity), or, if there is so much information available about a parameter that it is known with certainty (the sampling distribution of the estimator collapses to a point) leading to the rejection of any incorrect null.

The intuition is that economic agents hold beliefs that are subject to falsification by new information, in much the same way that they are in conventional statistical hypothesis testing. A change in beliefs thus requires new information that exceeds a threshold, modelled here by statistical significance. We assert that, at the individual level, beliefs about economic variables tend to be more subject to periods of inertia interspersed with occasional discrete shifts than what would be implied by rational expectations.

This is not a new idea in the philosophy of science (Kuhn, 1970), and it is consistent with the standard practice of most scientists, including many economists, who revise their theoretical beliefs only when classical hypotheses tests achieve statistical significance at conventional levels such as 0.05 or 0.01.[4] That being so, beliefs of economists about economics tend to be subject to periods of inertia interspersed with occasional discrete shifts. IE potentially explains deviations from strict rationality while still imposing a plausible and simple structure on expectations which is consistent with economists' own practices in statistical data analysis.[5]

One possible explanation for behavior consistent with IE is that, in many contexts, gathering and processing information to adjust beliefs may be costly relative to the marginal incentives to collect and process such information, in a way which is not very different from the adjustment of nominal prices:[6] IE may provide a 'near rational' state-dependent rule to decide whether to make the effort. In the language of Gigerenzer et al. (1999), it may embody a 'fast and

---

[4] This occurs despite the availability of an alternative approach based on Bayesian inference (e.g., Zellner, 1988). One of the authors has once had a paper rejected by a top tier journal because one of the main results was significant only at the 0.052 level.

[5] It is also consistent with modelling practice in the Markov Switching literature, where passing a threshold, however tentatively, leads to a new regime (Hamilton 1989).

[6] Romer (1996) makes this point in relation to the Lucas (1972) 'islands' model.

frugal heuristic' which may be rational in the presence of information-gathering and information-processing costs.

IE could be relevant for more sophisticated hypothesis testing, for example to detect the presence of autocorrelation, as in Rotheli (1998). Our approach is in spirit similar to Frydman and Goldberg's (1996, 2003) approach to expectations based on hypothesis testing over models.[7] Their research program, tracing its roots back at least to an informal discussion by Rappaport (1985), is more radical than ours (they allow for multiple models, and imperfect information), but at the cost of structural indeterminacy and complexity. Another way of viewing IE modelling is as complementary to Mankiw and Reis (2002, 2003) 'sticky information' approach: IE at the micro level might imply sticky information at a macro level.

We intend to explore these research avenues in the future. However, this paper has the narrow remit of explaining the idea of IE, finding some preliminary experimental evidence for it, and showing its impact on theorizing with two stylized models.

The paper is organized as follows. In section 2 we provide a general framework for IE, and demonstrate the conditions under which IE becomes RE; we illustrate our analysis with a Bernoulli data generation process, which we apply to a stylized account of currency collapse in section 3. In section 4 we outline an individual choice experiment that provides preliminary evidence for IE. In section 5 we illustrate the simplicity of using IE as a modelling tool by introducing IE into the Lucas (1972) 'islands' model. Section 6 concludes.

## 2. Inferential Expectations

In all IE models, there is a *cognitive target* (the variable or parameter that is believed to be in one of two states, described by the null hypothesis $H_0$ and the altenative hypothesis $H_1$), a *signal* (a model variable that provides information about the cognitive target), and a *test statistic* and *rejection region* which are defined conventionally.

Let $x$ be a parameter or random variable related in some way to a random variable $Y$. Granted some economic significance to $x$, agents form beliefs about it. Suppose that a data generating process for $Y$ emits $n$ stochastic signals $S_i$ (for $i = 1, \ldots, n$) which provide information about $x$. We assume that stochastic

---

[7] Our work can also be seen as related to Foster and Peyton Young's (in press) game-theoretical work on hypothesis testing by bounded-rational agents on their opponents' repeated games strategies.

signals are independent random draws from *Y*. The expectational task facing agents is to arrive at beliefs about *x* based on the signals.

The rational expectation is the mathematically best guess for *x*. The inferential expectation is the mathematically best guess for *x*, subject to a concern about changing beliefs (made operational by a Neyman-Pearson hypothesis test of size $\alpha$), and incorporating any testing shortcuts that qualify as a 'fast and frugal' heuristic. When the concern about changing beliefs becomes vanishingly small ($\alpha \to 1$), IE and RE coincide.[8] The *cognitive target* is *x*, the *signal* $S_i$, the *test statistic* some function of $S_i$, and the *rejection region* are the values of the test statistic that lead to a rejection of a Neyman-Pearson hypothesis test of size $\alpha$.

The simplest application is when *x* is the true mean of *Y*. In this case, the Maximum Likelihood (and Least Squares) estimator, $\bar{s}$, is the RE, since it is the mathematically best guess for *x*. Under RE, beliefs about *x* evolve continuously as the estimate $\bar{s}$ updates for every new signal. For example, a series of signals higher than the current sample mean will shift up the RE of *x*: see Figure 1.

(*Insert Figure 1 about here.*)

If *x* is the true mean of *Y*, and IE hold, *x* is the *cognitive target*, and the draws $S_i$ (for *i* = 1, …, *n*) are the *signals*. The *test statistic* is the sample mean $\bar{S}$ of all the $S_i$, which forms the basis for hypothesis tests about *x*.

To define the *rejection region*, we need to start with an initial belief about *x* (the null hypothesis). Assume agents first form a belief about *x* upon receipt of the *k*th signal (1 ≤ *k* < *n*), and this belief is $\bar{s}_k$ where the subscript refers to the number of individual signals used in the calculation. In Figure 1, this is the value *a*. This belief is maintained as a null hypothesis about the cognitive target. This null hypothesis is tested against the data upon the receipt of each additional signal. The null hypothesis is not rejected until the test statistic (for example, the *z* value if a normal approximation is used) passes a critical value determined by a standard statistical hypothesis test of size $\alpha$. In Figure 1, we assume that the hypothesis test is two-sided, defining the rejection region with probability $\alpha/2$ in each tail. Contingent on $\alpha$ and on the variance of the sampling distribution of $\bar{S}$ (which is decreasing in *n*), this could take many signals. If the critical value is

---

[8] This is true regardless of any shortcuts used in the testing procedure. If the test size is unity, then hypothesis testing is suspended (along with any shortcuts about distributional assumptions, etc.) because a Neyman-Pearson hypothesis test minimizes the size of the probability of a type II error given the test size (probability of a type I error). That is, given a test size of unity, the best way to minimize the chance of falsely believing the null is to *always* reject it.

reached after *i* signals, the mean is assumed to jump to the RE estimate $\bar{s}_i$. In Figure 1, the RE and IE of *x* are both equal to *c*.

Under the assumption of IE, there is first 'under-use' of information (by comparison with RE) and then 'over-use' (when beliefs change). As indicated in the introduction, RE is nested in IE in two ways:

1. RE is IE for the special case when α equals unity. This is clear from the fact that α equals unity implies a rejection for any value of $\bar{s}$. That being so, the RE belief is constantly embraced.
2. RE is IE for the special case when *x* is a parameter and *n* equals infinity (and memory is unbounded). When the number of signals is very large, the sampling distribution of $\bar{s}_n$ collapses to a single point at the true value. All nulls will be overturned (except if the null is correct), and RE beliefs embraced. In this limiting case, the RE belief will actually be the true parameter *x*.

IE requires the specification of a test size α and of a rejection region for the test given a given value of α. Standard classical inference theory (e.g., Hoel, 1984), and computational convenience - on the part both of modelling economists and presumably of economic agents -, can be used to determine the rejection region. One obvious shortcut to model rejection regions is to assume that agents use a Normal approximation.[9] Another one is to devise a statistical test based on Chebyshev's inequality.[10] Chebyshev's inequality says that:

$$P(\,|z|>k) \leq \frac{1}{k^2}$$

where *z* is a standardized random variable (the distance from the mean in units of standard deviations), and the weak inequality is relevant for a discrete random variable. If the probability of getting an observation more than *k* standard deviations away from the mean is less than $1/k^2$, we may set $1/k^2$ equal to α, and make a rare event statement.

For example, suppose a test of size α = 0.25 is required. The above inequality says that the chance of getting an observation more than 2 standard deviations away from the mean is less than 25%. Therefore, if such an observation is observed, a rare event has occurred and the belief can be changed with a chance of making a mistake (probability of a type I error) no greater than 25%.

---

[9] Naturally, when the Central Limit Theorem holds this is not a shortcut.
[10] On Chebyshev's inequality, see for example Davidson and MacKinnon (1993).

The main advantage from employing Chebyshev's inequality is that it just requires the computation of mean and variance, sidestepping the need for distributional assumptions, albeit at a loss of statistical power. Both shortcuts can be considered as consistent with a view of IE as a fast and frugal heuristic (Gigerenzer et al., 1999), and either one may be easier to work with in theoretical applications. Chebyshev's inequality will be used in our illustrative Lucas (1972) model with IE in section 5, and both IE with Chebyshev's inequality and IE with a Normal approximation will be tested experimentally in section 4.

## 3. Inferential Expectations and Sudden Currency Movements

We now illustrate our analysis using a Bernoulli data generating process to provide a stylized account of a sudden currency movement. There are three reasons for doing so. First, it may make the analysis more concrete. Second, while purely suggestive, the example of sudden currency movements is of intrinsic interest, given the difficulties in explaining currency movements with RE.[11] Third, we use an almost identical set-up in relation to the experiment described in section 4.

Let $x$ now be the probability that a currency is worthless tomorrow. Suppose a Bernoulli Data Generating Process emits $n$ stochastic signals $S_i = 1, 0$ (for $i = 1, \ldots, n$) about the value of $x$.[12] These could be Bloomberg reports, where 1 indicates that the currency will be worthless and 0 indicates that it will not. The average of the signals, $\bar{s}$, is now the proportion of ones, denoted $\hat{s}$, and $E(\hat{s}) = x$. That is, if the probability of the currency being worthless is 0.2 then, on average, twenty per cent of Bloomberg reports indicate that it will be worthless.

In this application the *cognitive target* is $x$, the *signal* is $S_i$, the *test statistic* is $\hat{s}$, and the *rejection region* is defined as it was in section 2, namely using either the Normal approximation or Chebyshev's inequality.[13] The test statistic based on $k$ signals is $\hat{s}_k$.

---

[11] One long-standing problem for RE is the vast literature debunking Uncovered Interest Parity (Frankel and Rose, 1995). Another more recent difficulty is that models of the 1997 Asian crisis struggle to reconcile RE with the apparent dearth of new information at the onset of the crisis. One response has been to speculate that agents did not understand some important relationships in the economy, but this really amounts to discarding RE (for example, Corbett and Vines, 1999). Another (controversial) approach is to propose multiple equilibria (Krugman, 1999), but this raises the question of how one moves between equilibria.

[12] Being a Bernoulli, the sample mean of these signals is an unbiased estimator for the chance of success on a single draw i.e. $E(\bar{S}) = x$. Bondarenko and Bossaerts (2000) have a somewhat similar set-up, but the parameter to be inferred is not part of the Data Generating Process for the signals.

[13] Strictly speaking, the rejection region could be determined by Binomial rejection regions for a small sample. However, unless the null belief of $x$ is 0.5, the Binomial is not symmetric, implying no

The Normal approximation method requires agents to reject the null hypothesis expectation $H_0 : x = \hat{s}_k$ (where $\hat{s}_k$ is is based on the first $k$ observations) whenever the updated proportion $\hat{s}_i$ is sufficiently different, that is when the $z$ statistic is higher in absolute terms than the critical value of $z$ for some given test size $\alpha$, i.e. $z_\alpha$:

$$\frac{\left|\hat{s}_i - \hat{s}_k\right|}{\sqrt{\dfrac{\hat{s}_k(1 - \hat{s}_k)}{i}}} > \left|z_\alpha\right|$$

If agents instead employ Chebyshev's inequality $P(|z| > k) \leq 1/k^2$, for a sample proportion we have

$$P\left(\frac{\left|\hat{s}_i - \hat{s}_k\right|}{\sqrt{\hat{s}_k(1 - \hat{s}_k)/i}} > k\right) \leq \frac{1}{k^2}$$

which, by setting $\alpha = 1/k^2$, becomes

$$P\left(\frac{\left|\hat{s}_i - \hat{s}_k\right|}{\sqrt{\hat{s}_k(1 - \hat{s}_k)/i}} > \frac{1}{\sqrt{\alpha}}\right) \leq \alpha$$

leading agents to reject the expectation $\hat{s}_i$ corresponding to the null hypothesis $H_0$ if

$$\frac{\left|\hat{s}_i - \hat{s}_k\right|}{\sqrt{\hat{s}_k(1 - \hat{s}_k)/i}} > \frac{1}{\sqrt{\alpha}}$$

In either case, only once the critical value is reached the mean jumps to the RE estimate $\hat{s}_i$.

To complete a stylized model of currency collapse, let tomorrow be the end period. The transversality condition is that the exchange rate is either unity with probability $1 - x$, or zero with probability $x$. Let interest rates in both economies be the same, and agents in both economies be risk neutral (so that they only care about expected returns). The transversality condition is that the end-period

---

uniformly most powerful test. Given a loss function, a full-blown optimization exercise could uncover an optimal rejection region, but we prefer the testing shortcuts in the text as they seem more like a 'fast and frugal' heuristic.

exchange rate equals 1 - $x$, and, by backward induction, the current value of the exchange rate must also be 1 - $x$.

Finally, suppose agents have IE about $x$ based on a given sample. A shift in their beliefs about $x$ (and hence in the current exchange rate) can occur simply because one piece of information – the proverbial straw that broke the camel's back – takes the sample proportion into the rejection region. If the null is overturned, IE agents change beliefs according to RE. If the RE belief about $x$ is much greater than the previously held belief, a currency collapse may occur as 1-$x$ shifts down. In the case of the 1997 Asian crisis, there could have been one small bit of information that, for a given rejection region, had the capacity to dramatically alter beliefs.[14]

## 4 Inferential Expectations and Experimental Evidence

### 4.1 *Introduction*

In this section we describe an individual choice experiment designed to test whether IE has significantly greater explanatory power than RE or, to put it differently, whether there are subjects for which assuming that $\alpha < 1$ provides a better fit. The individual choice design was thought to be best suited to test the idea of IE in its cleanest form, i.e. without having to worry about the strategic considerations that would arise from a strategic or market setting.

The basic structure of the experiment closely mapped the analytical framework of section 2 within an individual choice setting. There were two urns reflecting two possible states of the worlds, namely different combinations of white and orange balls. The true state of the world was chosen randomly, and subjects received signals about its nature by the means of random ball draws with replacement from the 'chosen urn'. The prior probability of an urn being chosen was 0.5 at the start of the experiment, but should have then evolved differently according to the observed sequence of white and orange balls being drawn and, importantly, according to different models of expectation formation. We next

---

[14] Two other factors may also be part of an IE-based explanation of the Asian crisis. First, agents may have become more inclined to change beliefs for a given number of signals ($\alpha$ became larger, reflecting less concern about Type I errors). Accordingly, investors could have suddenly discarded their conservatism about negative information coming out of Asia as $\alpha$ increased. Second, a deluge of information could have built up as the Asian crisis played out ($n$ increased) collapsing the sampling distributions and leading to the overturning of null beliefs. While both these explanations have some merit in describing developments towards the end of 1997, they are less compelling as an explanation of the *onset* of the crisis in the middle of that year. Radelet and Sachs (1998) argue that there was not much new information about Asia at the onset of the crisis, and, MacLeod (1998) makes the same point about Indonesia.

describe the experimental design in more detail, and then move to the experimental predictions and results.

4.2 *Experimental Design*

The experiment was run in at the School of Finance and Economics, University of Technology Sydney, in September 2003.[15] Recruitment was through lecture announcements, posters, and UTS Online (a local forum for electronic notices). Recruits were predominantly, though not exclusively, undergraduate students. There were six experimental sessions, three for each of the two experimental conditions; all sessions had six subjects except the last one, which had seven, for a total of 37 subjects. The experiment lasted about two hours, and paid an average of 31.42 Australian dollars (AUS $).[16] The experiment was in two stages, structurally unrelated to one another; in this paper we focus only on the first stage, which had six periods of fifteen rounds each and took over 75% of the session time.

At the start of the session subjects faced a table on the top of which there were two identical urns, a set of white and orange balls in a basket, and a screen. In the 0.7 condition, the experimenter (a) showed subjects that both urns were empty, (b) in front of the subjects, he took seven white balls and three orange balls and placed them in one of the two urns (Urn 1 in what follows) (c) and he took three white balls and seven orange balls and placed them in the other urn (Urn 2 in what follows); (d) he then hid both urns behind the screen. The 0.6 condition was identical to the 0.7 condition, except that Urn 1 got six white balls and four orange balls, and Urn 2 got four white balls and six orange balls.

At the start of each period subjects were reminded about the period number and then one of the two urns was randomly chosen by the flip of a coin in front of the subjects, and put on display. Let us label this urn the 'chosen urn'. It was made clear to the subjects that the probability of Urn 1 being chosen was 50% at the start of each period, but they were not told which urn had actually been chosen.

At the start of each round the experimenter drew a ball from the chosen urn, showed it to the subjects and then put it back in; subjects were asked to write down the ball color in correspondence to the correct period and round in their answer booklet, and then had to make a probability *guess*, between 0% and

---

[15] It was approved by the UTS Ethics Committee. The experimental instructions can be found in Appendix A.

[16] This is roughly equal to 25 US dollars.

100%, on how likely it was that the chosen urn was Urn 1. Subjects were told not to change choices made in previous draws.[17]

Once a period was completed, the following period got started with a new flip of the coin, up to the end of the 6th period. It was made clear to the subjects that the probability an urn was chosen was entirely independent of the probability that it had been chosen in previous periods. A questionnaire administered to the subjects at the start of the experiment ensured that this, and other key points, were clear.[18]

Payment was based on the guess made in a randomly chosen period and round picked at the end of the experiment. A standard quadratic scoring rule (e.g., Davis and Holt, 1993) was used in relation to this round to penalise incorrect answers: if the chosen urn was Urn 1, then subjects got $25 - 25 \times (guess - 1)^2$ AUS \$; if the chosen urn was Urn 2, then subject got $25 - 25 \times guess^2$ AUS \$. Subjects were provided with a payment table detailing the payment for each level of error, without need of any computation on their part (see Appendix A). There was also a participation fee of 8 AUS \$.

## 4.3 *Experimental Predictions*

*Rational Expectations*. The prior probability was set at 0.5. As information flowed in, RE (or, equivalently, IE with $\alpha = 1$) predicted straightforward Bayesian updating depending on whether white or orange balls were drawn.

*Inferential Expectations*. The IE signal is the drawn ball; that is, $s_i = 1$ for a white ball and 0 for an orange ball. The cognitive target is the probability that the chosen urn is Urn 1, or, relatedly, the total probability of drawing a white ball, given the beliefs about the chosen urn, denoted $p_w$.[19] It is simpler to describe the test with the second cognitive target (the total probability of a white ball), since the test statistic becomes the sample proportion of ones, $\hat{s}_i$. With both urns equally likely at the start, the original null is that the total probability of a white ball is 0.5.[20] The appropriate test here is a two-sided test, and we use both the Normal approximation method and the Chebyshev's inequality method to determine the rejection region (see section 2). The Normal approximation method requires agents to maintain the belief corresponding to the null hypothesis until signal $i$ is received such that

---

[17] We shall return to this point towards the end of section 4.3.

[18] The experimenters gave clarifications to the subjects who got answers wrong on the questionnaire.

[19] There is a one-to-one correspondence between the two. For the 0.7 condition, $p_w = 0.7\ P(Urn1) + 0.3\ P(Urn2) = 0.7\ P(Urn1) + 0.3\ [1 - P(Urn1)]$. A similar expression holds for the 0.6 condition.

[20] For the 0.7 condition, $0.7 \times 0.5 + 0.3 \times 0.5 = 0.5$, and, $0.6 \times 0.5 + 0.4 \times 0.5 = .5$ for the 0.6 condition.

$$\frac{\left|\hat{s}_i - p_w\right|}{\sqrt{\dfrac{p_w(1-p_w)}{i}}} > \left|z_\alpha\right| \quad where \quad p_w = P(white\,|\,Urn1)P(Urn1) + P(white\,|\,Urn2)(1 - P(Urn1))$$

If the Chebyshev's inequality method is instead used, we have

$$\frac{\left|\hat{s}_i - p_w\right|}{\sqrt{\dfrac{p_w(1-p_w)}{i}}} > \frac{1}{\sqrt{\alpha}}$$

In what follows we label $IE_N$ the predictions of IE complemented with Normal approximation and $IE_C$ the predictions of IE complemented with Chebyshev's inequality. In both cases we estimate the value of $\alpha$ corresponding to each experimental subject by using a least squares method, i.e. by minimizing the sum of squared errors between predictions and observations. That is, we consider all the choices made across rounds and periods by each subject (90 in the full sample) and we find the *subject-specific* value of $\alpha$ that mimimizes the sum of squared differences between IE ($IE_C$ or $IE_N$) and such choices.[21] These $\alpha$ values will be those used in comparing the performance of IE against RE and AE. We also employ the least squares method to estimate $\alpha$ values that best fit each period as played by each subject. That is, we consider the 15 choices made by a given subject in a given period, and we find the *period-specific* value of $\alpha$ that mimimizes the sum of squared errors between predictions and observations. Thus, for each subject there are six period-specific values of $\alpha$, and one can analyze whether these period-specific values followed any particular dynamic pattern in the experiment. These period-specific values of $\alpha$ are estimated only for the purpose of testing Hypothesis 3 below; whenever we do not specify otherwise, we shall be referring to subject-specific $\alpha$ instead.

*Adaptive Expectations.* As the IE model has one degree of freedom relative to RE, we also tested IE against another hypothesis on belief formation with one degree of freedom as well, in the form of a traditional adaptive expectations model (AE).

$$P(Urn1)_t = (1 - \beta)\,P(Urn1)_{t-1} + \beta\,s_i$$

---

[21] Or, equivalently, the mean sum of square error computed by observation; see Appendix B (section B.1) for details.

P(Urn1)$_0$ commences at 0.5 and, as before, $s_i$ takes the values zero and unity. The AE formula treats white balls as evidence for Urn 1 and orange balls as evidence for Urn 2, and is bounded between zero and unity.[22] $\beta \in [\ 0,\ 1\ ]$ provides the required degree of freedom. We determine the value of $\beta$ corresponding to each experimental subject by the least squares method.

*Experimental Hypothesis*. We can compute the expectations profile for RE, AE, IE$_N$, and IE$_C$ agents in relation to each session, using the sequence of observed ball draws and the procedures described so far. All our experimental hypotheses are designed to test the performance and robustness of IE.

HYPOTHESIS 1. IE$_N$ and IE$_C$ perform significantly better than RE and AE.

HYPOTHESIS 2. Mean $\alpha$ values do not significantly differ between the 0.6 and 0.7 conditions.

HYPOTHESIS 3. Mean period-specific $\alpha$ values tend to converge to 1 as the experiment progresses and the subjects have opportunities to learn about the nature of the task.

Hypothesis 2 is an obvious test of robustness in our $\alpha$ estimates to changes in the task they are estimated from. Hypothesis 3 is also a test of robustness, and aims to verify the absence of any obvious convergence towards greater RE play (i.e., IE with $\alpha = 1$) across the 90 rounds of the experiment. Relatedly, in testing Hypothesis 1 and 2 we used not only the 'full' sample from all six periods but also an 'experienced' sample which removes the observations from periods 1 and 2, thus allowing subjects to get some practice and experience about the nature of the task. We also considered a 'restricted' sample of observations where periods in which subjects altered their choices (notwithstanding our instructions to the contrary), and periods where some misperceptions occurred in the recording of the color of the balls, were removed.[23] Overall, in order to check the robustness of our results, we employed four samples: the full sample, the experienced sample, the restricted sample, and the experienced restricted sample.

---

[22] A drawback of this algorithm is that the steady state value of P(*Urn* 1) converges to the true probability of a white ball in the chosen urn $\hat{p}_w$, rather than to 0 or 1, the probability that Urn 1 was chosen (i.e., in the steady state $P(Urn1) = (1 - \beta)\ P(Urn1) + \beta\ s_i$ implies that $P(Urn1) = s_i$ and $E(s_i) = \hat{p}_w$ is equal to 0.7 or 0.3 for the 0.7 condition, or 0.6 or 0.4 for the 0.6 condition). However, making an adjustment so that the measure converges to 0 or 1 leads, implausibly, to a formula that could result in negative probability estimates.

[23] A total of nine periods were removed in this way, six from choice alteration and three from apparent misperception. Five of the nine periods removed were in periods 1 or 2.

4.4 *Experimental Results*

*Estimation of* α *values*. Figure 2 provides histograms for the distribution of α for both $IE_N$ and $IE_C$, in the various samples.

(*Insert Figure 2 about here*.)

In the experienced (full) sample mean α values were 0.585 (0.635) for $IE_N$ and 0.767 (0.813) for $IE_C$; mean values in the corresponding restricted samples were virtually identical. Table 1 shows the percentage of subjects displaying α < 1 and α < 0.9 in the various samples.

(*Insert Table 1 about here*.)

A non-negligible fraction of agents had α < 1 in both cases: for example, in the experienced sample, 14 out of 37 subjects (0.378) seems to have employed α < 1 for $IE_C$, a number rising to 29 out of 37 (0.757) for $IE_N$. It is interesting to look also at the α < 0.9 fraction of agents, as this may remove α estimation cases which are virtually indistinguishable, in terms of goodness of fit and predictions, from α = 1. For $IE_C$, all but one of the 11-14 subjects for whom α < 1 have also α < 0.9. For $IE_N$, over half of the subjects have α < 0.9.

The clear differences in the distributions of αs between $IE_N$ and $IE_C$ may suggest that, somewhat worryingly, $IE_N$ and $IE_C$ may bear little relation to one another. However, while there were differences in the distributions of the αs, Pearson *r* ($IE_N$, $IE_C$) is equal to 0.887, 0.843, 0.890 and 0.845 in the full, experienced, restricted, and experienced restricted samples respectively[24] (*P* < 0.001). Although $IE_N$ values tend to be lower than $IE_C$ values, $IE_N$ and $IE_C$ predictions tend to follow each other closely.

Another way of looking at the data is to constrain α to take only one possible value below 1, so as to classify subjects in one of just two categories, RE holders and IE (with α < 1) holders. Table 2 illustrates the results of this exercise.

(*Insert Table 2 about here*.)

---

[24] Spearman ρ ($IE_N$, $IE_C$) is equal to 0.822, 0.774, 0.831 and 0.775 in the full, experienced, restricted, and experienced restricted samples respectively (*P* < 0.001).

The results show that, even under the constraint of only one value $\alpha < 1$, between 16 and 22% are estimated to have IE, and $\alpha$ values are around 0.05 in six cases out of eight.

*Hypothesis 1 and related results*. We test Hypothesis 1 by computing the mean square error (MSE) values between choices and predictions according to each algorithm. One problem in doing so is the very likely non-independence of observations made by each subject; i.e., it is not meaningful to compute $F$ ratios using, say, $90 \times 37$ degrees of freedom on the numerator and denominator, as this is bound to overstate any significance of differences by treating each observation as independent.[25]

We address this problem in two ways. First, we compute MSE values *by subject* (see Appendix B for details), thus enabling us to do a $F$ test with only (37, 37) degrees of freedom: Figure 3 compares MSE values and Table 3 illustrates $F$ tests results for the various measures and samples. Second, we use MSE values computed by observation, but only as the basis for a nonparametric sign test, again only with 37 degrees of freedom.[26]

(*Insert Figure 3 and Table 3 about here.*)

Some results from the $F$ tests emerge:
(1) both $IE_N$ and $IE_C$ outperform RE and AE, and robustly so in all samples;
(2) we cannot reject the statement that $IE_C$ and $IE_N$ perform equally well;
(3) we cannot reject the statement that RE and AE perform equally well.

The results of the nonparametric sign tests are illustrated in Table 4.

(*Insert Table 4 about here.*)

$IE_N$ now outperforms $IE_C$: for example, in the experienced sample, out of 37 subjects $IE_N$ performed worse than $IE_C$ twice, tied for twenty-two subjects, and performed better 14 times ($P = 0.007$, two-tailed). $IE_C$ and $IE_N$ outperform AE at $P \leq 0.05$ in one-tailed tests in all four samples (in two-tailed tests, $P < 0.05$ in two samples and $P = 0.1$ in the other two samples). RE outperforms AE in two of the four samples. The following balance therefore emerges from the sign tests:

---

[25] With some 2000-3000 degrees of freedom on numerator and denominator, virtually any $F$ ratio different from 1 is statistically, and spuriously, significant.
[26] We perform robustness checks by running $F$ tests on MSE values computed by period (i.e., by treating observations in different periods by the same subjects as independent observations) and by running sign tests on MSE values computed by period and on MSE values computed by subject. See Appendix B for details.

(1) once again, both IE$_N$ and IE$_C$ outperform RE and AE across samples (more unequivocally so in relation to RE);
(2) IE$_N$ performs better than IE$_C$;
(3) RE may or may not outperform AE.

We conclude that the evidence from both parametric and nonparametric tests is in broad agreement in showing support for Hypothesis 1. An electronically available appendix[27] contains the mean choices and predictions according to each model of expectation formation by session and period; Figure 4 exemplifies the kind of aggregate dynamics observed by reproducing the graphs from session 4 (a 0.7 condition session).

(*Insert Figure 4 about here.*)

In period 1 RE performs better in the first seven periods, but IE does better on average afterwards. In periods 2 and 6, IE may be doing a better job in capturing the lower variability of choice relative to RE. In periods 3, 4 and 5 IE clearly does a better job at tracking mean choices than RE. Relative to RE, there appears to be a lower mean sensitivity of IE (with $\alpha < 1$) predictions to new information (though exceptions exist).

*Hypothesis 2*. As shown by Table 5, mean $\alpha$ values are surprisingly stable between the two conditions.

(*Insert Table 5 about here.*)

The greatest variability, although still a statistically non-significant one, occurs for the full sample and for the restricted sample in relation to IE$_N$. In all other cases mean $\alpha$ values are within at most 0.06 of one another. For example, in the experienced sample, in relation to IE$_N$ the mean $\alpha$ is 0.597 in the 0.6 condition and 0.573 in the 0.7 condition. Using t tests, we can never reject the hypothesis that mean $\alpha$ values are the same between conditions. This is in contrast to the instability of $\beta$ estimates in the calibration of the AE model.

*Hypothesis 3*. The fact that the evidence discussed so far is robust to the removal of the first two periods, as in the experienced sample, already suggests that IE may not tend to converge to RE as experience grows. This holds notwithstanding the fact that Figure 3 suggests, at least in relation to IE$_N$, a reduction in the noise in the data as one moves to the experienced sample.

---

[27] The appendix is at http://www.economics.ox.ac.uk/Research/Ree/MZ/MenziesZizzoWebAppendix.pdf.

Figure 5 shows the absence of any convergence to RE across the ninety rounds of the experiment. Pearson and Spearman correlation coefficients between mean $\alpha$ values and period number are never statistically significant no matter the sample, the IE measure and the correlation technique used, and are virtually equal to 0: they range between – 0.021 and 0.012.

(*Insert Figure 5 about here*.)

4.5 *Experimental Evidence: Some Conclusions*

The evidence we have presented is preliminary, being as it is from a small individual choice experiment. Nevertheless, we believe that the evidence is suggestive in a variety of ways. In a sample where AE does not fare well in comparison to RE notwithstanding its extra degree of freedom, IE does perform significantly better. Depending on how the critical region is determined, either 27-35% or 57-65% of the agents display test size $\alpha < 0.9$. If we constrain $\alpha$ to take only one possible value less than 1, between 16 and 22% of the subjects have $\alpha < 1$, and in six cases out of eight this is in the region of 0.05. It would be tempting, if probably premature in the light of the preliminary nature of this research, to link this finding to the scientific practice of awarding particular weight to the 0.05 significance level in classical hypothesis testing.

Mean $\alpha$ values were stable across conditions and across time. We found that $IE_N$ had the edge relative to $IE_C$ in the nonparametric sign tests, though no statistically significant difference can be found using $F$ tests; the sets of $\alpha$ estimates are highly correlated with one another. In any case, the design itself may have been biased *against* IE, since the only new thing happening between successive guesses was the provision of a piece of information, thereby potentially biasing subjects towards doing something with it in terms of their guesses.

## 5. Inferential Expectations and the Lucas Supply Curve

Notwithstanding the many incarnations of the Phillips curve, economists continue to believe in a short-run trade-off between inflation and unemployment (Mankiw, 2000). In this section we illustrate the impact of IE on a standard 'islands' set-up (Lucas, 1972). This is done purely to show how IE can be used as a modelling tool.

Log prices are decomposed into an average price $p$ and a relative price.

$$p_i = p_i - p + p = r_i + p \qquad\qquad p = \frac{\sum_i p_i}{n}$$

Using standard microfoundations, island labour supply is increasing in the inferred relative price, based on the revealed price $p_i$ on island $i$. We make a simplifying assumption about the parameter in the labour supply function.

$$l_i = q_i = \frac{1}{\gamma - 1} E(r_i \mid p_i) = E(r_i \mid p_i) \ \ if \ \ \gamma = 2$$

The problem is solved in Lucas (1972) by assuming a bivariate Normal distribution for the observed island price and the relative price, and finding the mathematical expectation for the relative price. The model was one of the first to include Rational Expectations (RE) into a macroeconomic setting. The Lucas supply function appears as a result of aggregation across islands.

$$E(r_i \mid p_i) = \frac{V_r}{V_r + V_p}(p_i - E(p)) \tag{1}$$

$$y = \frac{\sum q_i}{n} = \frac{\sum l_i}{n} = \frac{V_r}{V_r + V_p}(p - E(p)) \tag{2}$$

Our point of departure comes in the expectations equation 1. We give the agents the same parameter information set.[28] However, we assume that discovering that the distribution is Normal and working out the signal extraction problem involves psychic (and/or pecuniary) costs that exceed the benefits. Therefore the agents resort to a fast and frugal heuristic, by conducting a preliminary hypothesis test: they take extreme values of the island price as evidence of a relative price change.

Informally, they do not use the information of the island price unless they have evidence that it has changed in an important way. It is as if they 'turn a blind eye' to the change in the island price unless it is large. Only if this is large do they bother to do the signal extraction problem.

Formally, the *cognitive target* is *p - E(p)* (or *p*, since *E(p)* is given), the *signal* is $p_i$, the *test statistic* is also $p_i$, and the *rejection region* is defined below.

The implicit null and alternative hypotheses are:

---

[28] We assume *E(p)* is given as information, and does not have to be worked out.

$H_0$:  $p_i = E(p)$, so $p_i - E(p) = 0$
$H_1$:  $p_i = p_{i\ observed}$, so $p_i - E(p) \neq 0$.

One possible mechanism that justifies IE in this setting is that the benefits of obtaining information are positively correlated with the distance of the island price away from $E(p)$.[29]

Aiming to find an information-cheap way to decide if it is worth calculating RE, we assume that agents use Chebyshev's inequality.

$$P\left( \left| \frac{p_i - E(p)}{\sqrt{\sigma_p^2 + \sigma_{r_i}^2}} \right| \geq k \right) \leq \frac{1}{k^2} \qquad and\ \ setting\ \ \alpha = \frac{1}{k^2}$$

$$P\left( \left| \frac{p_i - E(p)}{\sqrt{\sigma_p^2 + \sigma_{r_i}^2}} \right| \geq \frac{1}{\sqrt{\alpha}} \right) \leq \alpha$$

That being the case, the rejection event $R$ ('reject $H_0$') occurs when:

$$| p_i - E(p) | \geq \sqrt{\frac{\sigma_p^2 + \sigma_{r_i}^2}{\alpha}} \ .$$

If they do reject, they then pay the costs of discovering the distribution and working out the signal extraction problem. This takes them back to the rational expectations solution.[30] Thus, when Inferential Expectations (IE) is applied to the expectations equation we obtain a modified Lucas supply curve.

$$IE(r_i \mid p_i) = \frac{V_r}{V_r + V_p}(\Phi_i - E(p))$$

$$where\ \ \Phi_i \quad = \quad p_i \quad if\ \ R \quad (i.e. \mid p_i - E(p) \mid \geq \sqrt{\frac{\sigma_p^2 + \sigma_{r_i}^2}{\alpha}})$$

$$= \quad E(p) \ \ if\ \ not-R \quad (i.e. \mid p_i - E(p) \mid < \sqrt{\frac{\sigma_p^2 + \sigma_{r_i}^2}{\alpha}})$$

---

[29] This is in keeping with the interpretation of 'sticky information' as analogous to 'sticky prices', with the costs of making an adjustment requiring to be compared to the marginal benefits, to which we already referred to in the introduction.

[30] The procedure is slightly different if the level of significance is unity. Given that a unit probability implies the null must be overturned, the RE solution is embraced, even if the rejection inequality evaluated with $\alpha$ equal to unity is not true.

$$y = \frac{\sum q_i}{n} = \frac{\sum l_i}{n} = b(IE(p) - E(p)), \qquad b = \frac{V_r}{V_r + V_p}$$

$$\textit{where} \qquad IE(p) = \frac{\sum_{i=1}^{n} \Phi_i}{n}$$

The new supply function is the same as Lucas', except that *IE(p)* has replaced the general price index *p*. The index *IE(p)* replaces any statistically insignificant island price (i.e., in *not-R*) with the true-mean price.[31] The logic is simple; non-volatile prices are simply discounted by producers when setting output.

The IE Lucas Supply Curve is a locus of aggregate-price aggregate-output pairs. For large (positive or negative) values of aggregate log price, the IE Supply curve will have less of an output response for a change in the aggregate price level, since *IE(p)* usually lies between *p* and *E(p)*.[32]

The IE Phillips curve will therefore be flatter for extreme values of *p,* giving less response in output for price changes; for values of *p* closer to *E(p),* it is less clear what the function will look like (see Figure 6).

(*Insert Figure 6 about here.*)

The interpretation is straightforward. Under IE, increases in prices in the neighbourhood of expected price do little for output because producers (who, in the monetary misperceptions story, are making incorrect output decisions) discount their relevance. However, beyond some point, producers start to take notice of the inflation and reductions in unemployment are possible. Monetary policy (or any aggregate demand policy) may have a 'band of inaction'.

We conclude this section with some comparative statics. The departures from the Lucas supply curve envisaged above become less pronounced as *IE(p)* approaches *p*. This, in turn, occurs when the hypothesis test size α approaches unity. Furthermore, if some agents use IE and some RE (i.e. IE with α = 1), *IE(p)* approaches *p* as the proportion of rational agents rises.

---

[31] As an aside, this contrasts sharply with the practice of removing volatile items from price indexes. Here, the non-volatile items are removed.

[32] This intuition follows from the *sample* version of Chebyshev's inequality, which says that no more than $100/k^2$ per cent of the data can lie more than *k sample* standard deviations away from the *sample* mean. Since *p* is a sample average, the individual $p_i$s must be clustered around *p*, with the degree of clustering determined by Chebyshev's inequality. That being so, setting individual $p_i$ draws equal to *E(p)* will tend to drag *IE(p)* closer to the true *E(p),* since many of the individual $p_i$ values are between *p* and *E(p)*. The simulations confirm this intuition.

To illustrate these implications of our model, we drew 200 Normal $p_i$ samples of for 5 islands, and calculated the Lucas Supply function, together with the IE supply function. The sample was constructed so that $p \sim N(0, 0.25^2)$.

Initially, twenty percent of the agents (that is, one in five islands) had IE with a test size of 0.25. $\alpha = 0.25$ is a conservative estimate in the light of the experimental results of section 4.4 for the case where only one value of $\alpha$ less than 1 is allowed; the percentage of 20% fits with the empirical estimation exercise (see Table 2). The results of the simulation are in the top graph of Figure 7, where $y$ is the vertical axis and $p$ is the horizontal axis.

(*Insert Figure 7 about here.*)

The intuition outlined earlier is confirmed; for large values of $p$ the IE points tend to have less response on $y$. Due to leverage, this implies that an OLS line through the IE points will be flatter.[33] In the bottom graph of Figure 7 we increase the test size to 0.75. This must imply that nulls are overturned more easily for the given draw of data, and so there is less difference between the two lines.

Finally, we make 80 per cent of agents have an $\alpha$ equal to .25, and the other 20 per cent have RE, i.e. $\alpha = 1$.

(*Insert Figure 8 about here.*)

Clearly, if most of the agents have IE, there is a considerable band of inaction for policymakers.[34]

## 6. Conclusion

This paper has presented a new model of belief formation, which we labeled *inferential expectations*. The basic idea of inferential expectations is that beliefs are maintained or revised using a Neyman-Pearson hypothesis test. They are rejected in favour of the rational expectation prediction as the new null hypothesis only when the rejection region, determined by the test size $\alpha$, is reached. This is congruent with the scientific practices of most scientists, including many economists, in forming and revising their beliefs in academic research: for example, the achievement of a 0.05 significance level is often

---

[33] The impact on the slope of an individual OLS observation depends upon the distance from the mean of the independent variable; observations furthest away from the mean have the greatest impact.

[34] We do not consider the other case when IE becomes RE. That is, when the number of pieces of information become infinite. This is so because the basic structure of the islands model prohibits it. Agents look at one piece of information only (the island price).

assigned considerable relevance to be satisfied that a particular hypothesis is supported. It is also congruent with a view of decision-making as characterized by information-gathering and information-processing costs, and hence by the usefulness of fast and frugal heuristics.

In all IE models, there is a *cognitive target* (the variable or parameter that is believed to be in one of two states, described by the null and the alternative hypothesis), a *signal* (a model variable that provides information about the cognitive target), and a *test statistic* and *rejection region* which are defined conventionally. Rational expectations is a special case of inferential expectations when the test size $\alpha$ is equal to 1, or when there is so much information available that the sampling distribution collapses to a point.

Our individual choice experiment showed a significant improvement of fit relative to rational expectations and a simple model of adaptive expectations with one degree of freedom. We used two ways of determining rejection regions, one based on a standard normal approximation and the other on Chebyshev's inequality. Both are closely related to one another, and, furthermore, $\alpha$ estimates are robust to the specific experimental condition subjects were faced with in different sessions. Across six periods of fifteen rounds each, there is no evidence suggesting a progressive convergence towards rational expectations. Depending on whether Chebyshev's inequality or a normal approximation is used, either 27-35% or 57-65% of the agents display test size $\alpha < 0.9$. We also performed an estimation exercise where we constrained $\alpha$ to take only one possible value less than 1; we found that in most cases this value was in the region of 0.05 significance level for 16-22% of the subjects.

We illustrated the implications of inferential expectations with reference to currency collapse, offering a possible mechanism for the sudden currency movements in the 1997 Asian Crisis. The most parsimonious way in which IE may have explained the evolution of beliefs in the context of this crisis is that small bits of information may have acted as the proverbial straw that broke the camel's back, by bringing agents into the rejection region and hence to reject their null hypothesis forecast about the value of the exchange rate.

We also exemplified the use of inferential expectations as a modelling tool in the context of the Lucas (1972) islands model; we found that replacing rational expectations with inferential expectations causes monetary policy (or other aggregate demand policy) to have a 'band of inaction'.

We intend to develop IE further, both theoretically and empirically. The goal of this paper was simply to define IE, present experimental evidence in favour of it, and demonstrate its potential as a modelling device.

**Acknowledgements**

## Appendix A. Experimental Instructions

<u>Instructions for 0.7 Condition</u>

**Welcome to the experiment!**

The experiment is divided into two parts, *Stage 1* and *Stage 2*. Your final winnings will be equal to the Stage 1 Payment, the Stage 2 Payment and a participation fee of 8 dollars. (All winnings will be rounded to the nearest 5 cents).

You are playing Stage 1 first. You can see two identical urns on the table, and a set of white and orange balls in a basket; you can also see a screen. The experimenter will shortly do the following:
(a) show you that the urns are empty;
(b) take **seven white balls** and **three orange balls**, and put them in one of the two urns; let us label this urn **Urn 1**;
(c) take **three white balls** and **seven orange balls**, and put them in the other urn; let us label this urn **Urn 2**;
(d) hide both urns behind the screen.

There are six *periods* in Stage 1. You have received an answer booklet with a sheet for each period.

*At the start of each period*, the experimenter announces the period number and writes it on the board. Then one of the two urns will be randomly chosen, by the flip of a coin, independently of what urns were chosen in previous periods. You will not be able to see whether this *chosen urn* is Urn 1 or Urn 2, but you will be asked to guess how likely you think it is that the chosen urn is Urn 1.

There are *sixteen draws* in each period. At the start of each draw the experimenter announces the draw number and writes it on the board. *In Draw 0*, which happens at the start of the period, your best probability guess that the chosen urn is Urn 1 would have to be 50%: this is because *at the start* of each period the chosen urn is picked randomly afresh. This Draw 0 probability guess has been printed into the answer booklet for you.

For Draws 1 through Draw 15 inclusive:

1. *first*, the experimenter draws a ball from the chosen urn and announces whether it is white or orange; *please write the ball colour on the answer sheet, in the line corresponding to the correct period and draw*; the experimenter then puts the ball back into the chosen urn;

2. *second*, you have to answer the following question: "how likely is it that the chosen urn is Urn 1? (Remember, Urn 1 is the urn with 7 white and 3 orange balls). Please choose a probability over the range 0% (definitely not) to 100% (definitely certain)"; *please put your guess in the line in the answer booklet corresponding to the correct period and draw*.

At the end of the period the experimenter hides the chosen urn again behind the screen. If you are in periods 1 through 5, you should move on to the answer sheet for the following period.

If you are in period 6, please wait until the sheets are collected and the material for Stage 2 is distributed.

*Stage 1 Payment*. It is important that you try to make your best probability guesses, both because it is important for the value of the experiment, and because your final winnings depend on it. At the end of the experiment the experimenter will randomly choose a winning draw to reward your performance. The experimenter will roll a die to choose the period, and pick randomly from a third urn (with balls numbered between 1 through 15) to choose the winning draw. Your Stage 1 Payment will depend on your choice in the draw corresponding to the number on the ball which has been picked. In relation to this draw, the experimenter will take your choice and compare it with the true chosen urn for that draw. If in the winning draw the chosen urn was Urn 1, then the correct probability of the chosen urn being Urn 1 is 100%; if the chosen urn was Urn 2, then the correct probability of the chosen urn being Urn 1 is 0%. Your Stage 1 Payment will then be equal to

$$25 - 25 \times (guess - correct\ probability)^2$$

that is, to 25 dollars minus a penalty. The penalty will be equal to the square of the error, that is of difference between the guess and the correct probability, multiplied by 25. The Stage 1 Payment will be higher the more correct your guess is. The enclosed table provides Stage 1 Payment values corresponding to some possible error levels.

Please stay seated throughout the experiment. It is essential, for the scientific value of the experiment, that you (a) do **not** communicate in any way with other participants during the experiment; (b) do **not** change your guesses for previous draws. You are liable to be expelled from the experiment, and forfeit all winnings (including the participation fee), if you do not comply with these simple rules.

This is an individual choice experiment: your choices have no influence on the winnings of other participants, and similarly the choices of other participants have no influence on your winnings. If you have any question, please raise your hand until an experimenter comes close to you, and then ask with a low voice. This may be a good time to ask questions, but feel free to raise your hand to ask questions at any time.

Stage 1 Payment Table

**Payment = 25 – 25 × (*guess – correct probability*)$^2$**

| Error | Stage 1 Payment | Error | Stage 1 Payment | Error | Stage 1 Payment |
|---|---|---|---|---|---|
| 0% | 25 | 34% | 22.11 | 68% | 13.44 |
| 1% | 25 | 35% | 21.94 | 69% | 13.1 |
| 2% | 24.99 | 36% | 21.76 | 70% | 12.75 |
| 3% | 24.98 | 37% | 21.58 | 71% | 12.4 |
| 4% | 24.96 | 38% | 21.39 | 72% | 12.04 |
| 5% | 24.94 | 39% | 21.2 | 73% | 11.68 |
| 6% | 24.91 | 40% | 21 | 74% | 11.31 |
| 7% | 24.88 | 41% | 20.8 | 75% | 10.94 |
| 8% | 24.84 | 42% | 20.59 | 76% | 10.56 |
| 9% | 24.8 | 43% | 20.38 | 77% | 10.18 |
| 10% | 24.75 | 44% | 20.16 | 78% | 9.79 |
| 11% | 24.7 | 45% | 19.94 | 79% | 9.4 |
| 12% | 24.64 | 46% | 19.71 | 80% | 9 |
| 13% | 24.58 | 47% | 19.48 | 81% | 8.6 |
| 14% | 24.51 | 48% | 19.24 | 82% | 8.19 |
| 15% | 24.44 | 49% | 19 | 83% | 7.78 |
| 16% | 24.36 | 50% | 18.75 | 84% | 7.36 |
| 17% | 24.28 | 51% | 18.5 | 85% | 6.94 |
| 18% | 24.19 | 52% | 18.24 | 86% | 6.51 |
| 19% | 24.1 | 53% | 17.98 | 87% | 6.08 |
| 20% | 24 | 54% | 17.71 | 88% | 5.64 |
| 21% | 23.9 | 55% | 17.44 | 89% | 5.2 |
| 22% | 23.79 | 56% | 17.16 | 90% | 4.75 |
| 23% | 23.68 | 57% | 16.88 | 91% | 4.3 |
| 24% | 23.56 | 58% | 16.59 | 92% | 3.84 |
| 25% | 23.44 | 59% | 16.3 | 93% | 3.38 |
| 26% | 23.31 | 60% | 16 | 94% | 2.91 |
| 27% | 23.18 | 61% | 15.7 | 95% | 2.44 |
| 28% | 23.04 | 62% | 15.39 | 96% | 1.96 |
| 29% | 22.9 | 63% | 15.08 | 97% | 1.48 |
| 30% | 22.75 | 64% | 14.76 | 98% | 0.99 |
| 31% | 22.6 | 65% | 14.44 | 99% | 0.5 |
| 32% | 22.44 | 66% | 14.11 | 100% | 0 |
| 33% | 22.28 | 67% | 13.78 | | |

Answer Booklet: Content of the Sheet for Each Period

When the experimenter draws a ball, write down the colour of the drawn ball in the middle column (if you find it convenient, you can just write W for white and O for orange).

How likely is it that the chosen urn is Urn 1? (Remember, Urn 1 is the urn with 7 white and 3 orange balls). Please choose a probability over the range 0% (definitely not) to 100% (definitely certain). Write down your answer in the Probability Guess column.

Do **not** change probability guesses corresponding to previous draws. If you do, you are liable to to be expelled from the experiment, and forfeit all winnings (including the participation fee).

If you discover that you have put your guesses in the wrong place (say, the wrong page or wrong row), please raise your hand.

| Draw | Drawn Ball Colour | Your Probability Guess |
|:---:|:---:|:---:|
| 0 | | 50% |
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |
| 11 | | |
| 12 | | |
| 13 | | |
| 14 | | |
| 15 | | |

Instructions for 0.6 Condition

These were identical to those for the 0.7 condition, except that 'six balls' ('6 balls') were replaced for 'seven balls' ('7 balls'), and 'four balls' ('4 balls') for 'three place' ('3 balls').

## Appendix B. Computation of mean sum of squares error and robustness analysis.

The mean sum of squares error (MSE) is equal to the sum of squares error (SSE) divided by the number of relevant datapoints.

B.1 *Sign Tests*

For the purpose of nonparametric *sign tests*, an algorithm *x* fits better than an algorithm *y* to predict the choices by a given experimental participant if she has a lower MSE if we were to predict her choices using algorithm *x* than if we were to predict her choices using algorithm *y*. Therefore we are interested in computing MSE values *at the level of* each experimental participant, i.e. at the level of the goodness of fit of each algorithm for each experimental participant.

Define *R* the number of rounds that are included in the sample for each participant; *r* = 15 the number of rounds in a given period (always 15); *q* the number of periods that are included in the sample for each participant. Clearly, *R* = *rq* by definition. There are three procedures to compute the MSE *at the level of* each experimental participant.

*MSE by observation*. It is possible to compute the mean squared difference by observation between prediction and observation. Then, in relation to each participant making a choice $p_i^{actual}$ in round *i* and to each corresponding theoretical prediction $p_i^{theory}$, it is possible to compute:

$$MSE = \frac{SSE}{R} = \frac{\sum_{i=1}^{R}(p_i^{actual} - p_i^{theory})^2}{R}$$

where *R* = *rq*: *R* = 90 for the full sample and for most subjects in the restricted sample (i.e., whenever *q* = 6); *R* = 75 for the nine subjects with contaminated periods (one each) in the restricted sample (i.e., whenever *q* = 5); *R* = 60 for the experienced sample and for most subjects in the experienced restricted sample (i.e., whenever *q* = 4); *R* = 45 for the four subjects with contaminated periods (one each) among periods 3-6 in the experienced restricted sample (i.e., whenever *q* = 3).

MSE by observation values provide the natural measure in relation to which to estimate sign tests, as in section 4.4; their mean for each sample and algorithm are illustrated in Table 3. SSE or MSE values computed by observation are also those minimized to estimate subject-specific α and β values for the IE and AE algorithms.

*MSE by period*. It is possible to compute the mean squared difference by period between predictions and observations. The relevant test statistic here is

$$MSE = \frac{SSE}{q} = \frac{\sum_{l=1}^{q}\left(\sum_{k=1}^{r} p_i^{actual} - \sum_{k=1}^{r} p_i^{theory}\right)^2}{q}$$

where $r = 15$ and the value of $q$ depends on the number of periods in the sample: $q = 6$ in the full sample and in most cases in the restricted sample; $q = 5$ for the nine subjects with contaminated periods (one each) in the restricted sample; $q = 4$ in the experienced and in most cases in the experienced restricted sample; $q = 3$ for the four subjects with contaminated periods (one each) among periods 3-6 in the experienced restricted sample.

*MSE by subject.* It is possible to compute for each subject the square of the sum of differences between predictions and observations. For the purpose of sign tests where the analysis is at the level of goodness of fit at the level of each experimental participant, the MSE by subject is, by construction, the same as the SSE by subject. The relevant test statistic is

$$MSE = SSE = \left(\sum_{i=1}^{R} p_i^{actual} - \sum_{i=1}^{R} p_i^{theory}\right)^2$$

where $R = rq$, with possible values equal to 90, 75, 60 and 45, as specified earlier under the description of the MSE by observation algorithm.

B.2 *F Tests*

For the purpose of the *F* tests, MSE and SSE, determined as for section B.1, are now *aggregated across* all 37 experimental participants. In other words,

$$SSE(F\ tests) = \sum_{37} MSE$$

$$MSE(F\ tests) = \frac{SSE(F\ tests)}{37} = \frac{\sum_{37} MSE}{37}$$

and MSE and SSE can again be computed by observation, by period or by subject (as specified in section B.1). The *F* tests in section 4.4 rely on MSE and SSE computed by subject, as these do not assume independence across observations (choices) either in the same or different periods by the same experimental participant.

B.3 *Robustness Tests*

*F* tests based on MSE and SSE computed by observation provide support for Hypothesis 1 but are meaningless, since independence across observations by the same subject, particularly in the same period, does not obviously hold, spuriously enhancing the significance of all *F* ratios (with 2000-3000 degrees of freedom, virtually any *F* ratio different from 1 is statistically significant).

*F* tests performed over MSE and SSE values by period do not require choices made by the same subject within the same period to be independent; they do still require that choices in different periods by the same subject should be treated as independent, though. Nevertheless, they can be of interest as a check of robustness of the analysis in section 4.4. Table 6 contains the results.

(*Insert Table 6 about here*.)

In agreement with Hypothesis 1, $IE_N$ and $IE_C$ always outperform RE and AE at $P = 0.001$ or better. RE outperforms AE in two samples out of four. $IE_N$ and $IE_C$ do not differ significantly in terms of goodness of fit.

As a further robustness check, we run sign tests using MSE values computed by period or by subject rather than by observation. The results are shown in Table 7.

(*Insert Table 7 about here*.)

$IE_N$ and $IE_C$ always outperform RE. They outperform AE in two samples out of four if MSE values by subject are used, and in all four samples if MSE values by period are used. $IE_N$ outperforms $IE_C$ in the MSE analysis by observation, and in two of the four samples in the MSE analysis by period. RE outperforms AE half of the times.

Overall, the robustness checks show support for Hypothesis 1, if more unequivocal in the comparison between IE and RE than in that between IE and AE. There is also some evidence suggesting that RE may be outperforming AE, and that, according to most nonparametric tests, $IE_N$ may be doing better than $IE_C$.

# References

Agenor, P., Miller, M., Vines, D., & A. Weber (1999) *The Asian Crisis: Causes, Contagion and Consequences*, Cambridge: Cambridge University Press.

Akerlof, G. A. and J. Yellen (1985) Can Small Deviations from Rationality Make Significant Differences to Economic Equilibria?, *American Economic Review* 75, 708-721.

Beckman, S. R. & D. Downs (1997) Forecasters as Imperfect Information Processors: Experimental and Survey Evidence, *Journal of Economic Behavior and Organization* 32, 89-100.

Blomqvist, H. C. (1989) The 'Rationality' of Expectations: An Experimental Approach, *Journal of Economic Psychology* 10, 275-299.

Bondarenko, O. & P. Bossaerts (2000) Expectations and Learning in Iowa, *Journal of Banking and Finance* 24, 1535-1555.

Camerer, C. (1995) Individual Choice, in J.H. Kagel & A.E. Roth (eds.), *Handbook of Experimental Economics*, Princeton: Princeton University Press, 587-703.

Cochrane, J.H. (1989) The Sensitivity of Tests of the Intertemporal Allocation of Consumption to Near-Rational Alternatives, *American Economic Review* 79, 319-337.

Corbett, J. & D. Vines (1999) The Asian Crisis: Lessons from the Collapse of Financial Systems Exchange Rates and Macroeconomic Policy, in P. Agenor, M. Miller, D. Vines, & A. Weber (eds.), *The Asian Crisis: Causes, Contagion and Consequences*, Cambridge: Cambridge University Press, 67-111.

Davidson, R. & J. G. MacKinnon (1993) *Estimation and Inference in Econometrics*, New York and Oxford: Oxford University Press.

Davis, D. D. & C. A. Holt (1993) *Experimental Economics*, Princeton: Princeton University Press.

Dornbusch, R. (1976) Expectations and Exchange Rate Dynamics, *Journal of Political Economy* 84, 1161-1176.

Dwyer, G. P. Jr., Williams A. W., Battalio R. C. & T. I. Mason (1993) Tests of Rational Expectations in a Stark Setting, *Economic Journal* 103, 586-601.

Foster, D.P. & H. Peyton Young (in press) Learning, Hypothesis Testing, and Nash Equilibrium, *Games and Economic Behavior*.

Frankel, J. A. & A. K. Rose (1995) Empirical Research on Nominal Exchange Rates, in G. Grossman & K. Rogoff, *Handbook of International Economics*, Amsterdam: Elsevier, 1689-1729.

Frydman, R. & M. D. Goldberg (2003) Imperfect Knowledge Expectations, Uncertainty-Adjusted Uncovered Interest Rate Parity, and Exchange Rate Dynamics, in P. Aghion, R. Frydman, J. Stiglitz & M. Woodford (eds.), *Knowledge, Information and Expectations in Modern Macroeconomics*, Princeton: Princeton University Press, 145-182.

Gigerenzer, G., P. M. Todd & the ABC Research Group (1999) Simple Heuristics That Make Us Smart, Oxford: Oxford University Press.

Goldberg, M. D. & R. Frydman (1996), Imperfect Knowledge and Behavior in the Foreign Exchange market, *Economic Journal* 106, 869-893.

Gruen, D. & G. D. Menzies (1995) Forward Discount Bias: Is it Near Rationality in the Foreign Exchange Market?, *Economic Record* 71, 157-166.

Hall, R. (1978) Stochastic Implications of the Life Cycle Permanent Income Hypothesis: Theory and Evidence, *Journal of Political Economy* 85, 971-987.

Hamilton, J. (1989) A New Approach to the Economic Analysis of Non-stationary Time Series Subject to Changes in Regime, *Econometrica* 57, 357-84.

Hoel, P. G. (1984) *Introduction to Mathematical Statistics*, 3rd Ed., New York: Wiley.

Kuhn, T. S. (1970) *The Structure of Scientific Revolutions*, 2nd Ed., Chicago: University of Chicago Press.

Krugman, P. (1999) Balance Sheets, the Transfer Problem and Financial Crises, *International Finance and Financial Crises; Essays in honour of Robert P. Flood Jr.*

Kydland, F. & E. C. Prescott (1982) Time to Build and Aggregate Fluctuations, *Econometrica* 50, 1345-1370.

Lucas, R. E., Jr. (1972) Expectations and the Neutrality of Money, *Journal of Economic Theory* 4, 103-124.

Maddock, R. & R. Carter (1982) A Child's Guide to Rational Expectations, *Journal of Economic Literature* 20, 39-51.

Mankiw, N. G. (2000) The Inexorable and Mysterious Tradeoff Between Inflation and Unemployment, HIER Discussion Paper 1905, Harvard University.

Mankiw, N. G. & R. Reis (2002) Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve, *Quarterly Journal of Economics* 117, 1295-1328.

Mankiw, N. G. & R. Reis (2003) Sticky Information: A Model of Monetary Nonneutrality and Structural Slumps, in P. Aghion, R. Frydman, J. Stiglitz & M. Woodford (eds.), *Knowledge, Information and Expectations in Modern Macroeconomics*, Princeton: Princeton University Press, 64-86.

MacLeod, R. (1998) Indonesia, in R. MacLeod & R. Garnaut (eds.), *East Asia in Crisis: from being a miracle to needing one?*, London: Routledge, 31-48.

Radelet, S. and J. Sachs (1998) The Onset of the East Asian Financial Crisis, NBER Currency Crisis Conference, 6-7 February 1998.

Rappoport, P. (1985) Unfalsified Expectations: An Alternative Perspective on Modelling Expectations in Macroeconomics, C.V. Starr Center for Applied Economics research report 85-16, New York University.

Romer, D. (1996) Advanced Macroeconomics, 2nd Ed., New York: McGraw Hill.

Rotheli, T. E. (1998) Pattern Recognition and Procedurally Rational Expectations, *Journal of Economic Behavior and Organization* 37, 71-90.

Schmalensee, R. (1976) An Experimental Study of Expectation Formation, *Econometrica* 44, 17-41.

Swenson, C.W. (1997) Rational Expectations and Tax Policy: Experimental Market Evidence, *Journal of Economic Behavior and Organization* 32, 433-455.

Woodford, M. (1991) Self-Fulfilling Expectations and Fluctuations in Aggregate Demand, in N. G. Mankiw & D. Romer (eds.), *New-Keynesian Economics*, vol. 2, Cambridge (MA): MIT Press, 77-110.

Zellner, A. (1988) Bayesian Analysis in Econometrics, *Journal of Econometrics* 37, 27-50.

Ziliak, S. T. & D. McCloskey (in press) Size Matters: The Standard Error of Regressions in the American Economic Review, *Journal of Socio-Economics*.

Figure 1. Example of IE dynamics (for $\alpha < 1$).
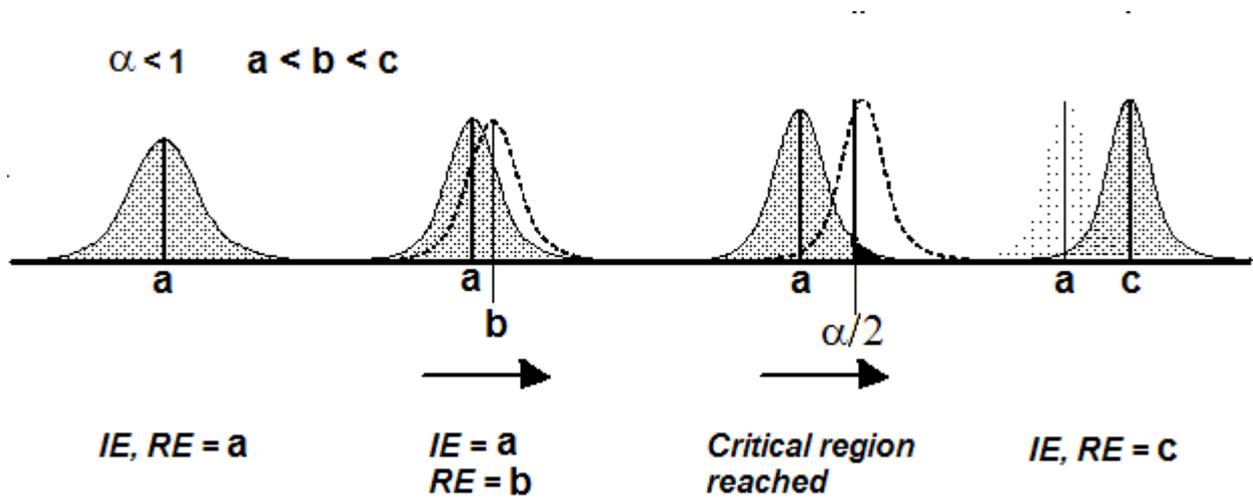


The example assumes that parcels of information having values above $c$ (where $a < b < c$) flow in sequentially. The IE ($\alpha < 1$) agent sticks with her IE = $a$ belief until the rejection region is reached. When the rejection region is reached, she switches to the RE belief $c$.

Figure 2. Histograms of α values for IE.
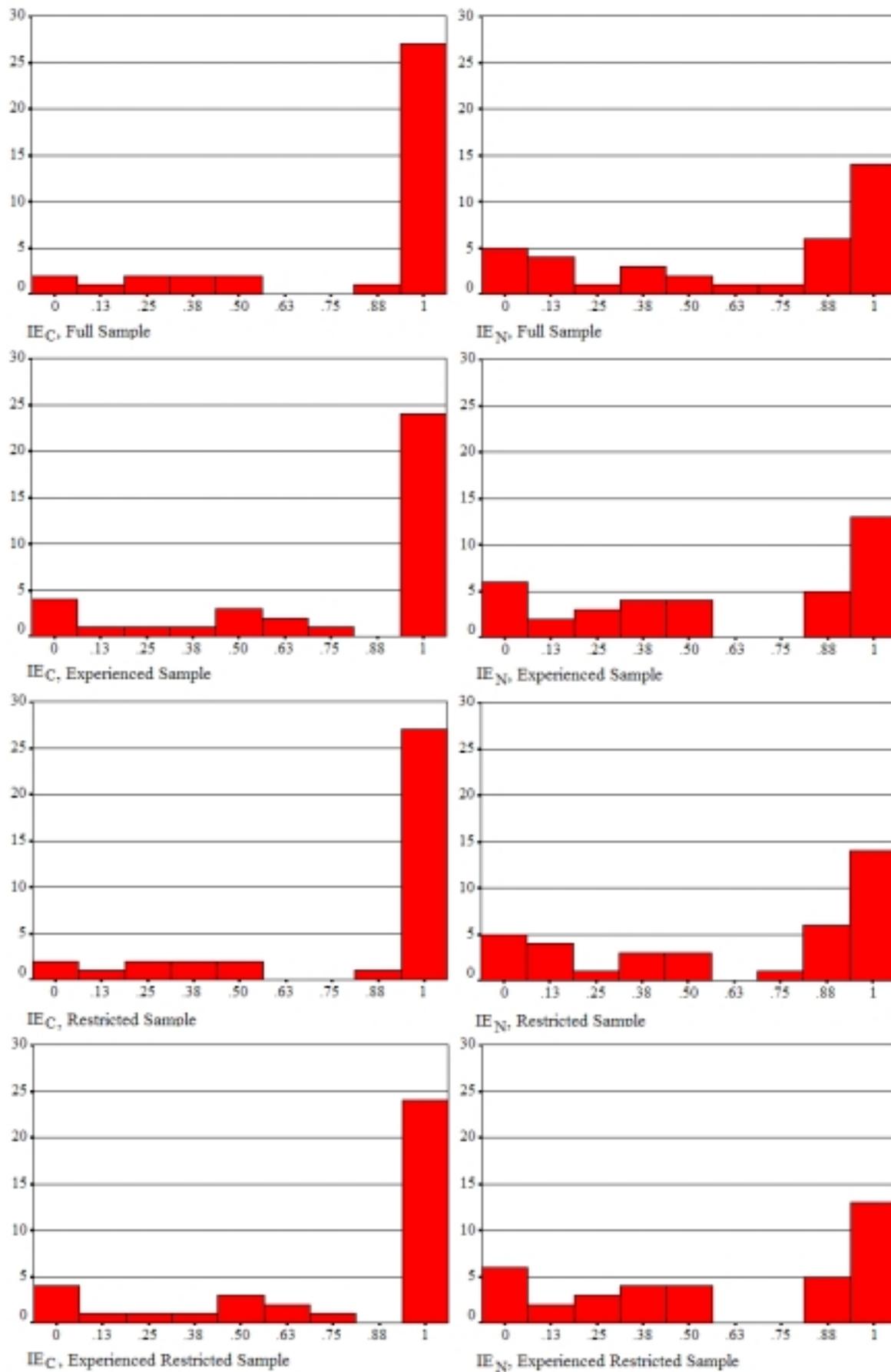
Figure 3. Goodness of fit of belief formation models: mean squared error.
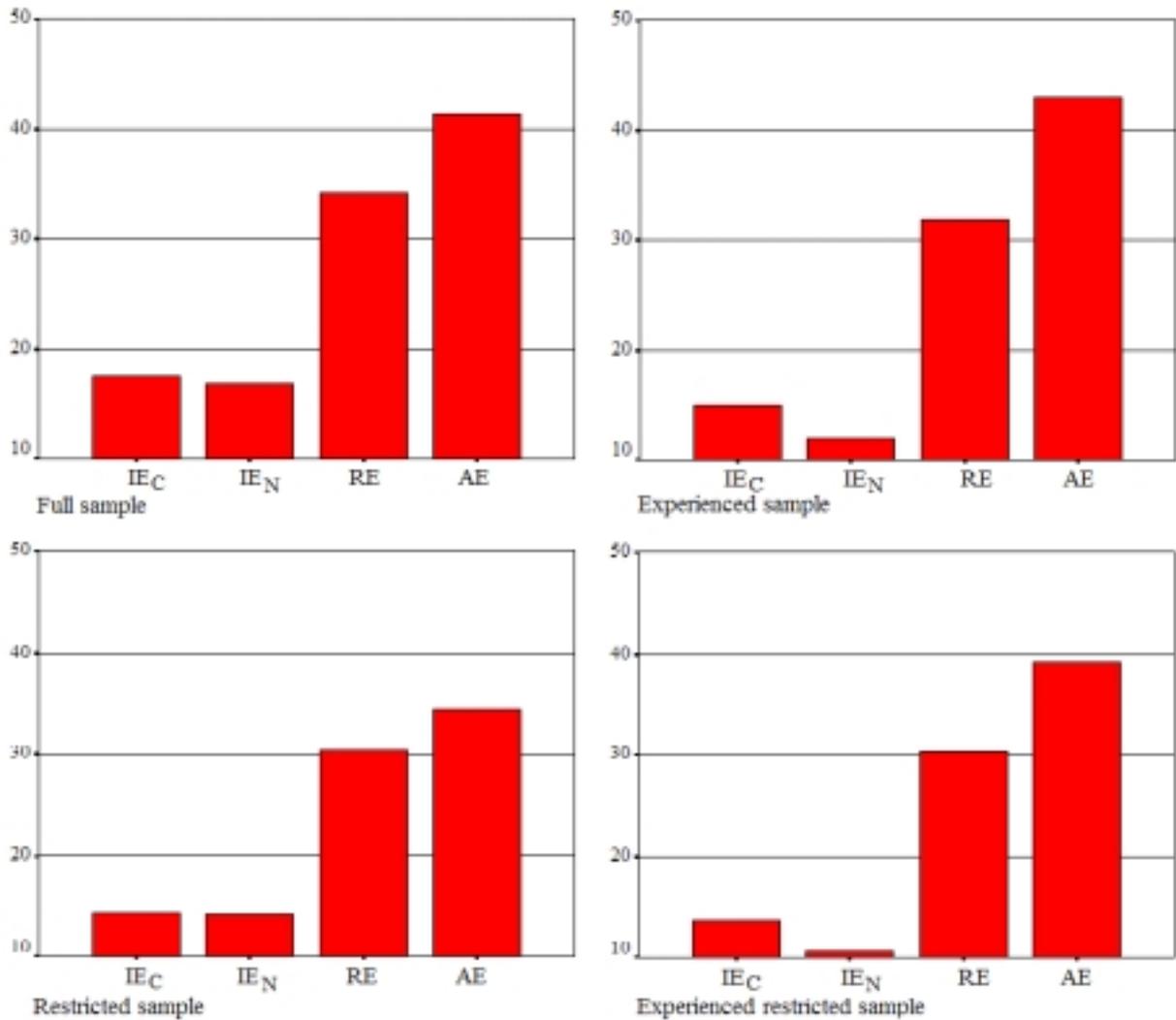


Mean sum of square error (MSE) values are computed by subject (see Appendix B, section 1) and averaged out across all experimental participants for each sample and algorithm.

Figure 4. Full sample mean choices and predictions: examples.
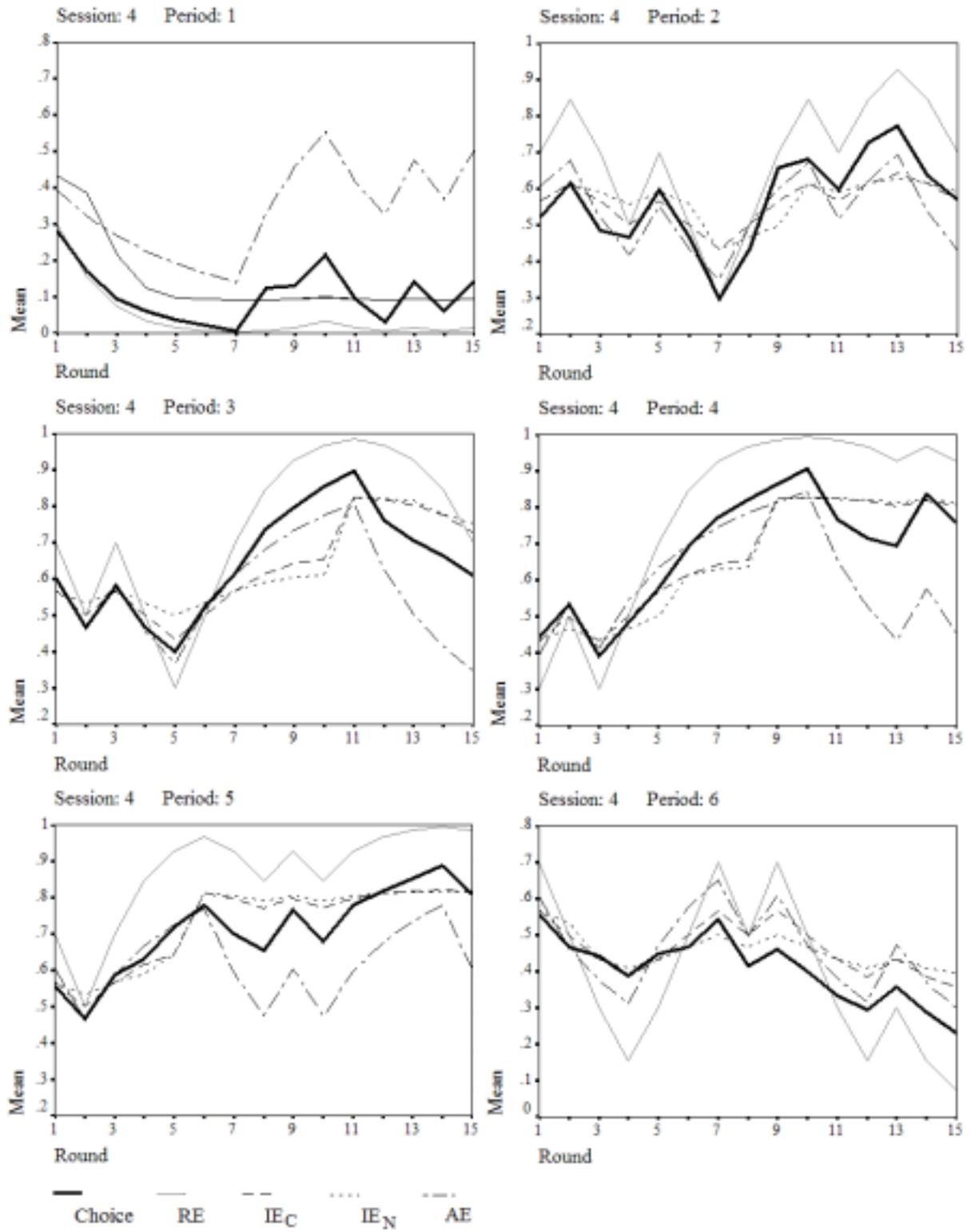
Figure 5. Mean α estimates by experimental period.

Full Sample



Restricted Sample
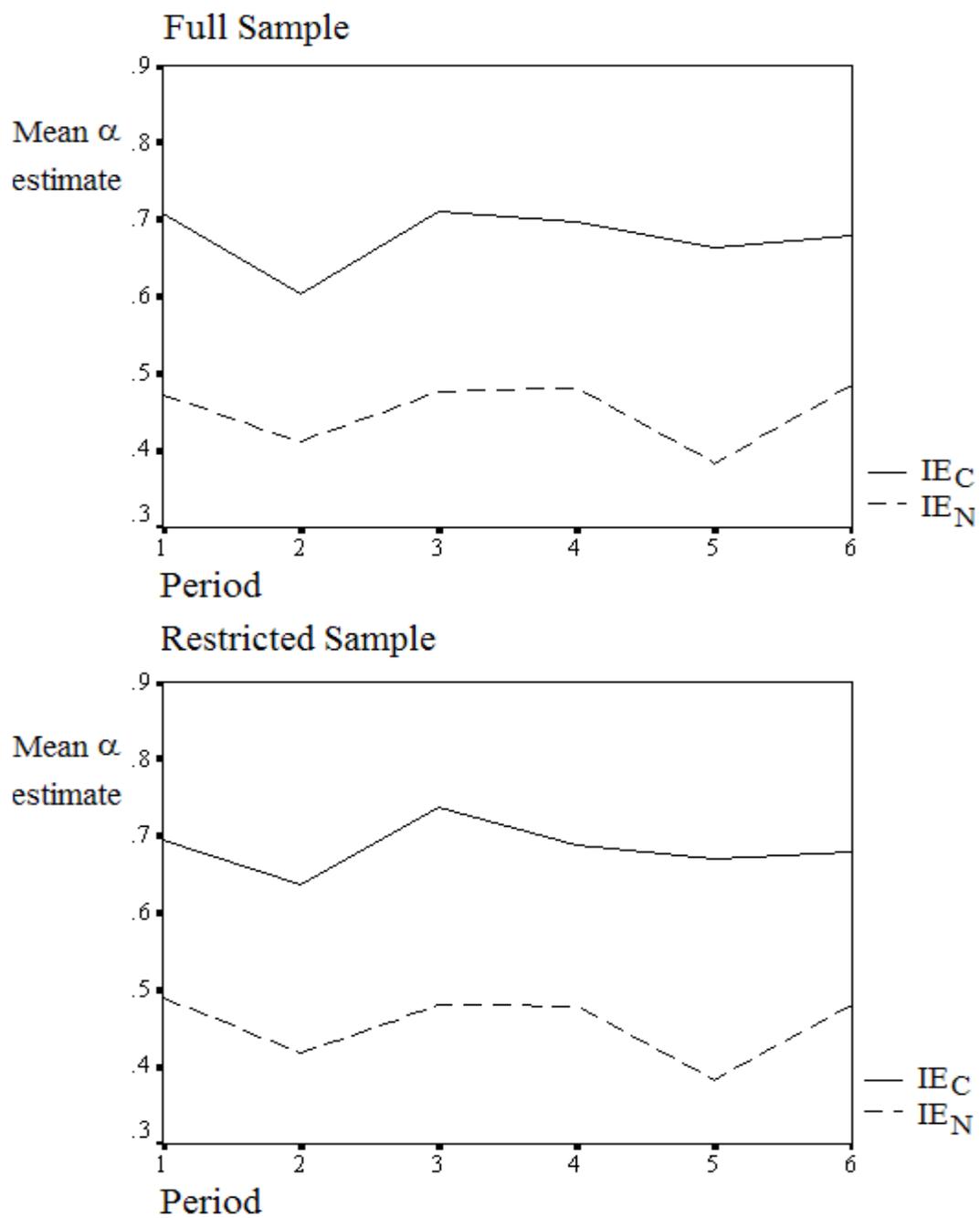


Figure 6. Comparison of Lucas with the IE Supply curve
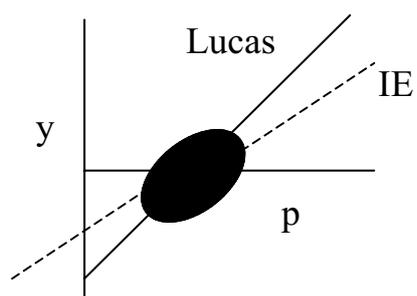
Figure 7. Computer simulations with 20% IE agents.



p~N(0, .0625)
(n=5, 20% IE agents, α=.25)



p~N(0, .0625)
(n=5, 20% IE agents, α=75)
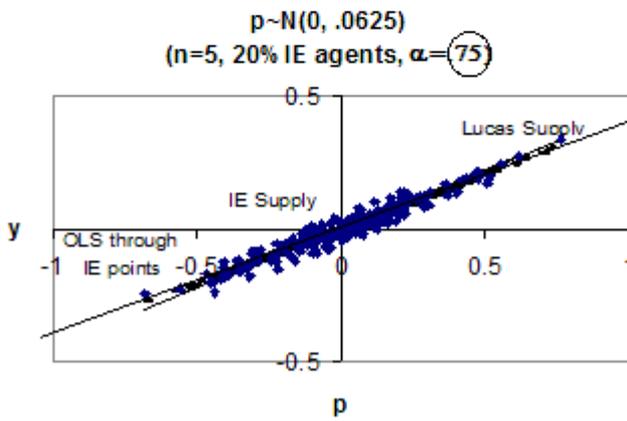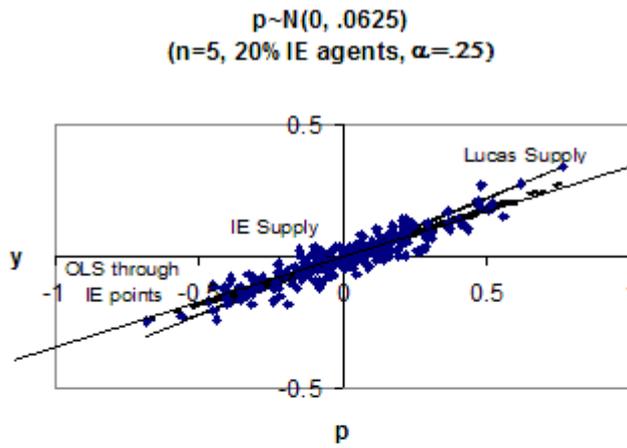
Figure 8. Computer simulation with 80% IE agents.



p~N(0, .0625)
(n=5, 80% IE agents, α=.25)
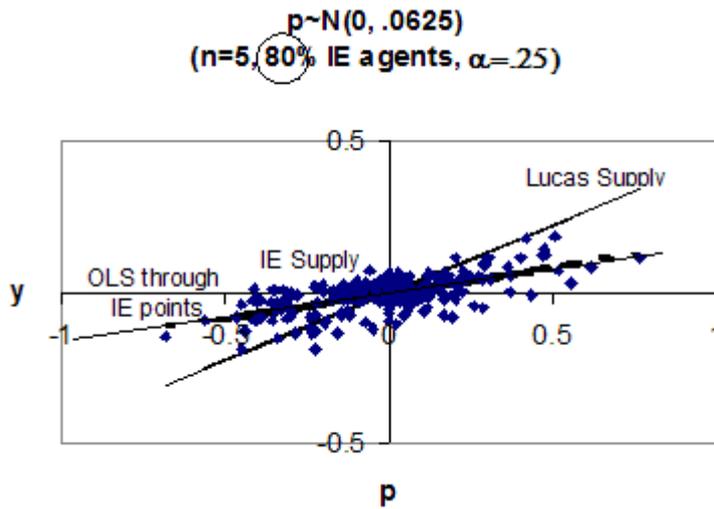
Table 1. Percentage of Subjects with $\alpha < 1$ or 0.9

| | | Sample | | | |
|---|---|---|---|---|---|
| | | Full | Exp | Restr | Exp+Restr |
| $\alpha < 1$ | $IE_C$ | 0.297 | 0.378 | 0.297 | 0.378 |
| | $IE_N$ | 0.757 | 0.784 | 0.757 | 0.784 |
| $\alpha < 0.9$ | $IE_C$ | 0.27 | 0.351 | 0.27 | 0.351 |
| | $IE_N$ | 0.568 | 0.649 | 0.568 | 0.649 |

The table displays the percentages of subjects, out of $n = 37$, for which $\alpha < 1$ or $\alpha < 0.9$. Full: full sample; Exp: experienced sample; Restr: restricted sample; Exp+Restr: experienced restricted sample. Percentages are computed out of $n = 37$.

Table 2. Constrained $\alpha$ Estimation.

| Sample | | $\alpha$ | %IE |
|---|---|---|---|
| Full | $IE_C$ | 0.27 | 16.216 |
| | $IE_N$ | 0.05 | 16.216 |
| Experienced | $IE_C$ | 0.01-0.09 | 16.22 |
| | $IE_N$ | 0.04 | 21.621 |
| Restricted | $IE_C$ | 0.27 | 16.216 |
| | $IE_N$ | 0.05 | 16.216 |
| Experienced Restricted | $IE_C$ | 0.01-0.09 | 16.22 |
| | $IE_N$ | 0.04 | 21.621 |

The table displays the estimated value of $\alpha$ if $\alpha$ is allowed to take only one value less than 1, and the percentage of subjects for which $\alpha < 1$ correspondingly applies. In two cases $\alpha$ is expressed as a range since any value of $\alpha$ within this range has an equally good fit.

Table 3. Goodness of fit of expectational models.

**Full Sample**

|     | SSE | RE | $IE_C$ | $IE_N$ | AE |
|-----|-----|-----|-----|-----|-----|
| RE | 1265.423 | 1.000 | 1.957* | 2.032* | 0.827 |
| $IE_C$ | 646.568 | 0.511 | 1.000 | 1.038 | 0.422 |
| $IE_N$ | 622.873 | 0.492 | 0.963 | 1.000 | 0.407 |
| AE | 1530.343 | 1.209 | 2.367** | 2.457** | 1.000 |

**Experienced Sample**

|     | SSE | RE | $IE_C$ | $IE_N$ | AE |
|-----|-----|-----|-----|-----|-----|
| RE | 1176.784 | 1.000 | 2.131* | 2.645** | 0.741 |
| $IE_C$ | 552.181 | 0.469 | 1.000 | 1.241 | 0.348 |
| $IE_N$ | 444.830 | 0.378 | 0.806 | 1.000 | 0.280 |
| AE | 1587.777 | 1.349 | 2.875*** | 3.569*** | 1.000 |

**Restricted Sample**

|     | SSE | RE | $IE_C$ | $IE_N$ | AE |
|-----|-----|-----|-----|-----|-----|
| RE | 1121.522 | 1.000 | 2.122* | 2.142* | 0.880 |
| $IE_C$ | 528.614 | 0.471 | 1.000 | 1.009 | 0.415 |
| $IE_N$ | 523.661 | 0.467 | 0.991 | 1.000 | 0.411 |
| AE | 1274.604 | 1.136 | 2.411** | 2.434** | 1.000 |

**Experienced Restricted Sample**

|     | SSE | RE | $IE_C$ | $IE_N$ | AE |
|-----|-----|-----|-----|-----|-----|
| RE | 1121.535 | 1.000 | 2.233** | 2.853*** | 0.776 |
| $IE_C$ | 502.183 | 0.448 | 1.000 | 1.278 | 0.347 |
| $IE_N$ | 393.090 | 0.350 | 0.783 | 1.000 | 0.272 |
| AE | 1445.823 | 1.289 | 2.879*** | 3.678*** | 1.000 |

The first data column contains the sum of squares error (SSE) in relation to each expectation model. The values in the 2nd to 5th column are equal to $F = a / b$, where $a$ is the mean SSE (MSE) corresponding to the row expectation model and $b$ is the mean SSE (MSE) corresponding to the column expectation model. MSE values are computed by subject (see Appendix B). *: significant at the 0.05 level; ** : significant at the 0.01 level; ***: significant at the 0.001 level.

Table 4. Goodness of fit of expectational models using sign tests.

**Full Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| IE$_C$ | IE$_N$ | 1 | 23 | 13 | 0.002 |
| IE$_C$ | RE | 11 | 26 | 0 | 0.001 |
| IE$_C$ | AE | 26 | 0 | 11 | 0.021 |
| IE$_N$ | RE | 23 | 14 | 0 | 0 |
| IE$_N$ | AE | 26 | 0 | 11 | 0.021 |
| RE | AE | 25 | 0 | 12 | 0.049 |

**Experienced Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| IE$_C$ | IE$_N$ | 2 | 22 | 13 | 0.007 |
| IE$_C$ | RE | 14 | 23 | 0 | 0 |
| IE$_C$ | AE | 24 | 0 | 13 | 0.1 |
| IE$_N$ | RE | 24 | 13 | 0 | 0 |
| IE$_N$ | AE | 24 | 0 | 13 | 0.1 |
| RE | AE | 23 | 0 | 14 | 0.188 |

**Restricted Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| IE$_C$ | IE$_N$ | 1 | 23 | 13 | 0.002 |
| IE$_C$ | RE | 11 | 26 | 0 | 0.001 |
| IE$_C$ | AE | 26 | 0 | 11 | 0.021 |
| IE$_N$ | RE | 23 | 14 | 0 | 0 |
| IE$_N$ | AE | 26 | 0 | 11 | 0.021 |
| RE | AE | 25 | 0 | 12 | 0.049 |

**Restricted Experienced Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| IE$_C$ | IE$_N$ | 0 | 23 | 14 | 0 |
| IE$_C$ | RE | 13 | 22 | 2 | 0.007 |
| IE$_C$ | AE | 24 | 0 | 13 | 0.1 |
| IE$_N$ | RE | 24 | 13 | 0 | 0 |
| IE$_N$ | AE | 24 | 0 | 13 | 0.1 |
| RE | AE | 23 | 0 | 14 | 0.188 |

Each cell value indicates the number of subjects for which, in relation to any two algorithms $x$ and $y$ (and a given sample), $MSE(x) < MSE(y)$, or $MSE(x) = MSE(y)$, or $MSE(x) > MSE(y)$, where MSE is the mean sum of squares error by observation: see Appendix B. $P$ values are approximated to three decimal places.

Table 5. Mean $\alpha$ and $\beta$ value by condition.

| Mean $\alpha$ or $\beta$ | Full Sample | | | Experienced Sample | | |
|---|---|---|---|---|---|---|
| | Condition | | | Condition | | |
| | 0.6 | 0.7 | t | 0.6 | 0.7 | t |
| $IE_C$ mean $\alpha$ | 0.819 | 0.808 | 0.096 | 0.736 | 0.796 | 0.502 |
| $IE_N$ mean $\alpha$ | 0.703 | 0.571 | 1.022 | 0.597 | 0.573 | 0.183 |
| AE mean $\beta$ | 0.136 | 0.225 | 3.204** | 0.121 | 0.229 | 3.812*** |
| | Restricted Sample | | | Experienced Restricted Sample | | |
| | Condition | | | Condition | | |
| Mean $\alpha$ or $\beta$ | 0.6 | 0.7 | t | 0.6 | 0.7 | t |
| $IE_C$ mean $\alpha$ | 0.82 | 0.808 | 0.111 | 0.736 | 0.796 | 0.502 |
| $IE_N$ mean $\alpha$ | 0.696 | 0.571 | 0.959 | 0.597 | 0.573 | 0.183 |
| AE mean $\beta$ | 0.136 | 0.226 | 3.245** | 0.121 | 0.23 | 3.877*** |

The table contains the mean of the $\alpha$ ($\beta$) values estimated for each subject and sample in relation to $IE_C$ and $IE_N$ (AE), by experimental condition. t tests estimate the significance of the difference between each pair of $\alpha$ or $\beta$ estimates. **: significant at the 0.01 level; ***: significant at the 0.001 level.

Table 6. Goodness of fit of expectational models using MSE by period.

**Full Sample**

| | SSE | RE | $IE_C$ | $IE_N$ | AE |
|---|---|---|---|---|---|
| RE | 754.606 | 1.000 | 1.615*** | 1.757*** | 0.724 |
| $IE_C$ | 467.114 | 0.619 | 1.000 | 1.088 | 0.448 |
| $IE_N$ | 429.480 | 0.569 | 0.919 | 1.000 | 0.412 |
| AE | 1041.742 | 1.381** | 2.230*** | 2.426*** | 1.000 |

**Experienced Sample**

| | SSE | RE | $IE_C$ | $IE_N$ | AE |
|---|---|---|---|---|---|
| RE | 514.228 | 1.000 | 1.717*** | 2.135*** | 0.834 |
| $IE_C$ | 299.496 | 0.582 | 1.000 | 1.244 | 0.486 |
| $IE_N$ | 240.839 | 0.468 | 0.804 | 1.000 | 0.391 |
| AE | 616.705 | 1.199 | 2.059*** | 2.561*** | 1.000 |

**Restricted Sample**

| | SSE | RE | $IE_C$ | $IE_N$ | AE |
|---|---|---|---|---|---|
| RE | 714.673 | 1.000 | 1.637*** | 1.807*** | 0.719 |
| $IE_C$ | 436.692 | 0.611 | 1.000 | 1.104 | 0.439 |
| $IE_N$ | 395.502 | 0.553 | 0.906 | 1.000 | 0.398 |
| AE | 994.479 | 1.392** | 2.277*** | 2.514*** | 1.000 |

**Experienced Restricted Sample**

| | SSE | RE | $IE_C$ | $IE_N$ | AE |
|---|---|---|---|---|---|
| RE | 502.767 | 1.000 | 1.739*** | 2.179*** | 0.844 |
| $IE_C$ | 289.104 | 0.575 | 1.000 | 1.253 | 0.485 |
| $IE_N$ | 230.706 | 0.459 | 0.798 | 1.000 | 0.387 |
| AE | 595.631 | 1.185 | 2.060*** | 2.582*** | 1.000 |

MSE: mean sum of squares error *by period*. *P* values are approximated to three decimal places. **: significant at the 0.01 level; ***: significant at the 0.001 level.

Table 7. Further sign tests of the goodness of fit of expectational models.

**Full Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 1 | 23 | 13 | 0.002 |
| $IE_C$ | RE | 11 | 26 | 0 | 0.001 |
| $IE_C$ | AE | 27 | 0 | 10 | 0.009 |
| $IE_N$ | RE | 23 | 14 | 0 | 0 |
| $IE_N$ | AE | 27 | 0 | 10 | 0.009 |
| RE | AE | 25 | 0 | 12 | 0.049 |

**Experienced Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 1 | 14 | 22 | 0.001 |
| $IE_C$ | RE | 12 | 23 | 2 | 0.013 |
| $IE_C$ | AE | 25 | 0 | 12 | 0.049 |
| $IE_N$ | RE | 24 | 13 | 0 | 0 |
| $IE_N$ | AE | 25 | 0 | 12 | 0.049 |
| RE | AE | 21 | 0 | 16 | 0.511 |

**Restricted Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 1 | 23 | 13 | 0.002 |
| $IE_C$ | RE | 11 | 26 | 0 | 0.001 |
| $IE_C$ | AE | 27 | 0 | 10 | 0.009 |
| $IE_N$ | RE | 23 | 14 | 0 | 0 |
| $IE_N$ | AE | 27 | 0 | 10 | 0.009 |
| RE | AE | 25 | 0 | 12 | 0.049 |

**Restricted Experienced Sample**

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 1 | 14 | 22 | 0.001 |
| $IE_C$ | RE | 12 | 23 | 2 | 0.013 |
| $IE_C$ | AE | 25 | 0 | 12 | 0.049 |
| $IE_N$ | RE | 24 | 13 | 0 | 0 |
| $IE_N$ | AE | 25 | 0 | 12 | 0.049 |
| RE | AE | 21 | 0 | 16 | 0.511 |

MSE: mean sum of squares error *by period*. P values are approximated to three decimal places.

### Full Sample

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 7 | 23 | 7 | 1 |
| $IE_C$ | RE | 10 | 26 | 1 | 0.012 |
| $IE_C$ | AE | 20 | 0 | 17 | 0.742 |
| $IE_N$ | RE | 17 | 14 | 6 | 0.035 |
| $IE_N$ | AE | 20 | 0 | 17 | 0.742 |
| RE | AE | 19 | 0 | 18 | 1 |

### Restricted Sample

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 7 | 23 | 7 | 1 |
| $IE_C$ | RE | 10 | 26 | 1 | 0.012 |
| $IE_C$ | AE | 21 | 0 | 16 | 0.511 |
| $IE_N$ | RE | 17 | 14 | 6 | 0.035 |
| $IE_N$ | AE | 21 | 0 | 16 | 0.511 |
| RE | AE | 19 | 0 | 18 | 1 |

### Experienced Sample

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 2 | 22 | 13 | 0.007 |
| $IE_C$ | RE | 11 | 23 | 3 | 0.057 |
| $IE_C$ | AE | 27 | 0 | 10 | 0.009 |
| $IE_N$ | RE | 20 | 13 | 4 | 0.002 |
| $IE_N$ | AE | 27 | 0 | 10 | 0.009 |
| RE | AE | 26 | 0 | 11 | 0.021 |

### Restricted Experienced Sample

| x | y | MSE(x)<MSE(y) | MSE(x)=MSE(y) | MSE(x)>MSE(y) | 2-tail P |
|---|---|---|---|---|---|
| $IE_C$ | $IE_N$ | 2 | 22 | 13 | 0.007 |
| $IE_C$ | RE | 11 | 23 | 3 | 0.057 |
| $IE_C$ | AE | 27 | 0 | 10 | 0.009 |
| $IE_N$ | RE | 20 | 13 | 4 | 0.002 |
| $IE_N$ | AE | 27 | 0 | 10 | 0.009 |
| RE | AE | 26 | 0 | 11 | 0.021 |

MSE: mean sum of squares error *by subject*. *P* values are approximated to three decimal places.