Econometrica, Vol. 44, No. 1 (January, 1976)

AN EXPERIMENTAL STUDY OF EXPECTATION FORMATION

BY RICHARD SCHMALENSEE[1]

This paper reports on an experimental study of expectation formation and revision in a time series context. In an adaptive expectations framework, it is shown that the speed of adjustment seems to fall in turning point periods. Expectations are considered as probability density functions, and a scoring system is devised and employed that gives subjects an incentive to report a measure of the dispersion of these functions. This measure, which is inversely related to the confidence with which expectations are held, seems to be inversely related to past forecasting performance.

1. INTRODUCTION

THIS PAPER REPORTS an empirical exploration of the way individuals form and hold expectations about future values of time series variables. In the application of economic models in which expectations about the future play a major role in determining behavior, these expectations are rarely directly observable, and the econometrician is generally forced to assume that a technical rule generates expectations as a simple function only of past observations. One way to see what sort of technical rules make sense in such applications might be to attempt to use this indirect approach to discriminate among possible functional forms. Usually, however, this is computationally burdensome and not terribly revealing. Another approach, currently receiving attention, involves direct analysis of real-world expectations data.[2] A third approach, and the one followed here, is to create and analyze an experimental situation in which the rule followed must be technical because no information other than the past history of the time series in question is available.

The main reason for the attractiveness of the experimental approach here, however, lies in the two aspects of expectation formation with which this study is principally concerned. The first of these concerns the influence of turning points in a time series context. The basic hypothesis is due to F. M. Fisher [9, p. 48]:

> ... a plausible way for decision makers to behave when frequent policy revisions are costly is to pay attention (perhaps unconsciously) ... to the turning points in the variable or variables whose future behavior they desire to predict. We argued that this was plausible because turning point years are interesting years in economic time—they contain information as to the general course of the variable in question. Further, they stand out—they subconsciously suggest themselves to the eye and the mind of the decision maker.

[1] I am indebted to Jacques Drèze, Richard Emmerson, John Hooper, Wolfhard Ramm, Dennis Smallwood, and a referee for helpful comments, to the University of California, San Diego, Department of Economics and the Academic Senate Committee on Research for financial support, and to Harold Nelson for assistance with the U.C.S.D. version of the Econometric Software Package. I would also like to thank those who commented on an earlier version of this study when it was presented to seminars at the University of California, San Diego, and the University of Minnesota, and at the 1971 Winter Meetings of the Econometric Society in New Orleans. I, of course, retain responsibility for any shortcomings that remain.

[2] See, for instance, Hirsch and Lovell [16] and Turnovsky [25].

The notion that turning point periods are special has not been much explored; I am aware of no empirical studies relying on this idea except those presented by Fisher in [9].

Fisher tested his hypothesis by means of an experiment in which the costs of making decisions were large.[3] It is of some interest to see if turning points are also special when the costs of decision making are small. We investigate this point below. We also see if turning points still seem to stand out when explicit account is taken of the possible operation of more commonly assumed expectation formation mechanisms. As discussed below, it is convenient for these purposes to use Fisher's time series.

The second aspect of expectational behavior with which we are concerned focuses on the form in which expectations about the future are held. The basic concept was stated clearly by J. R. Hicks [13, p. 125] in 1939:[4]

> ... people rarely have *precise* expectations at all. They do not expect that the price at which they will be able to sell a particular output in a particular future week will be just so-and-so much; there will be a certain figure, or range of figures which they consider most probable, but deviations from this probable value on either side are considered to be more or less possible.

The modern theory of behavior under uncertainty certainly responds to this point; there is a large and rapidly expanding literature on behavior under uncertainty, assuming that information about the future can be described in terms of subjective probability distributions. It is by now universally recognized that parameters of these distributions other than the mean can affect behavior in important ways. In order to apply models of choice under uncertainty to real situations in a dynamic context, it is clearly necessary to understand what determines at least the location and dispersion of individuals' subjective distributions.

Yet as far as I know, empirical studies of expectation formation have all persisted in describing expectations in terms of a single number, corresponding to the location parameter of the relevant subjective distribution. No attempt has been made heretofore to measure a dispersion parameter, corresponding to the confidence with which expectations are held, or to investigate the determinants of such parameters. A first step toward filling this gap is taken here. It clearly requires an experimental approach, since no published expectations series has associated with it any index of confidence.

The next section describes the experiment conducted, the data it generated, and the basic notation employed in the remainder of the article. We then examine the determination of the experimental subjects' best estimates, their point expectations. In Section 4, we analyze the confidence attached to those forecasts. The article concludes with a brief summary of our findings and a discussion of their implications.

Results of experimental studies are not, of course, on a par with those obtained by analyzing the actual behavior of economic actors. The experimental situation must, to some extent, be artificial. The subjects studied often differ systematically

[3] The experiment and the tests based on it are described in Fisher [9, Ch. 3].

[4] For an amplification of this same point in more modern language, see Hicks [14, p. 70].

from the real world actors whose behavior is of interest.[5] It is, however, difficult to judge the importance of these problems in any particular study—including this one.

2. THE EXPERIMENT

A total of twenty-three subjects participated in this experiment in January, 1971. Ten freshmen and sophomores taking the introductory economics course at the University of California, San Diego, composed the first group of subjects, while the second group, which participated one week later, consisted of twelve graduate students in economics and the spouse of one of these students. The subjects were competing for moderate (up to $10) cash prizes. It was decided to pool the two groups because their average performances, using the scoring system described below, did not differ noticeably.[6]

Each subject was given twenty-five observations on the deflated British wheat price, beginning in 1857. This series, shown in Figure 1, was taken from Fisher [9, pp. 66–7] and was the same series used in Fisher's experiments. The subjects were told that the years were not 1901–1925, but that the series gave actual wheat prices, corrected for cost of living changes, for a country with free trade in wheat and large imports, over a period with no major political changes. The subjects were given graph paper on which these observations had been plotted against time. Five year averages for years 1–5, 2–6, . . . , 21–25 were also presented and plotted.

The subjects were first required to write down their best estimate, call it F, of the average of this series for the five-year period 26–30. They were also asked to bracket their forecast by writing down a second number, call it B, such that they felt it likely that the true average for this period, $\bar{A}$, would lie between $(F - B)$ and $(F + B)$.

Prizes were awarded to those with the lowest cost, where cost had the following three components. The first was the absolute value of the difference between F and $\bar{A}$. The second component was twice B. The third component was zero unless $\bar{A}$ was below $(F - B)$ or above $(F + B)$, in which case it was ten times the amount by which $\bar{A}$ fell short of $(F - B)$ or exceeded $(F + B)$. They were told, for reasons that will become clear below, that a good rule of thumb was to pick the range B so that they felt the odds were about four to one that the actual average would lie within the indicated interval. The scoring system was discussed until all subjects said they understood it.

After the forecasts for the period 26–30 were completed, the actual price for year 26 and the five year average for the period 21–26 were announced. It was suggested that these quantities be recorded in the spaces provided on the

[5] See Friedman [10] for a discussion of these issues.

[6] Ignoring the graduate student spouse, the mean cost achieved by the second group over the twenty-five trials considered below exceeded that of the first by a tiny amount. (This implies slightly worse performance on the part of the graduate students.) The t statistic associated with this difference was only .109, however. Classifying the spouse as a graduate student, the corresponding t statistic was .361, while when she was placed with the first group, a t statistic of .201 was obtained. As her cost was only 1.36 sample standard deviations away from the overall mean, no obvious argument for excluding her from the sample was present.

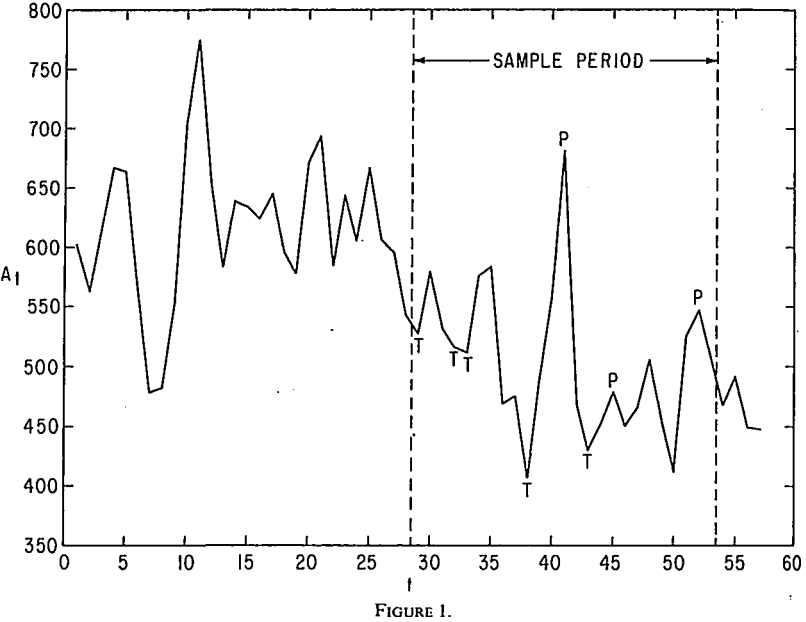

FIGURE 1.

answer sheet and that they be plotted on the graph paper. Subjects were then asked to forecast the five-year average for years 27–31 and to attach a range to their forecast as before. The experiment continued in this fashion, a year at a time, until forecasts for the period 53–57 (1909–1913) were made.

Each subject's total cost was thus based on twenty-eight forecasts and ranges. Various remarks during the experiment indicated, however, that several subjects had not really understood what was going on at the outset of the exercise. Consequently, we have chosen to consider only the last twenty-five observations for each subject, yielding a grand total of 575 observations, though the third forecasts and ranges have been employed as lagged values in various regressions reported below.

Let A_t and $\bar{A}_t$ be the actual price in year t and the actual five-year average for the period beginning with year t, respectively. A subject's forecast for the five-year period beginning in year t is denoted F_t, and the range attached to it is denoted B_t. Let C_b be the cost per unit of B_t, which equals two in our experiment, and let C_0 be the cost per unit distance that $\bar{A}_t$ lies outside the interval $[(F_t - B_t), (F_t + B_t)]$. This latter cost equals ten in our experiment.

We assume that each subject maximizes expected utility in the face of a subjective density function $g(\bar{A}_t)$, with distribution function $G(\bar{A}_t)$. The Appendix analyzes the optimal choice of F_t and B_t under these conditions; the two most important results are the following. First, if $g(\bar{A}_t)$ is symmetric about the point $\bar{A}_t = \mu$, the subject should set $F_t = \mu$, regardless of his attitude toward risk.

Second, if the subject is neutral toward risk, the following condition should be satisfied regardless of the shape of $g(\bar{A}_t)$:

$$(2.1) \qquad [1 - G(F_t + B_t)] + [G(F_t - B_t)] = C_b/C_0.$$

That is, the subjective probability that the actual average will lie outside the indicated interval should equal the ratio C_b/C_0, .20 in our case, in order to minimize expected cost.

The assumption of symmetry of the subjective distribution would not appear especially important, since F_t can be thought of as an indicator of the location of that distribution in any case. The assumption of neutrality toward risk is, however, quite crucial to the analysis. Given the relatively small prizes at stake and the small impact of any one decision on a subject's overall cost, this assumption seems sensible. Without a definite assumption about attitudes toward risk, of course, there is no way to identify the parameters of the subjective probability distribution influencing any decision. With the assumption of risk neutrality, B_t becomes a direct measure of the dispersion of $g(\bar{A}_t)$, and thus an inverse measure of the confidence attached to F_t.

In order to relate this dispersion to observable quantities, like actual forecast errors, it seemed useful to transform the B_t into estimates of the standard deviations of the subjective distributions. From standard tables, if $g(\bar{A}_t)$ is Gaussian, the subjective probability that $\bar{A}_t$ will lie outside the interval $[(F_t - B_t), (F_t + B_t)]$ is .20 when B_t is approximately 1.28 times the standard deviation of $g(\bar{A}_t)$. Thus $B_t/1.28$ may be used as an estimate of the subjective standard deviation of $\bar{A}_t$. We modified this estimate by multiplying it by $\sqrt{5}$ to obtain a quantity S_t which corresponds roughly (the correspondence is obviously exact only if the years are felt to be independent) to the subjective standard deviation associated with each of the next five years. While the interpretation of S_t as a standard deviation depends on these assumptions, it should be clear that it provides a valid measure for the dispersion of the subjective distribution regardless of the latter's shape.

Besides the series F_t, A_t, $\bar{A}_t$, and S_t, four dummy variables are employed in what follows. By means of an auxiliary experiment, Fisher [9, pp. 53–55] identified eight years in our twenty-five-year period that strongly appeared to be turning points at the time the observations corresponding to them occurred. Five of these, labeled with T's in Figure 1, were perceived as trough years, and three, labeled with P's in Figure 1, were perceived as peaks. We define DTR_t as a dummy variable equal to one in the years following these perceived troughs and zero otherwise.[7] Similarly, DPK_t is equal to one in the years following perceived peaks and zero otherwise. The remaining dummy variables are defined as follows:

$$(2.2) \qquad DTP_t = DTR_t + DPK_t,$$

$$(2.3) \qquad DTM_t = DTR_t - DPK_t.$$

[7] Thus DTR_t equals one only when $(t - 1)$ was perceived as a trough year by Fisher's subjects when they saw A_{t-1}. (This was the last value of A seen by the subjects in our experiment before they recorded F_t.) Fisher excluded two of these perceived troughs from further consideration for somewhat special reasons. Some preliminary analysis was done with two versions of DTR, the one presented in the text and one that equalled one only in the years Fisher did not exclude. Differences in results were negligible, however, and we dropped the second version.

3. FORECAST VALUES

In the twenty-five-year period of interest, the standard deviation of $\bar{A}_t$ was 27.69, and the root mean squared (RMS) error of the naive model $F_t = A_{t-1}$ was 68.15. Only fourteen of the twenty-three subjects achieved a lower RMS error than this model; the mean RMS error was 65.60, the median was 62.20, and the range was 39.52 to 97.07. On the other hand, the simple correlation between $\bar{A}_t$ and A_{t-1} was negative in this period, while the forecasts of twenty of twenty-three subjects were positively correlated with $\bar{A}_t$; the mean correlation coefficient was .3275, the median was .3514, and the range was $-.3232$ to .6422. This suggests that the subjects generally did a reasonable job of anticipating movements in $\bar{A}_t$ but that their forecasts were consistently biased; in fact, they were generally bullish.[8] Pooling all 575 observations, the mean of $B_t = (F_t - \bar{A}_t)$ was about 7.6 per cent of the mean of $\bar{A}_t$. Using a simple t test on B_t, its mean was significantly greater than zero at better than the .0001 per cent level. Similarly, the mean of B_t was positive for all subjects, significant at the one per cent level for sixteen, and significant at the five per cent level for nineteen. As Figure 1 shows, A_t, and hence $\bar{A}_t$, generally fell after year twenty-five. These subjects, by and large, underestimated the strength and permanence of this trend.

The basic hypothesis that underlies our investigation of the mechanism generating these forecasts differs both from that of Fisher [9] and from that underlying most related empirical work. Fisher suggested that, because costs of decision making are often sizeable, important alterations in expectations take place only in turning point periods. Most applied econometric work, on the other hand, assumes that some adaptive or extrapolative expectation formation mechanism operates in all periods, whether or not turning points occur. When decision making costs are modest, as they are in our experiment, it seems sensible to follow the second approach by allowing expectations to change in non-turning point periods according to some simple mechanism. On the other hand, if Fisher is right and turning point periods are special, we would expect a different mechanism to operate in such periods. In the remainder of this section, therefore, we apply alternative standard expectation formation mechanisms to all years and then investigate differences in behavior in turning point periods.

The most commonly encountered models of expectation formation are the Metzler [18]–Ferber [8] extrapolative model,

$$F_t = \alpha + A_{t-1}[\beta + \gamma(A_{t-1} - A_{t-2})/A_{t-2}] + \varepsilon_t, \tag{3.1}$$

in our notation, and the Cagan [3]–Nerlove [20, 21] adaptive model,

$$F_t - F_{t-1} = \alpha + \gamma[\beta A_{t-1} - F_{t-1}] + \varepsilon_t. \tag{3.2}$$

(Note that the value of A_{t-1} was the only new information given the subjects between their decisions on F_{t-1} and on F_t.) In order to compare these two basic

[8] This characteristic would appear to set these subjects apart from real-world business forecasters; see Hirsch and Lovell [16, pp. 62–73].

specifications, they were first rearranged so that $(F_t - A_{t-1})$ appeared on the left-hand side of both, yielding

$$(3.3) \qquad F_t - A_{t-1} = \alpha + \gamma[A_{t-1}(A_{t-1} - A_{t-2})/A_{t-2}] + (\beta - 1)A_{t-1} + \varepsilon_t$$

and

$$(3.4) \qquad F_t - A_{t-1} = \alpha + (1 - \gamma)[F_{t-1} - A_{t-1}] + \gamma(\beta - 1)A_{t-1} + \varepsilon_t.$$

To remove heteroscedasticity, all terms in these equations were divided by A_{t-1}, by F_{t-1}, and by $[A_{t-1}F_{t-1}]^{1/2}$. When these models were estimated using the data for each individual subject, the adaptive specification had a lower standard error for twenty-one out of twenty-three using the first two scaling factors and for twenty out of twenty-three subjects using the third. If the two models were really equivalent for all subjects, so that the probability was .50 that either would perform better for any particular subject, the probability of either performing better in at least twenty out of twenty-three tries would be about .0002. These comparisons led us to drop (3.1) from further consideration.

Glejser's [11] test for heteroscedasticity was applied to the three versions of the linear adaptive model (3.4) by regressing the absolute value of the residual vector on A_{t-1} and F_{t-1} for each version for each subject. At the ten per cent level, the hypothesis of no relation between these variables and the size of the residual could be rejected for only one subject when F_{t-1} was used as the deflator, while the version using $[F_{t-1}A_{t-1}]^{1/2}$ produced three rejections, and that using A_{t-1} produced five rejections. Only the last of these suggests difficulties.[9] As a further check, Durbin's [7] large sample test for first-order serial correlation in the presence of a lagged dependent variable was applied to all these regressions.[10] Again working at the ten per cent level, the null hypothesis of serial independence was rejected once for the versions using F_{t-1} and $[F_{t-1}A_{t-1}]^{1/2}$ as deflators and twice when A_{t-1} was employed. These tests indicate that all three versions of (3.4) are relatively sound. Because it seems the soundest, in what follows we work exclusively with the version in which division by F_{t-1} is used to remove heteroscedasticity.

Logarithmic versions of the extrapolative and adaptive models are

$$(3.5) \qquad \ln [F_t] = \ln [\alpha A_{t-1}] + \gamma \ln [A_{t-1}/A_{t-2}] + \varepsilon_t,$$

and

$$(3.6) \qquad \ln [F_t/F_{t-1}] = \ln [\alpha] + \gamma \ln [\beta A_{t-1}/F_{t-1}] + \varepsilon_t.$$

As above, these were rearranged so they had the same dependent variable, yielding

$$(3.7) \qquad \ln [F_t/A_{t-1}] = \ln [\alpha] + \gamma \ln [A_{t-1}/A_{t-2}] + \varepsilon_t,$$

[9] If the null hypothesis were true for all subjects and if the tests were independent so that the probability of rejecting the null hypothesis at the ten per cent level were .10 for all subjects, the probability of observing at least five rejections in twenty-three tries would be .0731, while the probability of observing at least three rejections would be .408.

[10] Durbin [7] proposes two asymptotically equivalent tests; we used the two-sided test based on his equation (11). As the small sample properties of this test are not known, its use here is justified only by the absence of a simple alternative.

and

$$\ln [F_t/A_{t-1}] = \ln [\alpha\beta^\gamma] + (1 - \gamma) \ln [F_{t-1}/A_{t-1}] + \varepsilon_t. \tag{3.8}$$

Once again the adaptive model performed much better than its rival; equation (3.8) had a smaller standard error than (3.7) for twenty-one of the twenty-three subjects, and the log-linear extrapolative model (3.5) was dropped from further consideration. The tests for heteroscedasticity and autocorrelation discussed above were also applied to equation (3.8). At the ten per cent level, the null hypothesis of homoscedasticity was rejected for four subjects, and the null hypothesis of no first-order serial correlation was rejected for one subject. Both these results support the basic soundness of the equation.[11]

In order to compare equation (3.4) (divided by F_{t-1}) and equation (3.8), the squared correlation coefficient between the predicted and actual values of F_t was computed for both specifications for all subjects. The linear model, which involved three parameters, performed better than the logarithmic model, which involved only two parameters, for only eight of the twenty-three subjects. The differences were generally tiny, however; the linear model had a slightly larger mean R^2 (.8354 versus .8341) and a slightly lower median (.8755 versus .8770). Both models perform well, considering that individuals' behavior if being modeled, and there seems very little difference between them. We consider specifications based on both in what follows.

From the very nature of the experimental situation, all subjects received approximately the same stimuli. Consequently, one might expect the disturbance terms in the various subjects' equations to be contemporaneously correlated. As an heuristic check on this possibility, Bartlett's [1] test for orthogonality was applied to the contemporaneous residual covariance matrices calculated from (3.4) (divided by F_{t-1}) and (3.8).[12] Using the normal approximation to the chi-square (with 253 degrees of freedom) yielded normal deviates of 5.68 for (3.4) and 3.80 for (3.8). Both strongly suggest the presence of contemporaneous correlation, and they thus imply that efficiency would be gained by applying Zellner's [26] generalized least squares estimator for "seemingly unrelated regressions." This method is consequently applied in what follows.

Equation 1 in Table I exhibits the coefficient means (across subjects) and the associated t statistics produced by GLS estimation of (3.4). For this specification, and for all others shown in Table I, Zellner's [26] test for aggregation bias rejected the hypothesis that the coefficients are the same for all subjects at at least the .01 per cent level. Hence, the means of the individual estimates are presented as summary information.[13] For all three coefficients, F tests rejected the hypotheses

[11] Under the assumptions of footnote 9, the probability of obtaining four or more rejections of the null hypothesis at the ten per cent level is .193.

[12] Since the small sample distribution of the estimated covariances depends in a complicated way on the correlations among the independent variables in the various equations (see, for instance, Theil [24, pp. 321–322]), this test is not exact.

[13] These statistics may be viewed as inefficient estimates of the means of the distributions of the coefficients in the population from which our subjects are a sample. See Swamy [23, pp. 147–150] for an (asymptotically) efficient estimator of such means; his approach was not employed because it would have required inversion of several 575 × 575 matrices.

TABLE I

SYSTEM GLS ESTIMATES OF LINEAR ADAPTIVE EXPECTATIONS MODELS: MEANS OF SUBJECTS' ESTIMATES[a]

Equation	Constant	DTM_t	$(F-A)_{t-1}$	$DTP_t(F-A)_{t-1}$	A_{t-1}	DTM_tA_{t-1}	DTR_tA_{t-1}	DPK_tA_{t-1}
	Independent Variables; All Divided by F_{t-1}							
1.	14.18		.6078		−.01657			
	(8.09)		(.00575)		(.0161)			
2.	5.238		.6354					
	(1.16)		(.00819)					
3.			.6412		.009990			
			(.00567)		(.00225)			
4.	5.444	.04715	.6038					
	(1.24)	(3.37)	(.00777)					
5.	6.060		.6152	.03839				
	(1.19)		(.0104)	(.0136)				
6.	6.483	−2.832	.6059	.09399				
	(1.26)	(4.49)	(.0114)	(.0513)				
7.			.6356		.01084	.001652		
			(.00571)		(.00245)	(.00610)		
8.			.6122	.1333	.01254	.009222		
			(.0104)	(.0540)	(.00246)	(.00864)		
9.			.6024	.1308	.01191		.009925	.008418
			(.00733)	(.0539)	(.00260)		(.00613)	(.00864)

[a] Quantities in parentheses are standard errors. The dependent variable is $(F_t - A_{t-1})/F_{t-1}$.

that the coefficient is zero for all subjects at at least the 1 per cent level. (This was true for all coefficients in Table I not discussed below.) The t statistics associated with the first and third mean values are small, however, because the first coefficient was negative for twelve subjects and the third was positive for eleven, while the second coefficient was positive for twenty-one subjects.

In the face of this, it seemed reasonable to simplify by assuming either $\alpha = 0$ or $\beta = 1$ in equations designed to detect turning point effects. (This makes some sense, of course, as $\alpha \neq 0$ and $\beta \neq 1$ are both ways of describing the difference between F and A in equilibrium.) Accordingly, the second and third specifications presented in Table I were estimated. The intercept in equation 2 was negative for only six subjects, while the second coefficient in equation 3 was negative for only five subjects. There seems little difference between these specifications, and we analyze both in what follows.

TABLE II

SYSTEM GLS ESTIMATES OF LOGARITHMIC ADAPTIVE EXPECTATION MODELS: MEANS OF SUBJECTS' ESTIMATES[a]

Equation	Constant	Independent Variables DTM_t	DTP_t	$\ln(F_{t-1}/A_{t-1})$	$DTP_t \ln(F_{t-1}/A_{t-1})$
1.	.01052			.6344	
	(.00232)			(.00721)	
2.	.01110	.001416		.6314	
	(.00255)	(.00720)		(.00752)	
3.	.01232		.008814	.6137	.1109
	(.00254)		(.00876)	(.0106)	(.0524)
4.	.01152			.6230	.02616
	(.00237)			(.00854)	(.0354)

[a] Quantities in parentheses are standard errors. The dependent variable is $\ln(F_t/A_{t-1})$.

System GLS estimates corresponding to (3.8) are presented as equation 1 in Table II. Again, since the hypothesis of coefficient equality across subjects was strongly rejected for all specifications shown, means of individual estimates are presented as summary information. Similarly, the hypotheses that either of the coefficients are zero for all subjects were convincingly rejected. (This was again true for all coefficients in Table II not discussed below.) The intercept was positive for all but six subjects, and the slope was positive for all but two.

We thus have three fairly standard models of expectation formation that perform well for this sample of subjects and that are almost equivalent on statistical grounds. We now examine the impact that turning points seem to have on the parameters of these models.

Consider first equation (3.4) when $\beta = 1$. After years that are perceived as troughs, one might expect forecasts to be higher than otherwise. Similarly, forecasts might be lower than usual at peaks. The most obvious way to model this is to allow α to rise immediately after perceived troughs and to fall immediately after perceived peaks. The fourth equation in Table I allows for this effect. The mean of the estimates of the second coefficient is positive but not significantly

different from zero. The hypothesis that this coefficient is zero for all subjects is quite convincingly rejected, however: the F statistic with 23 and 506 degrees of freedom is 8.07. The estimates for nine of the subjects are negative. There is thus reasonably strong evidence that α changes at turning points, but only weak evidence that the change is generally in the expected direction.

Equation 5 in Table I assumes that α is fixed, and it investigates the possibility that γ changes at turning points. When the actual series has been rising, adaptive forecasts will typically be below their equilibrium values. A fall in γ would thus lead to lower predictions at peaks and, correspondingly, higher forecasts when troughs are perceived. Again the evidence is mixed. The F statistic, with 23 and 506 degrees of freedom, corresponding to the hypothesis that the third coefficient is zero for all subjects is a highly significant 6.03, while the mean of the individual estimates, though it has the right sign, is not significant at any reasonable level. Seven of the individual estimates are negative.

Both these estimates suggest a skewed distribution of coefficients in the underlying population. Most of the mass of the densities seems to lie on the expected side of the origin, but there are noticeable tails on the other side.

Equation 6 in Table I allows for the occurrence of both changes, and it supports the second as against the first. The second coefficient, giving the mean estimated change in α, has the wrong sign, as do eleven of the individual estimates. The associated F statistic (2.76) is formally highly significant, but it is much smaller than that associated with the fourth coefficient (6.23). While eight of the estimates of this latter term are negative, the overall mean has the expected sign and is significant at five per cent on a one-tailed test.

Let us now consider the second special case of (3.4), that in which α is assumed zero. Equation 7 allows β to rise after troughs and falls after peaks, so that it corresponds closely to equation 4. And, as in that equation, the estimates are mixed. While the mean estimate of the third coefficient is virtually zero, the associated F statistic is 4.68, indicating that the hypothesis that the coefficient is zero for all subjects must be rejected. The seven subjects with negative estimated values draw down the overall mean enough to rob it of significance.

In equation 8, on the other hand, β is assumed constant while γ is allowed to fall at turning points. The second coefficient has a mean of the expected sign that is significant at the one per cent level on a one-tailed test. Eight of the individual estimates are negative, and the corresponding F statistic is 6.58. If β is in fact constant, the fourth coefficient should be a simple (but nonlinear) function of the first three for each subject.[14] These constraints were not imposed, however, and the mean of the estimates of the fourth coefficient has the wrong sign, as do thirteen of the individual estimates; the corresponding F statistic is only 2.78.

One possible reason for the weakness of this coefficient is that β also changes at turning points, rising at troughs and falling at peaks. Equation 9 allows for this

[14] Let b, c, d, and e be the four coefficients for some subject, and let γ' be the value of γ in turning-point periods. Then b is an estimate of $(1-\gamma)$, c of $(\gamma-\gamma')$, d of $\gamma(\beta-1)$, and e of $(\gamma'-\gamma)(\beta-1)$, so that four coefficients are being used to estimate three parameters. The obvious constraint is $e = -cd/(1-b)$.

possibility. Letting γ' be the value of γ in turning point years, β^p be the value of β in peak years, and β^t be the value of β in trough years, the fourth coefficient in equation 9 is an estimate of $[\gamma'(\beta^t - 1) - \gamma(\beta - 1)]$, while the fifth coefficient estimates $[\gamma'(\beta^p - 1) - \gamma(\beta - 1)]$. The mean estimate of the fourth coefficient exceeds that of the fifth, as we would expect, though the t statistic corresponding to this difference is only .152. In a now familiar pattern, however, the fifth coefficient exceeded the fourth for eight subjects, and an F test rejected the hypothesis that these coefficients are the same for all subjects at better than the one per cent level. On the other hand, the estimates of the second coefficient in equation 9 have all the solidity of those in equation 8. In this specification, as in the alternative assuming $\beta = 1$, it seems clear that γ falls in turning point periods, while the evidence for a change in the model's other parameter is less convincing.

Finally, we come to the log linear model, (3.8). Here α and β cannot be identified without further information, but we can allow α and/or β to rise in trough years and fall in peak years, thus causing corresponding changes in $\ln(\alpha\beta^\gamma)$. Equation 2 in Table II looks for such changes on the assumption that γ is constant, and it does not find them. The second coefficient is negative for eleven subjects, and its mean is not significant.[15] In equation 3, on the other hand, α and β are assumed constant, and γ is allowed to fall at turning points. Eight of the individual estimates of the fourth coefficient have the wrong sign, but the mean has the expected sign and is significant at the 2.5 per cent level on a one-tailed test, and the overall F statistic (7.89) is highly significant. We would expect the second coefficient in this equation to be negative, since we would expect β to generally exceed one, but the overall mean and fourteen of the individual estimates are positive. The mean is not significantly different from zero, and the F statistic relating to the hypothesis that the coefficient is zero for all units (2.77), while significant, is much smaller than that relating to the fourth coefficient.

There are two possible reasons for the weakness of the second coefficient. First, in spite of the dismal performance of equation 2, α and β may also be changing at turning points. Second, if β is quite close to unity, changes in γ will have little effect on the intercept in (3.8). We attempted to estimate an equation like 3 but with DTP_t replaced by DTR_t and DPK_t in order to investigate the first hypothesis. Unfortunately, the routine employed was unable to perform one of the matrix inversions required to compute the system estimates. Some evidence is provided by the ordinary least squares estimates, however. The coefficient of DTR_t exceeded that of DPK_t for seventeen of the twenty-three subjects, implying, as expected, that $\ln(\alpha\beta^\gamma)$ is generally larger in trough periods than in peak periods. Viewing each of the twenty-three estimates as an independent trial, if the true difference were zero for all so that the probability of the coefficient of DTR_t exceeding that of DPK_t were .5, the probability of obtaining seventeen or more "successes" in twenty-three trials would be .017.[16] Some indirect evidence is also

[15] As usual, however, the hypothesis that this coefficient is zero for all subjects is rejected at the one per cent level.

[16] Note that this sort of reasoning cannot be applied to the system estimates, since cross-equation dependence is explicitly present there.

provided by equation 4 in Table II, which embodies the alternate hypothesis that, because β is generally near unity, the intercept does not change at all at turning points. Imposing this restriction causes the mean of the third coefficient to fall dramatically from significance, even though only five of the individual estimates are negative.

Still, as in the other specifications, it seems clear that γ, the speed of adjustment, generally falls in turning point periods. The other parameters may also change, but the evidence from this sample is far from conclusive on that point.[17]

4. FORECAST CONFIDENCE

Just as the subjects' forecasts were generally biased upward, so they seemed to have been excessively confident of their projections. As we indicated in Section 2, if the subjects were minimizing expected cost, the subjective probability of the actual average, $\bar{A}$, falling outside the interval $[(F - B), (F + B)]$ would be .20 for all observations. Yet for 321 of the 575 observations, $\bar{A}$ did fall outside this interval. The probability of obtaining at least this number of observations with $\bar{A}$ outside the specified interval when the probability of such an occurrence is .20 is, using the normal approximation to the binomial, less than 10^{-9}. Under this same null hypothesis, the probability that any subject would encounter nine or more such events in twenty-five periods is about .05, yet eighteen of the twenty-three subjects had nine or more $\bar{A}$'s outside the corresponding intervals.

Three explanations for this bias suggest themselves. First, it is quite possible that our subjects did not understand the laws of probability or their application to the experimental situation. While this might lead to erratic behavior, however, it is hard to see why it would lead to a persistent bias in the observed direction. Second, as the Appendix shows, we might expect this bias if the subjects were generally risk lovers. Finally, as inexperienced forecasters, our subjects may have continuously overestimated their own abilities. The third of these seems as reasonable as the second, and even if the subjects generally preferred risk for some reason, there still may be something to be gained from analysis of changes in their level of confidence.

A second important characteristic of the S_t series is the extreme inertia it shows. For 247 of the 575 observations considered, about 43 per cent, S_t equalled S_{t-1}. That is, almost half the forecasts were written down with the same associated range as the previous forecast.[18] As there were no objective costs of changing B (and thus S_t), the only obvious explanation of this inertia is that the scoring system, as it related to the range attached to any forecast, was so complex that subjects were reluctant to make the mental effort of coping with it.

Because of this property of the data, we cannot treat all 575 observations on S_t as equivalent. It seems reasonable to suppose that when a change was made in

[17] It should be reported that a number of attempts were made to detect the effects of forecast confidence (measured by (S_t/A_{t-1})) on the parameters of these models, but no evidence of the existence of such influences was obtained.

[18] The median number of changes in S_t per subject in twenty-five trials was fourteen, and the mean was 14.3. The range was from four to twenty-four changes.

S_t, the new value accurately reflected perceived uncertainty. On the other hand, if $S_t = S_{t-1}$ it would appear doubtful that S_t was the best possible estimate of the relevant parameter of the subjective probability distribution. More likely, it was a bad estimate, but one that was not so bad that the subject felt it worth the effort to correct it.

We thus divide our analysis of forecast confidence into two parts.[19] The first considers models explaining S_t and applies them to the 328 "good" observations on this quantity. The second part focuses on the timing of changes in S_t. As will become clear, this section does not purport to present and verify "the" definitive theory of forecast confidence determination. Some seemingly plausible hypotheses have been confronted with the data, but conclusive results have not been obtained. As this is a new area of investigation, I could probably have continued to generate hypotheses, with an eye to the data, until more impressive statistics fell from the computer. I think, however, it is more useful to report the evidence obtained on a first set of hypotheses.

Our basic hypothesis is that the uncertainty a subject attaches to expectations about the future is related to the quality of his past forecasts. Specifically, we would expect a run of large forecast errors to lead a subject to doubt his ability and thus to increase the range attached to future forecasts. To make this notion more concrete, we must define "forecast error" and specify the functional form of the hypothesized relation.

Three definitions of forecast error were examined. Letting $E^{\alpha}_{t,\tau}$ be the error of type α associated with $F_{t-\tau}$ at the time the subject is preparing to write down F_t and S_t, the simplest definition is

$$E^{a}_{t,\tau} = |F_{t-\tau} - A_{t-\tau}|. \tag{4.1}$$

This definition implicitly assumes that each subject's feeling about any forecast is determined as soon as he hears the first of the five actual values to which it refers. It does not allow for any tendency to keep track of the five-year averages, which in fact determine his score. The second definition does allow for this:

$$E^{b}_{t,\tau} = \begin{cases} \left|F_{t-\tau} - \sum_{k=1}^{\tau} A_{t-k}/\tau\right|, & \tau < 5, \\ \left|F_{t-\tau} - \sum_{k=\tau-4}^{\tau} A_{t-k}/5\right|, & \tau \geqslant 5. \end{cases} \tag{4.2}$$

According to this definition, forecasts written down five or more periods earlier are compared with the corresponding actual five-year averages to judge their accuracy. For more recent forecasts, an estimate of the corresponding average is made using only those actual values currently available. Note that when $\tau = 1$ these two definitions give the same measure. Finally, it might be supposed that subjects also use their current expectations about the future to evaluate recent

[19] The general model of Dagenais [6] offers a conceptually more satisfactory approach to this sort of situation. Computationally, however, it is extremely burdensome, and, for that reason, it was not employed.

forecasts. As these expectations are reflected in the current forecast, we are led to our third definition:

$$E^c_{t,\tau} = \begin{cases} F_{t-\tau} - \left[(5-\tau)F_t + \sum_{k=1}^{\tau} A_{t-k}\right]/5, & \tau < 5, \\ E^b_{t,\tau}, & \tau \geqslant 5. \end{cases} \tag{4.3}$$

Suppose for some particular subject and some particular period t that $S_t \neq S_{t-1} = S_{t-2} = \ldots = S_{t-k} \neq S_{t-k-1}$. It seems logical that $S_t - S_{t-1}$ should reflect all the relevant evidence on the subject's forecasting ability that has been accumulated in periods $t-1, t-2, \ldots, t-k$. In particular, we assume that the quantity

$$D(\alpha, \lambda)_t = \left[\sum_{i=1}^{k} \lambda^{i-1}(E^{\alpha}_{t,i})^2 \Big/ \sum_{i=1}^{k} \lambda^{i-1}\right]^{1/2}; \quad 0 \leqslant \lambda \leqslant 1; \quad \alpha = a, b, c, \tag{4.4}$$

summarizes this information. $D(\alpha, \lambda)_t$ is the maximum likelihood estimate of the standard deviation of $E^{\alpha}_{t,\tau}$ in periods $t-1$ to $t-k$ when $\lambda = 1$ and E^{α} is assumed normally distributed with known mean zero. For smaller values of λ, more weight is placed on more recent periods, and for $\lambda = 0$, $D_t = E^{\alpha}_{t,1}$.

We initially assumed, in the spirit of adaptive models of expectation formation, that our subjects sought to establish some desired relation between $D(\alpha, \lambda)_t$ and S_{t-1}. When $S_t \neq S_{t-1}$, we assumed that the new uncertainty measure was generated according to

$$S_t - S_{t-1} = \gamma\{[\alpha + \beta D(\alpha, \lambda)_t] - S_{t-1}\} + \varepsilon_t \tag{4.5}$$

or equivalently,

$$S_t = \gamma\alpha + \gamma\beta D(\alpha, \lambda)_t + (1-\gamma)S_{t-1} + \varepsilon_t, \tag{4.6}$$

where ε_t was assumed homoscedastic and serially independent. Fifteen estimates of (4.6) were computed for each subject, assuming $\alpha = a, b, c$ and $\lambda = 1.0, .75, .50, .25, 0.0$.

Selection among the (α, λ) pairs was made on the basis of sums of squared residuals. Table III shows the frequency with which each pair of values was thus

TABLE III

DISTRIBUTION OF NONLINEAR PARAMETER ESTIMATES[a]

	λ					
Error Type (α)	1.0	.75	.50	.25	0.0	Total
a	1	0	1	1	1.5[b]	4.5
b	0	0	2	1	1.5[b]	4.5
c	7	1	0	1	5	14
Total	8	1	3	3	8	23

[a] Number of subjects for whom the indicated (α, λ) pair minimized the sum of squared residuals in (4.6).
[b] Since $(a, 0.0)$ and $(b, 0.0)$ are identical, the three subjects for whom this was the best specification were simply split between the two error types.

chosen. Notice that $\alpha = c$ was best for the majority of subjects; if the true probability of its being best for any one subject were $\frac{1}{3}$, the probability of its being selected fourteen or more times out of twenty-three trials would be about .007. Considering all values of λ, the sums of squared residuals with error types a and b could be compared ninety-two times, and type b was superior in fifty-two cases. Under the obvious null hypothesis, the probability of fifty-two or more "successes" is about .106, indicating that $\alpha = b$ was generally a better specification than $\alpha = a$. In 115 $a - c$ comparisons, c had the lower sum of squared residuals seventy-three times, while it was superior in seventy-one $b - c$ comparisons. Under the obvious null hypothesis, the probability of seventy-one or more successes is only about .006; this reinforces the impression of type c's superiority.

This is in fact what we had expected. The specification $\alpha = c$ assumes that subjects use all available information in evaluating past forecasts. When $\alpha = a$, they are assumed to use very little, while $\alpha = b$ is intermediate in this respect. It is comforting that the subjects generally seemed to use all the information at hand.

For the twenty-two subjects for which they could be calculated, F statistics were computed treating $(c, 1.0)$ and $(c, 0.0)$ as null hypotheses and formally assuming that one restriction was relaxed in estimating any alternative specification. These computations strongly indicated the flatness of the likelihood functions for most subjects. In the fifteen cases where $(c, 1.0)$ was not the preferred specification, it was rejected only twice at the ten per cent level, both times in favor of $(c, 0.0)$. In the seventeen cases where $(c, 0.0)$ was not preferred, it was rejected only three times at this same level, always in favor of $(c, 1.0)$. While these are hardly rigorous tests, they do add to the evidence suggesting that we would lose little by restricting our attention to $(c, 1.0)$ and $(c, 0.0)$. In simple pairwise comparisons, the first of these outperformed the second for eleven subjects, so it is not obvious how one would choose between them.

Several other aspects of these estimates of (4.6) require comments. The coefficient of D_t was positive, as expected, for eighteen subjects, and it was significant at the ten per cent level on a one-tailed test for fourteen.[20] Under the obvious null hypotheses, the probabilities of obtaining at least that many successes are .0053 and $<.0001$, respectively. The basic hypothesis that past performance influences confidence thus receives strong support. Further, this specification appeared free of heteroscedasticity; Glejser's [11] test, applied as in the last section, rejected the null hypothesis at the ten per cent level for only three subjects.

There were problems with this specification, however. Durbin's [7] large sample test for serial correlation was applied to the preferred regressions for each subject. The estimated variance of the estimated coefficient of the lagged dependent variable was too large to permit computation of the test statistic for five subjects, but for ten of the remaining eighteen, the null hypothesis of no first-order serial correlation was rejected at the twenty-five per cent level on a two-tailed test. For seven subjects, this hypothesis was rejected at the five per cent level. Even though the small sample properties of this test are unknown, these results required allowance for the possibility of serial correlation.

A second problem was the relative weakness of the coefficient of the lagged

[20] Degrees of freedom were not adjusted to take into account the estimation of α and λ.

dependent variable.[21] This coefficient was negative for ten subjects, though it was positive and significant at the ten per cent level on a one-tailed test for seven. While one might suspect that the latter result merely reflected bias in the presence of serial correlation, this coefficient was significant at the ten per cent level for only one of the ten subjects for whom the null hypothesis of serial independence was rejected at the twenty-five per cent level. Still, it seemed clear that we must allow for the possibilities that $\gamma = 1$ and that serial correlation is present for some subjects.[22]

Consequently, a second round of estimates was calculated. To ensure that hypothesis tests would have some power, we dropped the five subjects who made fewer than ten changes in S_t. These subjects accounted for six of the seven of our 328 total changes for which k, as in (4.4), was ten or greater, but they contributed only five of the thirty-five changes for which $3 \leqslant k \leqslant 7$. Three of these subjects had estimated values of λ not equal to zero or one; it seems difficult to interpret their disproportionate representation in this category. Finally, only one of the five excluded subjects had a negative coefficient of D_t.

The following equations were estimated for each subject with $\lambda = 1.0$ and with $\lambda = 0.0$:

$$S_t = \alpha + \beta D(c, \lambda)_t + \varepsilon_t, \tag{4.7a}$$

$$S_t = \alpha\gamma + \beta\gamma D(c, \lambda)_t + (1 - \gamma)S_{t-1} + \varepsilon_t, \tag{4.7b}$$

$$S_t = \alpha + \beta D(c, \lambda)_t + u_t, \tag{4.7c}$$

$$S_t = \alpha\gamma + \beta\gamma D(c, \lambda)_t + (1 - \gamma)S_{t-1} + u_t, \tag{4.7d}$$

where ε_t is assumed homoscedastic and serially uncorrelated, while u_t is assumed to be generated by a first-order autoregressive process with serial correlation coefficient ρ. This parameter was estimated by the Cochrane–Orcutt [4] iterative technique. As this approach drops the first observation, that observation was also dropped in estimation of (4.7a) and (4.7b) to ensure comparability.[23]

For each value of λ, simple F tests were used to select the preferred specification.[24]

[21] Of course, the well known small sample bias in this quantity, even in the absence of serial correlation, makes this finding less than completely surprising.

[22] One might suspect that large values of γ and small values of λ would go together, as both indicate heavy weighting of the most recent information. No such association was found in this sample, however. There was a tendency for subjects making more changes in S_t to have smaller estimated γ's; the simple correlation coefficient was significant at the ten per cent level. No significant correlations involving λ were observed in this sample.

[23] Estimates of (4.7a) with the addition of DPT_t were also computed to see if turning point effects could be detected here. For $\lambda = 1.0$, ten of the eighteen coefficients of this variable were negative, and four of these were significant at ten per cent on a two-tailed test. (As there were no a priori reasons to expect this coefficient to be of either sign, a two-tailed test is appropriate.) When $\lambda = 0.0$, a negative sign was encountered in twelve trials, and three of these were significant by the same test. These results suggested that subjects may have been more confident just after turning points, but the evidence seemed too weak to pursue the point.

[24] In the first round of tests, (4.7a) was treated as the restricted model, and each of the other three were used as the unrestricted model. If none of the restrictions could be rejected at the ten per cent level, the simple model was selected, while if only one of the other models was superior on this basis, it was chosen. If two or more of the tests were significant, another round of similar tests was run. Where models (4.7b) and (4.7c) were equivalent under this procedure, the one with the smaller standard error (deducting one degree of freedom in (4.7c) to allow for the estimation of ρ) was chosen.

Following Bischoff [2], the restriction $\rho = 0$ was treated as if it were linear. Somewhat surprisingly, (4.7a) was selected for eight of the eighteen subjects with $\lambda = 0$ and for 10 when $\lambda = 1$. Even more surprisingly, in view of the results of Durbin's test, (4.7d) represented a significant improvement over (4.7b) only once for each value of λ. The preferred specifications for the two values of λ were compared, and the one with the smaller standard error was chosen.[25] (In most cases, differences were minor, especially between (4.7b) and (4.7c).) The results are shown in Table IV; the statistics shown for specifications (4.7a) and (4.7b) were computed using all observations for each subject.

These are not terribly impressive. For subjects eight, ten, fourteen, and sixteen, the coefficient of D_t was negative not only here but in virtually all other specifications estimated. The basic model employed here does not seem to apply to those subjects. Some of the estimates of this coefficient for subject seventeen were positive, including that in the preferred version of (4.6), which had $\alpha = b$ and $\lambda = .50$, but it is quite possible, in view of the difficulty of drawing any firm conclusions on the basis of ten observations, that the model does not apply to him either. On balance, though, I think the evidence for the hypothesis that past performance generally influences current confidence suggests that it merits further study.

Considering the thirteen subjects with ten or more changes to whom the model does seem to apply, it is interesting to note that equation (4.7a) was selected for five of the six subjects for whom $\lambda = 0.0$ was the preferred specification. When $\lambda = 0.0$, only the quality of the most recent forecast is used to compute D_t, so that it is not surprising that older information, embodied in the last choice of S_t, does not affect current confidence. Similarly, this equation was selected for none of the seven subjects for whom $\lambda = 1.0$ was preferred. When $\lambda = 1.0$, all the information received since the last choice of S_t is reflected in D_t, so that it is not too surprising that even older information affects confidence. This suggests a natural division of this sample into those whose confidence is determined entirely by very recent experience and those whose confidence is affected by their entire past performance. It is not clear, however, how such a division, if it really exists, might be exploited.

Finally, we report on an investigation of the causes of changes in S_t. Our basic hypothesis was that differences between S_{t-1} and D_t that are significant, in a sense to be made precise below, are likely to lead to changes in S_t.[26] This hypothesis did

[25] One degree of freedom was subtracted in (4.7c) and (4.7d) to allow for the estimation of ρ. This selection procedure was not followed for subject thirteen in Table IV. For this subject, (4.7c) was only slightly superior to (4.7b) in the reduced sample, and $\lambda = 0.0$, which had negative coefficients of D_t, was slightly superior for both models to $\lambda = 1.0$, for which the coefficients were positive. In the overall sample, however, (4.7b) with $\lambda = 1.0$ was considerably better than the same equation with $\lambda = 0.0$. In the face of this evidence, (4.7b) with $\lambda = 1.0$ was selected as the preferred specification.

[26] A preliminary analysis of the impact of turning points on the occurrence of changes in S_t was also carried out. Of the 328 changes, 64 occurred in the five trough years. Under the null hypothesis that the probability of a change is independent of whether or not a trough is perceived, one would expect 65.6 changes in these years; the difference is clearly not significant. Similarly, 40 changes occurred in peak years, 39.4 would be expected under the null hypothesis, and the difference is negligible. These findings were sufficiently negative to discourage further search for turning point effects on this aspect of subject behavior.

TABLE IV

FORECAST CONFIDENCE MODELS FOR SUBJECTS WITH TEN OR MORE OBSERVATIONS[a]

Subject	Total Observations[b]	Constant	$D(c, \lambda)_t$	S_{t-1}	λ	ρ	R^2
1	18	6.443	.2326	.7116	1.0	—	.74
		(15.0)	(.0559)	(.115)			
2	13	22.55	.0539	—	1.0	−.65	.41
		(2.41)	(.0252)				
3	17	24.52	.1084	—	0.0	—	.19
		(4.93)	(.0583)				
4	21	33.00	.1591	—	0.0	—	.15
		(3.22)	(.0853)				
5	11	25.78	.1470	—	0.0	—	.07
		(5.87)	(.173)				
6	16	36.72	.2335	.4546	1.0	—	.73
		(14.0)	(.0553)	(.156)			
7	12	24.80	.3082	—	0.0	—	.23
		(4.82)	(.176)				
8[c]	16	29.12	−.0572	−1.076	0.0	.64	.72
		(2.90)	(.0297)	(.124)			
9	23	100.0	.3833	—	1.0	.35	.18
		(12.3)	(.282)				
10	19	97.20	−.2326	—	1.0	—	.07
		(13.1)	(.203)				
11	20	60.40	.0543	−.0639	1.0	—	.05
		(13.5)	(.0571)	(.243)			
12	14	67.20	.6442	—	0.0	—	.28
		(13.6)	(.300)				
13	14	132.7	.1484	−.1025	1.0	—	.29
		(23.6)	(.0694)	(.186)			
14	14	59.80	−.0499	—	1.0	—	.18
		(4.96)	(.0308)				
15	22	20.33	.2055	.5484	0.0	—	.33
		(11.6)	(.240)	(.181)			
16	24	76.15	−.2777	—	0.0	.46	.23
		(9.31)	(.210)				
17	10	124.8	−.2359	—	0.0	—	.02
		(25.0)	(.635)				
18	13	18.40	.1717	.6487	1.0	—	.69
		(12.9)	(.0780)	(.155)			

[a] Standard errors are shown in parentheses.
[b] Where $\rho \neq 0$, estimates are based on one less observation.
[c] Cooper's [5] large sample adjustment was applied to the standard errors.

not receive a great deal of support, but, since the problem investigated is something of a peculiarity of this sample, no effort was made to improve upon it.

Suppose, as before, that $S_{t-1} = \ldots = S_{t-k} \neq S_{t-k-1}$ and consider the quantity

$$C_t^n = [M_t D(\alpha, \lambda)_t / \bar{S}^n]^2, \quad \text{where} \quad M_t = \sum_{i=1}^{k} \lambda^{i-1}. \tag{4.8}$$

Under the null hypothesis that the $E_{t,\cdot}^z$ are normally and independently distributed with known mean zero and standard deviation $\bar{S}^n$, C_t^n is distributed as chi-square

with $M_t = k$ degrees of freedom when $\lambda = 1.0$. Under the same null hypothesis, C_t^n is distributed as chi-square with $M_t = 1$ degree of freedom when $\lambda = 0.0$.[27] If the subject would set $S_t = S_{t-1}$ if the null hypothesis were true, one would expect values of C_t^n large or small enough to reject this hypothesis at some reasonable level to lead to $S_t \neq S_{t-1}$.

This approach was applied to the thirteen subjects whose estimates in Table IV had positive coefficients of D_t: the values of λ indicated there were used. Two versions of $\bar{S}^n$ were employed. The simplest, call it $\bar{S}^1$, was simply S_{t-1}. The second version, call it $\bar{S}^2$, was obtained by solving each subject's estimated equation in Table IV for D_t as a function of S_{t-1}, assuming $S_t = S_{t-1}$. This quantity is thus the value of D_t for which the subject would be in equilibrium if the current S_{t-1} prevailed.

Linear interpolation was then used with standard tables to approximate[28]

$$P_t^n(C_t^n) = 2|\text{prob}\,(x \leqslant C_t^n) - 1/2|, \tag{4.9}$$

for x distributed as chi-square with M_t degrees of freedom and $n = 1, 2$, for each year in the twenty-five-year sample period for each of the thirteen subjects considered. It is clear that the P_t^n vary between zero and one and that they equal $(1 - 2\alpha)$, where α is the level of significance attained by the corresponding C_t^n on a one-tailed test. Values of P_t^n near one should, on the above reasoning, be likely to lead to changes in the range attached to a subject's forecast.

Since P_t^n is bounded, a linear probability model is not as inappropriate here as in most applications. Hence, we created a variable Δ_t for each subject that was equal to one when $S_t \neq S_{t-1}$ and zero otherwise, and we estimated equations in which this quantity was a linear function of P_t^n.[29] In ordinary least squares runs with P_t^1, only eight of the thirteen slope coefficients were positive, and the mean and median values of R^2 were .0618 and .0155. When P_t^2, which had been expected to perform better, was the independent variable, only four of the slopes were positive, and the mean and median values of R^2 were .0521 and .0334. It is true, as Morrison [19] has pointed out, that if the true probabilities of change do not vary much, one would expect low values of R^2 in estimates of this sort.[30] But we obtained truly miniscule values of this statistic, and the sign pattern of the slope coefficients is quite disturbing.

Ordinary least squares is well known to be an inefficient estimator of the linear probability model because of that model's basic heteroscedasticity. We thus reestimated these functions, using the fitted values from the least squares equations

[27] For intermediate value of λ, the moment generating function of this quantity is reasonably complicated. It is relatively easy to verify, however, that if $F(\lambda, k) \equiv (1 + \lambda^k)/(1 + \lambda)$, the random variable $Z_t = M_t[1 - F(\lambda, k)] + C_t/\sqrt{F(\lambda, k)}$ has the same mean and variance as a random variable distributed according to the gamma distribution with parameters $M_t/2$ and 2. As this reduces to the chi distribution when M_t is an integer, appropriate interpolation in standard tables of the chi-square could be employed as a first approximation.

[28] Interpolation was linear in C for C less than the median of the appropriate distribution, recognizing that $P(0) = 1$. Similarly, linear interpolation was performed on $1/C$ for C above the median, taking into account that $P(\infty) = 1$.

[29] Some experiments were also performed with quadratics, but no gain in predictive power was obtained.

[30] But see also Goldberger [12].

to estimate the standard deviations of the individual disturbances.[31] The sign pattern of the slopes was completely unchanged in this second round.

Two ways of summarizing these estimates seemed to be available. First, we employed Zellner's [26] generalized least squares estimator to estimate each of these (transformed) models for all subjects together. As in the last section, the similarity of the stimuli to which all were exposed suggested substantial contemporaneous error correlation. For both models, seven of the thirteen slopes were positive in this third round of estimates, and the hypothesis of identical coefficients for all subjects could be rejected at better than the one per cent level. The arithmetic means of the coefficients across subjects and the associated standard errors were as follows:

$$\underset{(.0484)\quad(.0673)}{\Delta_t = .5624 + .1674\ P_t^1},\tag{4.10a}$$

and

$$\underset{(.0436)\quad(.0608)}{\Delta_t = .6665 + .0152\ P_t^2}.\tag{4.10b}$$

These suggest that the probability of changing the range attached to any forecast is, at best, only slightly raised by even dramatic increases in the significance level attained by C_t. According to (4.10a), the probability of change varies from .562 to .730, while (4.10b) indicates a range of .667 to .682, while the actual mean of the dependent variable for all thirteen subjects was .658. These statistics might lead one to believe that the true probabilities do not vary much.

Bartlett's [1] test was applied to the estimated contemporaneous disturbance covariance matrices used in computing these estimates, however, and in both cases the normal approximation to the chi-square statistic with seventy-eight degrees of freedom was negative, implying that the hypothesis of disturbance orthogonality could not be rejected at any reasonable level. The small sample properties of this test applied to regression residuals are unknown, of course, and the GLS standard errors were generally noticeably smaller than the corresponding single equation statistics; the standard errors of the slopes fell by an average of about thirty per cent in both models. Still the possibility of disturbance orthogonality suggested the use of Swamy's [23, Ch. 4] random coefficient estimator for such situations. This approach assumes that the true coefficients for each subject are random drawings from an underlying population, and asymptotically efficient estimates of the mean vector of that population are produced. The estimated population means and the associated standard errors for the two models considered were as follows:

$$\underset{(.141)\quad(.206)}{\Delta_t = .6495 + .0175\ P_t^1},\tag{4.11a}$$

[31] See McGillivray [17] for the basic procedure and its large sample properties. We followed a suggestion of Smith and Cicchetti [22] and treated observations for which the least squares equation yielded an estimated probability outside the interval [0, 1] as if the estimate had been .99, thus assigning such observations large weights.

and

$$\Delta_t = \underset{(.175)}{.6547} - \underset{(.222)}{.0141}\ P_t^2. \tag{4.11b}$$

These are quite awful.

There seems to be no point in trying to argue for the superiority of (4.11) over (4.10) as a general representation of subject behavior, or vice versa. Equations (4.10) take into account disturbance covariances, but the arithmetic mean of the individual subjects' estimates is not a particularly efficient estimator of the population mean. Similarly, (4.11) result from a technique that is efficient only if the a priori unlikely hypothesis of disturbance orthogonality is correct. Overall, though, the bad results in (4.11), the rather disappointing sign pattern of the single-equation slope coefficients, and the unexpected superiority of P_t^1 over P_t^2 combine to suggest weakness in the basic approach. We have been assuming that the probability that S_t differs from S_{t-1} is a function only of the extent to which the latter is seen to be at odds with recent performance. As the results indicate, this assumption may be too restrictive. One might imagine, for instance, that S_t received thought when a subject found it easy, for one reason or another, to decide on F_t. Still, our main interest is in the determination of the level of S_t, rather than the timing of its changes.

A final set of experiments should be mentioned. Since the confidence with which expectations are held is so rarely directly observable, it seemed of some interest to see if an observable proxy for this quantity could be found. The only thing that came to mind, however, was some measure of the dispersion of the individual forecasts. A variety of measures of average confidence were related to a variety of measures of the dispersion in the F_t. While some significant correlations were obtained, none of the relations had sufficient explanatory power to justify reporting results in detail.

5. CONCLUSIONS AND IMPLICATIONS

In Section 3 it was found that turning point years seemed to be special years to the subjects in this experiment, as Fisher [9] had asserted, even though costs of decision making were negligible and account was taken of the operation of an adaptive expectations mechanism. In the three models examined, the best supported hypothesis was that the parameter γ, the speed of response in the adaptive structure, fell in turning point periods.

This finding has a number of implications. First, it suggests that the technical expectation formation mechanisms commonly assumed in empirical work are too simple. But the alternative mechanism suggested by this analysis is not much more complex, so it should be relatively simple to test its applicability to real-world data. The limitations of the experimental approach and the somewhat tentative nature of our statistical results suggest that further testing of this notion, using any or all of the three approaches mentioned in the Introduction, is in order.

In Section 4, we investigated the determinants of the confidence with which expectations were held. Some support was found for the basic hypothesis that a

forecaster's current confidence was affected by his past performance, and weaker evidence was found for the notion that our subjects were more likely to change their reported confidence when previous reports were strikingly at odds with their recent performance. This part of the study was clearly more tentative and exploratory than the analysis of Section 3, and it yielded fewer firm conclusions. But it did suggest, I hope, the possibility of refining this basic approach and applying it in other contexts. At a minimum, this study should indicate the need for and possibility of explicit consideration of at least two moments of economic actors' subjective probability distributions in applied work.

University of California, San Diego

Manuscript received September, 1973.

APPENDIX

If a subject makes decisions according to a utility of points function $U(x)$, where x is minus his cost in points, the expected utility associated with a particular choice of F and B is (dropping time subscripts for clarity)

$$E(U) = \int_{-\infty}^{F-B} U\{-(F-\bar{A}) - BC_b - C_0[(F-B)-\bar{A}]\}\,dG(\bar{A}) + \int_{F-B}^{F} U\{-(F-\bar{A}) - BC_b\}\,dG(\bar{A}) + \int_{F}^{F+B} U\{-(\bar{A}-F) - BC_b\}\,dG(\bar{A}) + \int_{F+B}^{+\infty} U\{-(\bar{A}-F) - BC_b - C_0[\bar{A}-(F+B)]\}\,dG(\bar{A}). \tag{A.1}$$

Making the necessary convergence and regularity assumptions (see Hildebrand [15, pp. 359–361]), we can differentiate $E(U)$ to obtain the two first-order conditions for a maximum:

$$\frac{\partial E(U)}{\partial F} = \left\{\int_{F}^{F+B} U'\{-(\bar{A}-F) - BC_b\}\,dG(\bar{A}) - \int_{F-B}^{F} U'\{-(F-\bar{A}) - BC_b\}\,dG(\bar{A})\right\} + (1+C_0)\left\{\int_{F+B}^{+\infty} U'\{-(\bar{A}-F) - BC_b - C_0[\bar{A}-(F+B)]\}\,dG(\bar{A}) - \int_{-\infty}^{F-B} U'\{-(F-\bar{A}) - BC_b - C_0[(F-B)-\bar{A}]\}\,dG(\bar{A})\right\} = 0, \tag{A.2}$$

and

$$\frac{\partial E(U)}{\partial B} = -C_b\left\{\int_{F-B}^{F} U'\{-(F-\bar{A}) - BC_b\}\,dG(\bar{A}) + \int_{F}^{F+B} U'\{-(\bar{A}-F) - BC_b\}\,dG(\bar{A})\right\} + (C_0 - C_b)\left\{\int_{-\infty}^{F-B} U'\{-(F-\bar{A}) - BC_b - C_0[(F-B)-\bar{A}]\}\,dG(\bar{A}) + \int_{F+B}^{+\infty} U'\{-(\bar{A}-F) - BC_b - C_0[\bar{A}-(F+B)]\}\,dG(\bar{A})\right\} = 0, \tag{A.3}$$

where primes denote differentiation. If the subject is neutral toward risk, these reduce to

$$\{G(F+B) + G(F-B) - 2G(F)\} + (1+C_0)\{1 - G(F+B) - G(F-B)\} = 0, \tag{A.2'}$$

and

$$-C_b\{G(F+B) - G(F-B)\} + (C_0 - C_b)\{G(F-B) + 1 - G(F+B)\} = 0. \tag{A.3'}$$

Two observations can be made at this point. First, if $g(\bar{A})$ is symmetric about the point $\bar{A} = \mu$, $F = \mu$ and any $B \geq 0$ will satisfy (A.2). In this case the two choices are sequential: (A.2) determines F, and (A.3) determines B. Second, equation (A.3') may be rewritten as

$$[1 - G(F + B)] + [G(F - B)] = C_b/C_0. \tag{A.4}$$

That is, risk neutrality implies that the subjective probability of $\bar{A}$ falling outside the range $[(F - B), (F + B)]$ should equal C_b/C_0, regardless of the shape of the distribution.

In order to derive a manageable set of second-order conditions, let us assume that $g(\bar{A})$ is symmetric about $\bar{A} = \mu$, so that $\mu = F$. It can be directly verified that in this case $\partial^2 E(U)/\partial F\,\partial B = 0$, so that the two second-order conditions for a maximum are

$$\begin{aligned} 2\frac{\partial^2 E(U)}{\partial F^2} = &\int_F^{F+B} U''\{-(\bar{A} - F) - BC_b\}\, dG(\bar{A}) \\ &+ (1 + C_0)\int_{F+B}^{+\infty} U''\{-(\bar{A} - F) - BC_b - C_0[\bar{A} - (F + B)]\}\, dG(\bar{A}) \\ &- U'[-BC_b]g(F) - C_0U'[-B - BC_b]g(F + B) < 0, \end{aligned} \tag{A.5}$$

and

$$\begin{aligned} 2\frac{\partial^2 E(U)}{\partial B^2} = &C_b^2\int_F^{F+B} U''\{-(\bar{A} - F) - BC_b\}\, dG(\bar{A}) \\ &+ (C_0 - C_b)^2\int_{F+B}^{+\infty} U''\{-(\bar{A} - F) - BC_b - C_0[\bar{A} - (F + B)]\}\, dG(\bar{A}) \\ &- C_0U'[-B - BC_b]g(F + B) < 0. \end{aligned} \tag{A.6}$$

A sufficient, but not necessary, condition for these to hold is that U'' be nonpositive almost everywhere.

In the symmetric case, it is possible to investigate the influence of attitudes toward risk on the choice of B. Define

$$\overline{U}' = \left\{\int_F^{F+B} U'\{-(\bar{A} - F) - BC_b\}\, dG(\bar{A})\right\} \Big/ \Big\{G(F + B) - G(F)\Big\}. \tag{A.7}$$

Employing (A.7) and the symmetry of $g(\bar{A})$, (A.3) may be written as

$$\begin{aligned} &-C_b[G(F + B) - G(F)]\overline{U}' \\ &\quad + (C_0 - C_b)\overline{U}'\int_{F+B}^{+\infty} \frac{U'\{-(\bar{A} - F) - BC_b - C_0[\bar{A} - (F + B)]\}}{\overline{U}'}\, dG(\bar{A}) = 0, \end{aligned} \tag{A.8}$$

or

$$-C_b[G(F + B) - \tfrac{1}{2}] + (C_0 - C_b)k[1 - G(F + B)] = 0, \tag{A.9}$$

where k is a constant which equals one if the subject is risk neutral and exceeds one if he is everywhere risk averse. Let $P = 2[1 - G(F + B)]$ be the subjective probability of $\bar{A}$ falling outside $[(F - B), (F + B)]$, and let $P_0 = C_b/C_0$ be the value of P that would be selected by a risk neutral subject Substituting into (A.9) and solving yields

$$P = \frac{P_0}{P_0 + k(1 - P_0)}. \tag{A.10}$$

A risk averse subject will thus select a smaller P (larger B) than a risk neutral subject when both are faced with the same subjective distribution. Similarly, a risk lover for whom (A.5) and (A.6) hold will select a larger P (smaller B) than a subject indifferent to risk.

REFERENCES

[1] BARTLETT, M. S.: "Tests of Significance in Factor Analysis," *British Journal of Psychology*, Statistical Section, 3 (1950), 77–85.

[2] BISCHOFF, C. W.: "Hypothesis Testing and the Demand for Capital Goods," *Review of Economics and Statistics*, 52 (1969), 354–368.
[3] CAGAN, P. D.: "The Monetary Dynamics of Hyperinflation," in *Studies in the Quantity Theory of Money*, ed. by M. Friedman. Chicago: University of Chicago Press, 1956.
[4] COCHRANE, D., AND G. H. ORCUTT: "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44 (1949), 32–61.
[5] COOPER, J. P.: "Asymptotic Covariance Matrix of Procedures for Linear Regression in the Presence of First Order Serially Correlated Disturbances," *Econometrica*, 40 (1972), 305–310.
[6] DAGENAIS, M. D.: "A Threshold Regression Model," *Econometrica*, 37 (1969), 193–203.
[7] DURBIN, J.: "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables," *Econometrica*, 38 (1970), 410–421.
[8] FERBER, R.: *The Railroad Shippers Forecasts*. Urbana, Illinois: Bureau of Economic and Business Research, University of Illinois, 1953.
[9] FISHER, F. M.: *A Priori Information and Time Series Analysis*. Amsterdam: North-Holland, 1962.
[10] FRIEDMAN, J. W.: "On Experimental Research in Oligopoly," *Review of Economic Studies*, 36 (1969), 399–416.
[11] GLEJSER, H.: "A New Test for Heteroskedasticity," *Journal of the American Statistical Association*, 64 (1969), 316–323.
[12] GOLDBERGER, A. S.: "Correlations Between Binary Outcomes and Probabilistic Predictions," *Journal of the American Statistical Association*, 68 (1973), 84.
[13] HICKS, J. R.: *Value and Capital*. Oxford: Clarendon Press, 1939.
[14] ———: *Capital and Growth*. Oxford: Oxford University Press, 1965.
[15] HILDEBRAND, F. B.: *Advanced Calculus for Applications*. Englewood Cliffs, New Jersey: Prentice-Hall, 1962.
[16] HIRSCH, A. A., AND M. C. LOVELL: *Sales Anticipations and Inventory Behavior*. New York: John Wiley, 1969.
[17] MCGILLIVRAY, R. G.: "Estimating the Linear Probability Function," *Econometrica*, 38 (1970), 775–776.
[18] METZLER, L.: "The Nature and Stability of Inventory Cycles," *Review of Economic Statistics*, 29 (1941), 113–129.
[19] MORRISON, D. G.: "Upper Bounds for Correlations Between Binary Outcomes and Probabilistic Predictions," *Journal of the American Statistical Association*, 67 (1972), 68–70.
[20] NERLOVE, M.: *The Dynamics of Supply: Estimation of Farmers' Response to Price*. Baltimore: Johns Hopkins University Press, 1958.
[21] ———: "Adaptive Expectations and Cobweb Phenomena," *Quarterly Journal of Economics*, 72 (1958), 227–240.
[22] SMITH, V. K., AND C. J. CICCHETTI: "Regression Analysis with Dichotomous Dependent Variables," paper presented at the meeting of the Econometric Society, December, 1972.
[23] SWAMY, P. A. V. B.: *Statistical Inference in Random Coefficient Regression Models*. Berlin: Springer-Verlag, 1971.
[24] THEIL, H.: *Principles of Econometrics*. New York: John Wiley, 1971.
[25] TURNOVSKY, S.: "Empirical Evidence on the Formation of Price Expectations," *Journal of the American Statistical Association*, 65 (1970), 1441–1454.
[26] ZELLNER, A.: "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57 (1962), 348–368.