Pergamon

Journal of International Money and Finance, Vol. 16, No. 6, pp. 969–987, 1997

Printed in Great Britain
0261-5606/97 $17.00 + 0.00

PII: S0261-5606(97)00015-6

The accuracy of OECD forecasts of the international economy: balance of payments

JCK ASH

Department of Economics, University of Reading, Reading, Berkshire RG6 6AA, UK

DJ SMYTH

Department of Economics, Louisiana State University, College of Business Administration, Baton Rouge, LA 70803, USA

AND

SM HERAVI*

Cardiff Business School, University of Wales, Cardiff, CF1 3AT, UK

This paper examines the accuracy of forecasts of the international economy made by the OECD. Our large data set, comprising some 5500 pairs of forecasts and outcomes, includes one-, two-, and three-step ahead semi-annual forecasts of eight components of the balance of payments for the G7 economies over a 20-year period. There is considerable variation in the accuracy of these forecasts. Although they are generally superior to naive and time-series predictions, there are some marked exceptions particularly as the forecast horizon lengthens. Forecasting error is overwhelmingly non-systematic. However, our study reveals numerous instances of forecasts which could be improved by a simple linear correction, or by incorporating information contained in known, recent forecast errors. The OECD's forecasts of services and private transfers, and official transfers are cause for particular concern: the accuracy of these forecasts is low, often below that of corresponding time-series forecasts, and rationality tests indicate that they are most prone to inefficiency and inconsistency. (JEL F17).

*The authors gratefully acknowledge the helpful comments and suggestions of the anonymous referees.

Since 1967 the Organization for Economic Cooperation and Development (OECD) has published semi-annual forecasts of economic activity in its seven largest Member countries — Canada, France, Germany, Italy, Japan, the UK and the USA. These forecasts, more recently extended to include all 24 members of the Organization, cover the major components of demand and output, inflation, and the balance of payments. According to Llewellyn *et al.* (1985, p. 98) the OECD aims to 'produce an integrated set of internationally consistent country forecasts, taking due account of linkages between economies'. Over the years the forecasting methods employed by the OECD have evolved from the systematic but relatively informal 'pooling or confronting' of Member country forecasts first mooted by McMahon (1965), to the current large INTERLINK system of formal macro-econometric models which ensures consistency in forecasting world trade flows, capital flows, and domestic economic developments. Llewellyn *et al.* (1985) explain the structure of INTERLINK and how the system is used for forecasting; OECD forecasting techniques are summarized in the Technical Appendix to each issue of *OECD Economic Outlook*; details of relevant research appear from time-to-time in the OECD Economics and Statistics Department's Working Papers and Occasional Studies (e.g. Richardson, 1988).

The OECD publishes its forecasts twice a year in the June/July and December issues of *OECD Economic Outlook*. The forecasts are for the half-year that is nearly over, and the two ensuing half-years. Thus a forecast published in June or July of year t covers the first and second halves of year t, and the first half of year $t + 1$; a forecast published in December of year t covers the second half of year t and the first and second halves of year $t + 1$.

In this paper we analyze the accuracy and various properties of these one-, two- and three-step ahead OECD forecasts, henceforth denoted by F1, F2 and F3, respectively. Problems of 'ragged edge' forecasting (Wallis, 1986, including lags in the availability of first estimates of outcome data and their subsequent revision, make F1 a genuine though very short-term forecast series. Llewellyn *et al.* (1985, pp. 75–76) state that 'it may not be possible to put together a satisfactory overall picture of the economy that is less than 4–6 months out of date'. We evaluate predictions of eight components of the balance of payments. A full list of variables is given in the Appendix. The sample period differs somewhat between variables and is shorter for the longer-term forecasts, but usually runs from the first half of 1967 through the second half of 1987. In all we inspect over 5500 pairs of forecasts and outcomes. The present paper is part of a larger study which also includes an analysis of 12 demand and output variables (Ash *et al.*, 1990, 1993).

Until now, no research into the accuracy of the OECD's disaggregated balance of payments forecasts has been published. Attention has focussed instead on forecasts of expenditure, output, inflation and occasionally the balance of payments as a single national aggregate. Moreover, with the exception of earlier studies by the present authors (Ash *et al.*, 1990, 1993; Smyth and Ash, 1975, 1981), previous analysts of OECD accuracy have worked with the Organization's annual forecasts, not the underlying semi-annual predictions which are more relevant to the timing of policy initiatives. None have inspected forecasts which are more than one-step ahead, though Holden *et al.* (1987) have investigated the way in which annual forecasts are revised between December and July issues of *Economic Outlook*. The OECD itself has published two post-mortems on its forecasts: Ballis (1989) for the period 1982–1987 covering all member countries, and Llewellyn and Arai (1984) for the period 1967–1982 for 14 OECD economies. These in-house studies contain a wealth of descriptive information relating to individual forecasts, but little in the way of formal econometric analysis. Accuracy is measured only by the mean absolute error, and there are no diagnostic checks beyond a regression of outcomes on forecasts by Llewellyn and Arai (1984), who also compare the OECD's performance with that of competing naive and time-series models. Garcia-Ferrer *et al.* (1987) and more recently Mittnik (1990) have produced time-series models which give better annual growth forecasts than the majority of OECD predictions for the seven countries analysed by Smyth (1983). Smyth's study also includes OECD forecasts of inflation and the balance of payments for the period 1968–1979: a range of accuracy measures are computed, and the OECD forecasts are compared with the predictions from a random walk model. Similar techniques are used by Holden *et al.* (1987) in their study of eight variables forecast by OECD for all 24 Member countries over the period 1976–1984. Artis (1988) compares OECD and International Monetary Fund forecasts of output, inflation and the balance of payments for the Group of Seven countries individually and in aggregate: over the 13 years between 1973 and 1985 he concludes (p. 20) that 'there is generally little to choose between the two sets of forecasts. . . the two organizations tend to make the same errors about the same variables for the same countries at the same time'.

Our examination of the OECD's track-record is what the Organization itself would describe as a practical test of its forecasts, not a technical test. In a technical test the forecasts are corrected for errors in the exogenous assumptions. The OECD forecasts usually assume unchanged government policies, nominal exchange rates, and dollar-denominated oil prices, unless intended policy changes are known in sufficient detail in advance of the forecasts. Because of lags in the implementation and impact of post-forecast policy changes, such measures are likely to have a significant effect on the accuracy of only the 18-month ahead F3 forecasts. In any case the OECD's own studies surprisingly suggest that forecast errors might have been even larger had they been based on actual rather than assumed values for policy instruments, exchange rates and oil prices (Ballis, 1989 p. 25). Thus partly for these reasons, and partly because of the impossibility for anyone other than the OECD forecasters themselves to separate out judgmental adjustments from exogenous assumptions, we do not attempt to correct the forecast data for the latter source of error.

I. Techniques of accuracy analysis

On two occasions each year, in the June/July and December issues of *OECD Economic Outlook*, the OECD publishes semi-annual forecasts for the ensuing

18 months (3 half-years). Predictions and outcomes for the current balance and its five components — exports, imports, the trade balance, net services and private transfers, and net official transfers — are published as seasonally adjusted levels of expenditure, in millions of US dollars at current prices. Granger and Newbold (1973, p. 44) point out that because the typical time-series of economic levels is a near random walk and because one random walk can appear to predict another independent random walk rather well, it is more meaningful to evaluate forecasts in terms of predicted and actual changes. We have therefore first-differenced the OECD data as follows. Forecasts and outcomes of exports and imports are expressed as percentage changes from the previous period at annual rates. Data on the trade balance, net services and private transfers, net official transfers and the current balance are analysed as simple first differences from the previous period. These balances or net items are volatile, comparatively small expenditures, sometimes fluctuating between positive and negative values in levels: percentage changes in these series are often large — infinite when calculated from a zero base, which is not infrequent. In addition, the OECD publishes forecasts of the volume of exports and imports as constant price percentage changes from the previous period, seasonally adjusted at annual rates, and it is in this form that we evaluate their accuracy.

Our choice of outcome data are the corresponding figures published 6 months after the event; thus a forecast for period t is evaluated against data published in $t+1$. These constant vintage outcomes are certainly subject to later revision. However, they are neither provisional flash estimates nor late revisions incorporating composition changes or changes in base. They probably most closely measure what the forecasters themselves are trying to predict.

Let F_t; $t = 1, 2,...,N$ be the forecast time-series and let A_t; $t = 1, 2...,N$ be the time-series of corresponding outcomes. The means of the forecasts and outcomes are denoted FM and AM, respectively, and the corresponding standard deviations F.S.D. and A.S.D. A popular and readily intelligible measure of overall accuracy, one consistent with a notional quadratic loss function and the obvious measure of forecast quality if one is using a least-squares criterion in model estimation, is the average squared forecast error

$$\langle 1 \rangle \qquad \mathrm{MSE} = \Sigma(F_t - A_t)^2/N$$

or the root mean square error

$$\langle 2 \rangle \qquad \mathrm{RMSE} = \sqrt{\Sigma(F_t - A_t)^2/N}$$

where this and subsequent summations Σ are for $t = 1$ to N.

In order to compare the accuracy of forecasts for different variables and/or different time periods, the RMSE should be standardised for the normal magnitude of the outcome, A, during the time period under consideration (Chatfield, 1988). Thus the standardised RMSE is the widely-used Theil (1966) inequality coefficient, U:

$$\langle 3 \rangle \qquad U = \mathrm{RMSE}/\sqrt{(\Sigma A_t^2/N)}$$

The inequality coefficient is zero only in the case of perfectly accurate forecasts, rises with inaccuracy, and has no upper bound. Note that if $F_t = 0$ for all t, $U = 1$. As all the variables which we examine here are changes not levels, $U = 1$ for any series of OECD forecasts is as inaccurate as a naive repetitive no-change prediction.

Theil inequality coefficients are calculated for OECD forecasts and also for competing predictions from an auto-regressive time-series model. Time-series models of order AR(1) through AR(5) were fitted to the data. Akaike's Information Criterion (Priestley, 1981 pp. 372–376) was used to select the model with best fit. This model was then sequentially re-estimated over the final 20 half-years of our sample period, adding one new observation at a time. After each re-estimation, forecasts were generated for up to three steps ahead, and compared with the corresponding forecasts made by the OECD.

A number of diagnostic checks are performed on the forecasts using the full sample period. The square of the numerator of the inequality coefficient is the MSE which, as Theil (1961) demonstrates, can be decomposed in two alternative ways. The decomposition preferred by Granger and Newbold (1973) results in the following inequality proportions:

$$\text{UM} = (\text{FM} - \text{AM})^2/\text{MSE} \tag{4}$$

$$\text{UR} = (\text{FSD} - \text{R.ASD})^2/\text{MSE} \tag{5}$$

$$\text{UD} = (1 - R^2).\text{ASD}^2/\text{MSE} \tag{6}$$

R is the sample correlation coefficient between predictions and outcomes. UM is the proportion of forecasting error (as measured by MSE) due to bias in the sense of over- or underpredicting the mean; UR is the regression proportion or the proportion of error due to misforecasting the systematic component of the variance of outcomes; UD is the disturbance or residual proportion. Both UM and UR tend to zero for the optimum predictor, and so UD should tend to unity.

Consider the regression of outcomes on predictions

$$A_t = \alpha_0 + \alpha_1.F_t + u_t \tag{7}$$

For a series of optimal forecasts the following conditions hold:

$$\alpha_0 = 0;\ \text{FM} = \text{AM};\ \text{UM} = 0$$

$$\alpha_1 = 1;\ \text{FSD} = \text{R.ASD};\ \text{UR} = 0$$

If these conditions do not hold, the forecasts' accuracy could, in principle,be improved by a simple linear correction.

Tests based on ⟨7⟩ must explicitly recognize that the timing of information available to the forecaster and the length of the forecast horizon have important implications for the random term u (see for example Holden and Peel, 1985). Figures for the previous half-year are available to OECD economists preparing their forecasts for publication in the June/July and December issues of *Economic Outlook*. Thus in this instance there is no problem of information lag. However, even with full current information a problem arises when the forecast horizon is more than one period ahead as is the case with the OECD's

F2 and F3 forecasts. Here the forecasts overlap, some intervening outcomes and errors are unknown, and errors are serially correlated. As Brown and Maital (1981) show, an s-step ahead forecast will have an error which follows a moving average process of order $(s - 1)$. The error term u_t in $\langle 7 \rangle$ follows a first-order moving average process for the F2 forecasts, and a second-order moving average process for F3. Thus we follow Blake *et al.* (1986) and Holden and Peel (1985), and use methods proposed by Hansen and Hodrick (1980) in order to obtain consistent unbiased estimates of α_0 and α_1 by OLS, and an efficient estimator of the parameter covariance matrix by GLS.

Two final diagnostic checks are reported in the text though not tabulated. A simple test of whether accuracy has improved or deteriorated during the sample period involves regressing the absolute value of each forecast error, relative to the corresponding outcome, against a time trend. For some purposes, such as the successful timing of changes in the direction of policy, it may be more harmful to make a smaller prediction error yet misforecast the direction of change than to make a larger, directionally-correct error. Therefore the overall prevalence of turning point errors is also calculated.

We subject the forecasts to a number of rationality tests — for unbiasedness, efficiency and consistency. Unbiased forecasts have the same mean as the actual outcomes. Efficiency requires that the forecasts utilise all relevant information available at the time that they are made; if they did not, accuracy could be improved by incorporating the extra unused information. Forecasts which span overlapping time periods should use known information consistently, such that one cannot predict the way in which these forecasts are revised. Consistent forecasts are therefore revised only in the light of new information. If the OECD forecasts do not possess these properties of rationality, their accuracy could be improved by using publicly available information apparently ignored by the forecasters. As well as summarizing the theory of rational expectations, Holden *et al.* (1985) describe a range of test procedures and survey the empirical evidence: they conclude that economic expectations and forecasts in general do not support the rational expectations hypothesis.

The standard test for unbiasedness is to estimate $\langle 7 \rangle$ and, following Holden and Peel (1985), to test the joint hypothesis that $\hat{\alpha}_0 = 0$ and $\hat{\alpha}_1 = 1$ using a $\chi^2(2)$ statistic. We perform this test but do not report the results here. The Hansen and Hodrick estimation procedure explicitly assumes that errors follow an $(s - 1)$ moving average process, and that no other error correlations are present at lags greater than $(s - 1)$. If this latter assumption were incorrect the above unbiasedness test would be inappropriate. Also, Holden and Peel (1990) demonstrate that testing the joint hypothesis that $\hat{a}_0 = 0$ and $\hat{a}_1 = 1$ is a sufficient but not necessary condition for unbiasedness. They propose a further test performed directly on the forecast errors:

$$\langle 8 \rangle \qquad A_t - F_t = E_t = \mu + u_t, \quad \text{TM is a } T(n-1) \text{ test for } \mu = 0$$

where μ is the mean error, and u_t follows the appropriate moving average process.

Efficient forecasts incorporate all relevant, available information. As a test of this property, Mullineaux (1978) proposes that the forecast error should be uncorrelated with any element in the set of information available at the time the forecasts are prepared. Important elements in this information set are past outcomes, forecasts and errors. Forecasts for s steps ahead ought to have error autocorrelations of order s or greater equal to zero, for otherwise the error E_t will be correlated with information available to the forecasters which could have been used to improve the forecast. We therefore compute Box–Ljung statistics for both the forecast errors, E_t, and the residuals from ⟨7⟩. The Box–Ljung statistic, BL, is defined as

⟨9⟩ $$BL = \frac{(N+2)\sum_{j=s}^{n}(N-j)^{-1}\hat{\rho}_j^2}{\left(1+2\sum_{j=1}^{s-1}\hat{\rho}_j^2\right)/N}$$

where the ρ_j are estimated autocorrelation coefficients, N is the number of observations, and $n = 4$ are the number of lags used in these tests. Note that a comparison of the performance of the OECD forecasts with that of the time-series model is a further test of relative efficiency.

Revisions to consistent forecasts cannot be predicted. Assuming the initial forecast exploits all available information and is unbiased, revisions should reflect new information only. The initial forecast and each revision to it should be an unbiased predictor of its successor. Thus XC is a $\chi^2(2)$ test for consistency in successive revisions to forecasts, a joint test on $\hat{\gamma}_0 = 0$ and $\hat{\gamma}_1 = 1$ in the regression

⟨10⟩ $${}_{t-s}F_t = \gamma_0 + \gamma_{1\,t-s-1}F_t + u_t \quad s = 1, 2$$

The results of our analysis are summarized in three tables showing, respectively, Theil inequality coefficients, the UD residual proportion, and three tests for rationality: The TM bias tests, the BL efficiency test on the forecast errors, and the XC test for consistency. For these latter three tests, * after a statistic denotes departure from the rationality criterion at the 5% significance level. The accompanying text discusses these specific results, and also comments on the more interesting findings from the other tests described above. A full report by Ash *et al.* (1995), presents alternative measures of the accuracy of OECD and time-series forecasts, for example using the mean absolute error criterion, and gives detailed results for the diagnostic checks and rationality tests. This paper is available to readers on request.

II. Accuracy of OECD forecasts

Table 1 shows inequality coefficients computed for the OECD forecasts over a constant sample period of 20 semi-annual observations for the decade 1978–1987. Their mean value, averaged across F1, F2, and F3 forecasts for all countries and variables, is 0.79. In comparison, the corresponding figure for the OECD's demand, output and inflation forecasts is 0.65 (Ash *et al.*, 1990, p. 385). However, this average conceals great variation in the accuracy with which individual series are forecast. For example, the most accurately forecast series

TABLE 1. Accuracy of OECD forecasts: *U* values

	Canada	France	Germany	Italy	Japan	UK	USA
X							
F1	0.656	0.611	0.400	0.765	0.432	0.596	0.320
F2	0.805	0.837	0.870	0.839	0.662	0.874	0.634
F3	0.800	0.941	0.976	0.870	0.854	0.881	0.727
M							
F1	0.680	0.476	0.451	0.722	0.562	0.475	0.381
F2	0.863	0.651	0.643	0.809	0.806	0.861	0.651
F3	0.873	0.723	0.793	0.795	0.822	0.888	0.805
X%							
F1	0.606	0.325	0.466	0.612	0.254	0.276	0.265
F2	0.787	0.853	0.859	0.822	0.603	0.703	0.573
F3	0.904	0.956	0.947	0.863	0.779	0.826	0.742
M%							
F1	0.512	0.275	0.271	0.523	0.411	0.451	0.252
F2	0.791	0.741	0.741	0.789	0.779	0.899	0.630
F3	0.815	0.869	0.884	0.811	0.808	0.898	0.813
TB							
F1	0.748	0.586	0.543	0.734	0.356	0.497	0.468
F2	1.221	0.884	0.913	1.000	0.836	0.871	0.734
F3	0.970	0.949	0.935	1.029	0.804	0.912	0.791
CB							
F1	0.760	0.540	0.589	0.707	0.345	1.007	0.438
F2	1.230	0.855	0.810	0.981	0.791	0.840	0.722
F3	0.989	1.009	0.927	1.042	0.814	0.910	0.731
STI							
F1	0.831	0.959	0.868	0.946	0.989	0.975	0.892
F2	0.928	1.007	0.971	1.087	1.056	1.063	1.030
F3	1.077	1.089	1.226	1.112	0.900	1.048	1.021
OT							
F1	0.972	0.861	0.831	0.958	0.820	0.983	0.850
F2	1.098	0.990	0.926	1.020	0.903	1.014	1.001
F3	0.999	0.990	0.955	1.006	0.918	1.100	0.971

— recording the lowest value for the inequality coefficient, 0.25 — is the F1 forecast of US current price imports. At the other extreme of the range are the F3 forecasts of the net flow of services and private transfers for Germany: the inequality coefficient is 1.23, worse than would have resulted from naively forecasting no-change.

In order to obtain an overview of the accuracy of these OECD forecasts, the inequality coefficients can be pooled and then averaged by variable across countries. When this is done, the forecast variables appear to fall into four pairs, each pair of variables being forecast with very similar accuracy. Most accurate are current price imports and exports: the inequality coefficient is 0.64 for both. Perhaps not surprisingly, import and export volumes are predicted rather less accurately, with inequality coefficients of 0.70 and 0.73, respectively. The third pair of forecast series are the two balancing items: the mean value of the inequality coefficient is 0.80 for the trade balance and 0.81 for the current balance. The least accurate balance of payments forecasts are for net changes in official transfers and net changes in services and private transfers: the respective inequality coefficients are 0.96 and 1.00, the latter as (in)accurate as the naive no-change predictor. Previous research has revealed a marked difference in the accuracy with which the OECD forecasts economic activity in the two North American neighbours, Canada and the USA (Ash *et al.*, 1993). This finding is confirmed and indeed underlined here, for the G7 economies the OECD's balance of payments forecasts are least accurate for Canada and most accurate for the USA. Mean values of 0.87 (Canada) and 0.69 (USA) result from averaging the inequality coefficients across variables.

As one would expect, the accuracy of the OECD's forecasts deteriorates with the length of the forecasts' horizon. Generally a sharp decline in accuracy occurs as prediction extends from one to two steps ahead: the pooled inequality coefficients are 0.61 for the F1 forecast and 0.86 for F2. There is only one instance of a series being better forecast two steps ahead — the current balance for the UK. With an inequality coefficient of 1.007 these short-run F1 forecasts of the UK current balance are no more accurate than a naive no change prediction, and are less accurate than not only the year-ahead forecasts but also those for 18 months ahead. With a pooled inequality coefficient of 0.91 for F3 forecasts, the decline in accuracy between two and three steps ahead is less pronounced. Indeed for 12 series (four for Canada, and three for both services and private transfers, and official transfers) the three-step ahead forecast is the more accurate.

If the Theil inequality coefficient is greater than unity the performance of forecasters predicting changes is worse than that of a very naive model, the assumption of no change from the previous period. In all, 24 of the 168 OECD inequality coefficients equal or exceed unity. All but one, for the UK current balance, are F2 or F3 forecasts. Their incidence is fairly evenly divided between the 1-year and 18 months ahead predictions, at 12 and 11 series, respectively, and 11 of these examples of poor forecasting relate to changes in services and private transfers. The OECD's success at very short-run forecasting is underlined by comparison with the time-series forecasts, the details of which are given in Ash *et al.* (1995). All but two of the OECD's F1 forecasts are more accurate in terms of a lower inequality coefficient. The exceptions are official transfers for the UK and USA. When the time horizon lengthens to 1 year and 18 months ahead, approximately 21% of the OECD's F2 forecasts and 32% of the F3 forecasts are beaten by the time-series predictor; these proportions are about double those reported for the OECD's demand, output and inflation forecasts (Ash *et al.*, 1990). Again it is the OECD's forecasts of services, and private and official transfers which are particularly vulnerable on this account.[1]

Any comparison between the OECD and time-series forecasts should be tempered by the knowledge that the simple autoregressive time-series model itself is clearly not a very accurate predictor. Its pooled inequality coefficients are 0.97, 1.00 and 1.00 for F1, F2 and F3 forecasts respectively. However, it does provide a useful benchmark against which to evaluate the relative efficiency of the OECD forecasts. Thus relative to the time-series forecasts the OECD's accuracy declines as the forecast horizon lengthens: the ratios of OECD to time-series pooled inequality coefficients are 0.63, 0.86 and 0.91 for F1, F2 and F3 forecasts. This is further evidence that the OECD has greater difficulty forecasting the balance of payments than demand, output and prices: from our earlier work the corresponding ratios are 0.53, 0.79 and 0.84 (Ash *et al.*, 1990, p. 386). The OECD's F2 and F3 forecasts of official transfers and F3 forecasts of services and private transfers are the only series for which the pooled relative inequality coefficient error exceeds unity. Over all variables the OECD's forecasts for Japan are the most accurate compared with the time-series model: pooling the forecasts yields an average value of 0.71 for the relative inequality coefficient. Not that Japan is intrinsically an easy economy to forecast, if high inequality coefficients for its time-series forecasts are the criterion. Canada on the other hand typically has lower inequality values for its time-series predictions, yet in comparison the accuracy of the OECD forecasts is worse. Indeed the Canadian F2 and F3 forecasts are the only ones which, on average, are (marginally) less accurate than our time-series predictions. Both Smyth (1983) and Llewellyn and Arai (1984) support our general finding that the OECD forecasts are superior to naive and simple time-series forecasts; using more sophisticated time-series models, Garcia-Ferrer *et al.* (1987) and Mittnik (1990) generate better forecasts than the OECD for economic growth in a majority of the G7 countries, though there is no other published evidence on the success or otherwise of forecasting the balance of payments using time-series methods.

Table 2 shows the random proportion, UD, of the mean square error in the OECD forecasts. This diagnostic check uses the full data set available: 1967–1987 for F1, 1968–1987 for F2, and 1973–1987 for F3. Judged by the size of these inequality proportions, forecasting error is predominantly non-systematic. For example the mean value of UD is 93.4% for the F1 forecasts, 92.5% for F2 forecasts, and 91.1% for F3 forecasts. These are commendably low values: indeed out of the 168 forecast series which we analyse only 10 series, or 6% of the total, record UD proportions $\leq 75\%$. When systematic error does occur it is usually due to misforecasting the non-random component of fluctuations (there are only three instances of underpredicting the average growth rate). In terms of mean UD, systematic error is most marked in F3 forecasts of services and private transfers.

Full details of the diagnostic analysis can be found in Ash *et al.* (1995). Here we merely highlight some of our more interesting findings. Although most forecasting error is random and large systematic errors are rare, a greater proportion of these OECD forecasts fail the optimality test; that is, in ⟨7⟩

TABLE 2. Error diagnostics of OECD forecasts: UD values

	Canada	France	Germany	Italy	Japan	UK	USA
X							
F1	0.956	0.994	0.806	0.959	0.997	0.871	0.989
F2	0.857	0.967	0.998	0.997	0.942	0.922	0.974
F3	0.890	0.814	0.897	0.926	0.994	0.874	0.968
M							
F1	0.887	0.878	0.956	0.921	0.994	0.788	0.992
F2	0.996	0.971	0.882	0.994	0.921	0.993	0.748
F3	0.986	0.959	0.728	0.990	0.891	0.997	0.974
X%							
F1	0.990	0.990	0.981	0.997	0.977	0.803	0.940
F2	0.984	0.996	0.933	1.000	0.746	0.972	0.904
F3	0.888	0.923	0.960	0.975	0.967	0.984	0.811
M%							
F1	0.935	0.950	0.936	0.987	0.889	0.975	0.935
F2	0.983	0.959	0.921	0.969	0.963	0.977	0.813
F3	0.988	0.965	0.980	0.674	0.709	0.974	0.959
TB							
F1	0.989	0.927	0.965	0.938	0.876	0.970	0.866
F2	0.641	0.880	0.884	0.986	0.956	0.992	0.931
F3	0.970	0.940	0.936	0.949	0.894	0.986	0.952
CB							
F1	0.981	0.993	0.991	0.893	0.836	0.856	0.879
F2	0.583	0.929	0.941	0.992	0.918	0.994	0.913
F3	0.959	0.953	0.926	0.937	0.899	0.998	0.941
STI							
F1	0.921	0.806	0.974	0.903	0.966	0.830	0.969
F2	0.940	0.908	0.985	0.836	0.899	0.823	0.942
F3	0.685	0.794	0.630	0.746	0.840	0.829	0.962
OT							
F1	0.967	0.996	0.955	0.968	0.987	0.863	0.971
F2	0.838	0.995	0.999	0.941	0.982	0.889	0.972
F3	0.907	0.997	0.993	0.974	0.968	0.786	0.999

either $\hat{\alpha}_0 \neq 0$ or $\hat{\alpha}_1 \neq 1$ or both. A simple linear correction would then in principle improve the accuracy of these forecasts. Nine of the F1 forecasts fail this test, 10 of the F2 forecasts, and 15 of the F3 forecasts: 20% overall. Nine out of 21 forecasts of services and private transfers fail on this account, with only the US forecasts fully passing this test. No single source of systematic error is apparent in these forecasts: in five instances the slope term in the regression of outcomes on predictions is negative, and in another three cases it is positive but significantly less than one, indicating that the systematic component of the variance in the series is overestimated by the forecasters. In general, the largest incidence of series which fail the optimality test occurs in the UK forecasts: only the UK forecasts of current price imports and the trade balance are fully optimal, whereas one-, two-, and three-step ahead forecasts of official transfers all either overpredict mean changes or misforecast systematic variations in the series, or both. In addition to these instances of significant systematic error, there are 74 series (out of a total of 168) for which there is no significant correlation between forecasts and outcomes. The heaviest incidence occurs among the F3 18-month ahead forecasts, as well as the F2 forecasts for services and private transfers, and official transfers. Apart from these two variables, this incidence is highest in the Canadian forecasts, and lowest in those for Japan and the USA.

Given the stochastic nature of economic processes, one would expect an optimal predictor to display a smaller variance than the associated outcome (Mincer and Zarnowitz, 1969; Granger and Newbold, 1973; Hatanaka, 1975; Samuelson, 1976). This is indeed the norm for these OECD forecasts; the standard deviation of forecasts exceeds the standard deviation of outcomes in only three out of a total of 168 pairs of series examined here. Fluctuations are underforecast in spite of the evidence, reported above, of some tendency towards countervailing inefficiency, building in greater systematic variation than is warranted.

Not surprisingly, the incidence of turning point errors generally increases with the time horizon of the forecasts: 18% of one-step ahead forecasts, 24% of two-step ahead forecasts, and 28% of three-step ahead forecasts are wrong-signed, that is, the direction of change is misforecast. This pattern holds for all variables except services and private transfers, and official transfers, for both of which the incidence of turning point errors in the F1 forecasts is over 30%, but declines as the outlook lengthens. This unexpected pattern in the incidence of turning-point errors is particularly marked for the official transfers forecasts: pooling data on all seven economies yields mean proportions of turning-point errors of 0.34, 0.16, and 0.15 for F1, F2 and F3 forecasts, respectively. Forecasts with the greatest incidence of turning-point errors are those for the current balance, the trade balance, and services and private transfers: in each case the proportion of turning-point errors averages approximately 30%. However, again there is wide variation, for example, at least half the three-step ahead predicted changes in the current balance for France and Italy are wrong-signed. On the other hand only 17% and 20%, respectively, of the Japanese trade balance and current balance F3 forecasts commit turning point errors, lower proportions than the corresponding F2 forecasts. Turning point errors are least evident in forecasts of current price exports and imports, averaging 17% overall.[2]

Our long sample period, from 1967 to 1987, includes shocks to the world economy from two major oil price rises. Holden *et al.* (1987) found that the oil price rises between 1979 and 1980 had no effect on the OECD's predictive accuracy. However, using the results of Smyth (1983) they show that the OECD's accuracy decreased at the time of the first oil shock between 1973 and 1974. The years covered by our analysis also include the breakdown of the Phillips curve, shifts in the personal savings ratio in response to changes in inflation, increasing financial linkage between individual OECD economies, and major changes in exchange rate policies. During this period the OECD's formal and informal forecasting procedures have evolved in response to these and other developments, not least as forecasters learn from past mistakes. In order to assess the extent to which the forecasts have improved or deteriorated over time, we regress the absolute value of the error, relative to the outcome, against a time trend. Overall there is evidence that the OECD's forecasting accuracy has declined as 52.3% of the series show absolute errors rising through the sample period. Altogether there are 12 instances of a significant (at the 5% level) deterioration in accuracy against only six significant improvements. Forecasts of Japanese import volumes and of UK current price exports and imports show the clearest evidence of a decline in accuracy.

Rationality criteria require forecasts to be unbiased, to make an efficient use of all available information, and to be revised in a consistent manner. We submit the forecasts to two tests for bias, and in Table 3 we tabulate the results for one of them, TM the *T*-test for a mean forecast error of zero which Holden and Peel (1990) recommend as a necessary and sufficient condition for unbiasedness. Of the 12 forecasts which fail this test a majority, seven, are biased upwards, that is they overestimate the mean value of the corresponding outcomes. The incidence of bias tends to increase with the forecasts' time horizon. There are one, two and nine cases for F1, F2 and F3, respectively. While instances of bias are not concentrated in the forecasts of any particular variable, it is noteworthy that constant and current price forecasts of imports are bias-free, as are forecasts of the current balance. Four F3 forecasts for France are biased — for export volumes, the trade balance, services and private transfers, and official transfers. None of the Italian forecasts show any evidence of bias.

A paper by Lai (1990) highlights the difficulty of interpreting the findings of unbiasedness. Tests of efficiency and consistency, details of which can be found in Ash *et al.* (1995), shed additional light on the forecasts' rationality. We summarize these results in Table 3, which shows the Box–Ljung test for autocorrelation in the forecast errors, BL, and the χ^2 statistic, denoted by XC, testing jointly for zero intercept and unit slope in ⟨10⟩: the regression of F1 forecasts on the corresponding F2 forecasts, and the F2 forecasts on the corresponding F3 forecasts. As for the TM test, * denotes failing these rationality criteria at the 5% level.

There are 14 instances of a series failing the Box–Ljung efficiency test, five of which are for USA forecasts, and five of which are for the OECD's forecasts of official transfers. F1, F2, and F3 forecasts of US official transfers all fail. It is here that there is greatest scope for improved use of information available at the time the forecasts are made. Overall, the incidence of inefficiency falls with

TABLE 3. Rationality of OECD forecasts

	Canada			France			Germany			Italy			Japan			UK			USA		
	TM	BL	XC	TM	BL	XC	TM	BL	XC	TM	BL	XC	TM	BL	XC	TM	BL	XC	TM	BL	XC
X																					
F1	1.13	2.06	1.36	−0.19	4.83	0.20	0.30	3.00	7.42*	1.16	2.85	1.35	0.11	5.97	0.08	1.35	2.51	4.62	−0.56	4.59	0.36
F2	−0.28	1.40	4.53	−1.67	5.48	0.83	0.13	5.29	0.24	0.69	0.84	2.98	0.87	9.84*	0.98	0.21	2.77	3.13	0.20	4.94	1.58
F3	−0.30	1.74	3.70	−5.59*	5.82	5.84	−3.07*	6.76	6.79*	0.71	2.68	1.13	0.50	4.08	6.88*	0.67	2.87	0.30	0.03	6.02	1.30
M																					
F1	1.06	8.32	3.71	1.35	8.40	4.37	0.57	2.40	1.42	1.29	3.49	2.70	0.40	3.11	0.18	1.36	5.17	8.33*	0.35	9.72*	0.24
F2	−0.09	5.27	4.68	0.23	7.32	2.13	−0.99	2.26	2.66	0.36	6.43	1.78	−1.25	2.37	3.70	0.17	1.03	0.58	0.37	4.42	12.2*
F3	−0.06	4.79	1.58	−0.66	6.78	2.94	−0.83	2.12	2.42	−0.06	4.53	3.19	−0.29	5.29	4.09	−0.71	1.26	2.48	−0.07	3.10	3.58
X%																					
F1	−0.52	0.94	0.38	−0.56	3.09	0.38	0.48	0.86	0.75	0.27	1.83	0.13	0.78	5.78	0.93	−0.76	1.29	9.52*	−0.76	14.0*	2.49
F2	−0.69	1.31	0.71	0.09	4.33	0.80	1.12	1.37	3.62	0.04	2.19	0.07	2.34*	1.83	13.9*	0.54	1.32	3.33	0.80	5.52	6.49*
F3	−4.70*	1.36	2.65	0.03	4.27	0.34	0.46	1.06	0.29	−0.06	4.08	7.60*	0.90	1.94	4.63	0.56	2.75	0.84	0.43	4.72	2.12
M%																					
F1	1.46	3.70	2.75	1.16	7.02	2.03	1.18	10.6*	2.68	0.71	8.44	0.51	−0.83	4.29	4.86	0.59	8.67	1.02	1.65	2.54	2.77
F2	0.58	5.06	0.05	0.96	0.15	1.08	1.05	0.33	1.54	0.88	2.77	0.88	0.87	5.83	2.52	0.69	2.28	1.48	1.80	6.12	9.14*
F3	0.17	2.53	1.66	0.68	1.30	1.83	0.47	1.47	0.21	0.72	0.90	4.16	0.54	2.53	2.40	0.56	2.21	11.4*	1.17	0.27	7.15*
TB																					
F1	0.63	10.0*	0.43	−1.64	2.60	3.04	1.15	4.36	1.39	0.31	3.86	2.52	1.25	0.77	5.41	−0.82	0.67	1.17	−1.76	2.81	5.96
F2	−0.02	8.91	11.1*	−1.79	3.85	1.41	1.25	2.84	6.45*	−0.17	3.05	1.04	0.78	6.57	0.79	−0.40	1.81	0.84	−1.28	5.30	2.64
F3	−0.84	9.75*	14.6*	−8.83*	5.08	6.78*	0.74	1.20	1.85	−0.52	1.54	7.37*	0.32	4.00	6.36*	−0.40	1.14	0.83	−5.07*	2.66	1.99
CB																					
F1	0.65	1.46	0.74	−0.27	8.24	0.26	0.56	2.85	0.35	0.51	5.65	4.55	1.52	1.75	7.52*	0.79	9.43	6.39*	−1.36	6.04	5.24
F2	0.07	12.5*	14.9*	−1.10	3.76	1.79	1.30	3.12	7.04*	−0.58	1.58	2.75	0.96	4.26	0.74	−0.08	3.31	1.56	−1.58	2.39	1.36
F3	−0.85	7.54	21.3*	−1.33	2.90	1.41	1.06	1.24	0.10	−0.81	1.02	7.25*	0.46	2.38	6.70*	0.05	1.29	1.62	−1.78	2.00	5.39

Table 3 continued on opposite page

TABLE 3 (*Continued*)

STI																					
F1	−0.41	3.13	3.01	1.08	1.33	8.15*	0.97	9.08	0.96	−0.25	15.8*	3.53	0.78	2.56	1.22	2.05*	3.25	7.27*	0.37	2.94	1.08
F2	−1.55	4.06	2.62	1.27	12.5*	10.1*	0.29	9.46	9.78*	−0.47	6.83	11.1*	0.88	4.22	3.31	0.68	2.00	21.1*	0.11	2.91	0.23
F3	−5.11*	7.80	52.6*	4.28*	4.26	11.6*	0.82	5.38	26.6*	−0.49	5.25	6.48*	0.66	4.18	41.1*	1.00	1.42	13.7*	0.04	1.49	7.76*
OT																					
F1	0.16	3.22	1.18	0.17	2.97	0.14	−0.74	3.68	1.65	0.69	0.87	1.08	−0.63	5.25	0.45	−1.02	13.8*	5.55	−0.97	20.6*	1.03
F2	0.31	7.69	23.1*	1.62	1.56	6.03*	−0.07	4.08	1.32	0.63	1.21	2.25	−0.94	16.2*	2.25	−2.40*	0.80	14.1*	−1.04	26.3*	0.43
F3	2.00*	2.77	37.7*	3.79*	1.18	34.9*	−0.41	1.25	11.4*	0.98	2.36	2.79	−0.94	9.03	2.13	−0.59	2.13	3.83	0.01	13.4*	17.7*

* Denotes a departure from the rationality criterion at the 5% significance level.

the forecast horizon: seven of the failures are for F1 forecasts, five for F2, and two for F1.

The consistency test analyses F2 forecasts as predictors of F1 forecasts, and F3 forecasts as predictors of F2. The test is failed by 38 of the F2 and F3 forecasts; these forecasts require subsequent 'predictable' revisions. The XC value reported for F1 forecasts is the bias test for the joint hypothesis of zero intercept and unit slope in ⟨7⟩; we do not discuss these results here. Although this is a higher incidence of failure than observed in the OECD's demand, output and prices forecast (Ash *et al.*, 1990, p. 390), the relative infrequency of forecasts which fail the consistency test indicates that in general the OECD's forecast revisions are random and themselves unpredictable. Thus if the initial forecast is unbiased and uses all available information efficiently, it is revised only upon receipt of new information. Failure tends to be concentrated within the forecasts of services and private transfers, and official transfers (20 out of the 38). Elsewhere, marked inconsistency is confined to the forecasts of German export volumes, US current price imports, the Canadian trade balance, and the current balance for both Canada and Japan.

Looking at the rationality tests as a whole, our findings point to specific areas of the OECD's modelling and/or forecasting procedures which require improvement. For example, certain forecasts fail more than one rationality test: export volumes for Germany (biased and inconsistent) and Japan (inefficient and inconsistent); current price US imports (inefficient and inconsistent); the Canadian trade balance and current balance (inefficient and inconsistent); forecasts of services and private transfers for Canada and the UK (biased and inconsistent), Italy (inefficient and inconsistent), and France (all three criteria); forecasts of official transfers for Canada (biased and inconsistent), the US (inefficient and inconsistent), France and the UK (all three criteria).

On the other hand, other OECD forecasts survive the rationality tests remarkably well. These include export volumes for Canada, Italy, the UK and USA, import volumes for all countries except the USA, current price exports from France, Germany and the UK, current price imports into Canada, France, Italy and Japan, the UK trade balance, the French, UK and US current balance, and Italian official transfers.

III. Conclusions

In this paper we have subjected a large set of OECD balance of payments forecasts to a battery of tests for accuracy, error diagnostics and rationality. Our results indicate that the OECD is less successful in forecasting balance of payments variables than demand, output and prices. There is considerable variability in the accuracy with which different balance of payments variables are forecast for different G7 countries. Current price exports and imports are forecast more accurately than the corresponding volumes, and significantly more accurately than the trade balance and the current balance. Least accurate are forecasts of net flows of services and private transfers, and net official transfers.

Compared with forecasts generated by a simple autoregressive time-series model and with the naive no-change prediction the OECD's track record is usually one of success, albeit modest success. The very short-run F1 forecasts mostly outperform our competing predictions. However, over 20% of the 1-year ahead forecasts, and over 30% of the 18-month ahead forecasts are less accurate than the time-series model, and approximately 20% of these F2 and F3 forecasts are even beaten by the naive, no-change prediction. So the longer the economic outlook, the lower is the OECD's absolute and relative accuracy. Forecasting error is overwhelmingly non-systematic, and this commendably low systematic proportion is maintained as the forecasts' time horizon lengthens. Over the 20 years covered by our study, there is more evidence of deterioration in the OECD's accuracy than of improvement.

Although systematic error does not usually make a large contribution to inaccuracy, the diagnostic checks and rationality tests reveal numerous instances of forecasts which could in principle be improved by a simple linear correction, or by incorporating information contained in known, recent forecast errors.

Our study therefore pinpoints specific areas for future research. Chief among these must be the OECD's procedures for forecasting services and private transfers, and official transfers. The accuracy of the forecasts is typically low, often below that of corresponding time-series forecasts. Rationality tests indicate that these forecasts are most prone to inefficiency and inconsistency. The work of Garcia-Ferrer *et al.* (1987) and Mittnik (1990) suggests that this might be a fruitful area for more sophisticated time-series analysis, particularly since economic theory sheds relatively little light on the international flow of transfer payments.

Rational forecasts should be unbiased, efficient and consistent. Overall, only 22 out of 56 series pass all rationality criteria when the F1, F2 and F3 forecasts for each variable are amalgamated. An alternative and, from the OECD's viewpoint, a more sanguine perspective on rationality is gained by treating the F1, F2 and F3 series as separate forecasts of each variable. Then, 52 out of 168 forecasts fail at least one rationality test at the 0.05 significance level, approximately a 30% rate of failure — or a 70% success rate, depending on the reader's point of view. Whatever one's judgement on the OECD forecasters for these results, their track-record gives no support for the general empirical validity of the rational expectations hypothesis.

Appendix: List of variables

X, Volume of exports, percentage change.
M, Volume of imports, percentage change.
X%, Exports at current prices, percentage change.
M%, Imports at current prices, percentage change.
TB, Change in trade balance.
CB, Change in current balance.
STI, Change in services and private transfers, net.
OT, Change in official transfers, net.

F1, One step ahead forecasts.
F2, Two step ahead forecasts.
F3, Three step ahead forecasts.

Notes

1. Also the *U* statistic and its variants are descriptive, and cannot be used to test hypotheses per se. Thus even though these OECD forecasts are generally better than the time-series model, a statistical test might show that thedifferences are not statistically significant.
2. Although the percentage of turning point errors seems to be relatively low, a statistical test using a contingency table to test for the independence ofthe predicted and actual changes might show that the forecasts would not have been valuable to the users (see, for example Steckler, 1994).

References

Artis, M. J. (1988) *How Accurate is the World Economic Outlook? A Post Mortem on Short-Term Forecasting at the International Monetary Fund*. International Monetary Fund, Research Department, Staff Studies for the World Economic Outlook.

Ash, J. C. K., Smyth, D. J. and Heravi, S. M. (1990) The accuracy of OECD forecasts of the International Economy: demand, output and prices. *International Journal of Forecasting* **6**, 379–392.

Ash, J. C. K., Smyth, D. J. and Heravi, S. M. (1993) The accuracy of OECD forecasts for Canada and the United States. *North American Journal of Economics and Finance* **4**(2), 179–210.

Ash, J. C. K., Smyth, D. J. and Heravi, S. M. (1995) The accuracy of OECD forecasts of the International Economy: balance of payments. Discussion Papers in Quantitative Economics and Computing, Series E, No. 37. Department of Economics, University of Reading, 1995, and *Forecasting the International Economy*, Basingstoke: Macmillan, forthcoming.

Ballis, B. (1989) *A post-Mortem on OECD Short-Term Projections from 1982 to 1987*. OECD, Department of Economics and Statistics, Working Paper No. 65.

Blake, B., Beenstock, M. and Brasse, V. (1986) The performance of UK exchange rate forecasters. *The Economic Journal* **96**, 986–999.

Brown, B. W. and Maital, S. (1981) What do economists know? An empirical study of experts' expectations. *Econometrica* **49**, 491–504.

Chatfield, C. (1988) Editorial: apples, oranges and mean square error. *International Journal of Forecasting* **4**, 515–518.

Garcia-Ferrer, A., Highfield, R. A., Palm, F. and Zellner, A. (1987) Macroeconomic forecasting using pooled International data. *Journal of Business and Economics Statistics* **5**, 53–67.

Granger, C. W. J. and Newbold, P. (1973) Some comments on the evaluation of economic forecasts. *Applied Economics* **5**, 35–47.

Hansen, L. P. and Hodrick, R. J. (1980) Forward exchange rates as optimal predictors of future spot rates: an econometric analysis. *Journal of Political Economy* **88**, 829–853.

Hatanaka, M. (1975) The underestimation of variations in the forecast series: a note. *International Economic Review* **16**, 151–160.

Holden, K. and Peel, D. A. (1985) An evaluation of quarterly national institute forecasts. *Journal of Forecasting* **4**, 227–234.

Holden, K. and Peel, D. (1990) On testing for unbiasedness and efficiency of forecasts. *The Manchester School* **58**, 120–127.

Holden, K., Peel, D. and Sandhu, B. (1987) The accuracy of OECD forecasts. *Empirical Economics* **12**, 175–186.

Holden, K. Peel, D.A. and Thompson, J.L. (1985) *Expectations: Theory and Evidence*. Macmillan, London.

Lai, K.S. (1990) An evaluation of survey exchange rate forecasts. *Economics Letters* **32**, 61–65.

Llewellyn, J. and Arai, H. (1984) *'International Aspects of Forecasting Accuracy'*. OECD Economics Studies, No. 3.

Llewellyn, J., Potter, S. and Samuelson, L. (1985) *Economic Forecasting and Policy — The International Dimension*. Routledge and Kegan Paul, London.

McMahon, C. W. (1965) A general survey of the procedures and problems of official short-term economic forecasting. In *Techniques of Economic Forecasting*. OECD, Paris, Ch.1.

Mincer, J. and Zarnowitz, V. (1969) The evaluation of economic forecasts. In *Economic Forecasts and Expectations*, ed J. Mincer, pp. 1–46. National Bureau of Economic Research, New York.

Mittnik, S. (1990) Macroeconomic forecasting using pooled International data. *Journal of Business and Economic Statistics* **8**, 205–208.

Mullineaux, D. J. (1978) On testing for rationality: another look at the Livingston price expectations data. *Journal of Political Economy* **86**, 329–336.

Priestley, M. B. (1981) *Spectral Analysis and Time Series*. Academic Press, London.

Richardson, P. (1988) *The Structure and Simulation Properties of OECD's INTERLINK Model*. OECD Economic Studies No. 10.

Samuelson, P. A. (1976) Optimality of sluggish predictors under ergodic probabilities. *International Economic Review* **17**, 1–7.

Smyth, D. J. (1983) Short-run macroeconomic forecasting: the OECD performance. *Journal of Forecasting* **2**, 37–49.

Smyth, D. J. and Ash, J. C. K. (1975) Forecasting gross national product, the rate of inflation and the balance of trade: the OECD performance. *Economic Journal* **85**, 361–364.

Smyth, D. J. and Ash, J. C. K. (1981) The underestimation of forecasts and the variability of predictions and outcomes. *Bulletin of Economic Research* **33**, 37–44.

Steckler, H. O. (1994) Are economic forecasts valuable? *Journal of Forecasting* **13**, 495–505.

Theil, H. (1961) *Economic Forecasts and Policy*. Amsterdam, North Holland.

Theil, H. (1966) *Applied Economic Forecasting*. Amsterdam, North Holland.

Wallis, K. F. (1986) Forecasting with an econometric model: the 'ragged edge' problem. *Journal of Forecasting* **5**, 1–14.