7

COMPARATIVE ADVANTAGES OF TYPES OF FORECASTS

ADVANTAGES OF EXPERTISE IN SECTORAL FORECASTS

For one of the forecast sets (A) it is possible to compare three methods of arriving at a forecast of GNP. Each of the participants submits a forecast of total GNP. The mean of these forecasts, which we now designate A(M), is the set used for group A so far in our tables. In addition, each of the members of the group forecasts a particular sector or aspect of the economy. By consolidating these forecasts of each expenditure sector of GNP (personal consumption, government spending, etc.), a single forecast of total GNP is obtained, to be denoted A(S). The third procedure is to project the trend in real potential GNP with the aid of labor force and productivity estimates and to adjust the result by means of the average group A forecasts of unemployment and prices. The forecast thus derived will be called A(L). The forecasts A(S) and A(L) are available for semiannual or annual periods in 1955–58 and for the fourth quarters of the coming year since 1959. In Table 22, they are compared with each other and with some other fourth-quarter forecasts.

Among the three methods, the sum-of-sectors forecast A(S) has the smallest average error, followed by the labor-input productivity forecast A(L), while A(M) ranks third in terms of both level and change errors. The superiority of A(S) and A(L) over A(M) is sufficient to make their average errors smaller than those for the other forecasts for which we have a comparable record.[1] As between the A(S) and A(L) models, while the former produced smaller absolute errors, the latter

[1] Incidentally, the fourth-quarter forecasts on which the comparison is based have in general substantially larger errors than the annual forecasts made at the same time (which would be expected, of course, since their spans are longer).

TABLE 22

Selected Summary Measures of Error in GNP Forecasts of Fourth Quarter of Following Year, 1955-63 and 1959-63

(billion dollars)

Line	Forecast Set[a]	Forecasts of Levels: Mean Absolute Error, $\overline{|E|}$ (1)	Forecasts of Levels: Mean Error, $\bar{E}$ (2)	Forecasts of Levels: Root Mean Square Error (M_P) (3)	Forecasts of Changes, Root Mean Square Error (M_P) (4)
		Period Covered: 1959-63[b]			
1.	A(M)	14.4	−7.1	16.6	15.8
2.	A(S)	10.4	−8.6	12.4	10.1
3.	A(L)	11.5	−2.9	14.1	14.7
4.	C	14.0	−1.1	15.2	16.3
5.	D	16.6	−10.1	19.2	18.5
6.	E	18.2	−13.4	21.0	20.8
7.	G	15.6	+12.9	17.2	20.2
		Period Covered: 1955-63[c]			
8.	A(M)	14.2	−9.8	17.3	15.0
9.	A(S)	10.6	−8.0	13.2	10.8
10.	A(L)	10.7	−5.1	13.7	12.8

[a]A(M): group mean forecast for set A; A(S): sum-of-sectors forecast for A; A(L): labor-input productivity forecast for A. See text on these models. The rest of the code is familiar from the tables and text above.

[b]All the forecasts refer to the fourth quarters of the year. The number of forecasts was five.

[c]The forecasts for 1955-56 refer to the second half of the year; those for 1957-58, to the year as a whole; and those for 1959-63, to the fourth quarter of the year. The number of forecasts was nine.

apparently avoided much of the underestimation bias of the other forecasts by allowing more strongly and more accurately for the upward trend in GNP.

Thus the sum-of-sector forecasts by the group's experts in the particular fields prove, on the average, superior to the mean of the global forecasts of GNP by all the individuals in the group. It is important to distinguish this situation from the one considered earlier, where total GNP forecasts were compared with sectoral forecasts by the same individuals or groups. There it was recognized that partial cancellation of errors in the aggregation by sectors helped explain the smaller over-all errors in the GNP forecast than in most of the component forecasts. But here we compare two alternative forecasts with the same degree of aggregation, both of which are presumably helped by the offsetting sectoral errors. If A(S) is better than A(M), this suggests that the sectoral forecasts by individual experts contained in A(S) are more accurate than the mean sectoral forecasts of all the group members contained in A(M). Since these same individuals also participate in the A(M) forecasts, one can also say that their predictions in the particular areas assigned to them apparently have been better than their predictions in the other areas.

This result, therefore, suggests that reliance on expertise in particular sectors may yield better forecasts. However, our evidence is scanty and it would be desirable to develop further data bearing on this point.

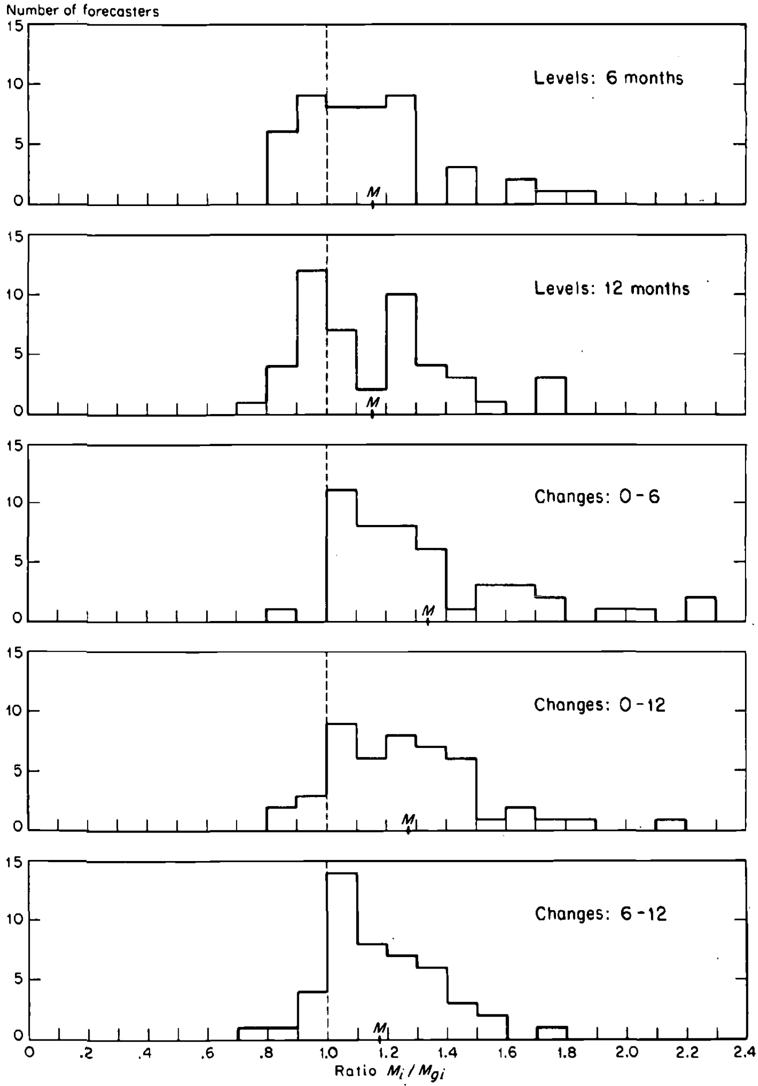
AGGREGATE AND INDIVIDUAL FORECASTS

For one large group of business economists, whose forecasts are designated as set D, an intensive analysis was made of the individual member forecasts as well as the group average forecasts. The errors of the former were compared with the errors of the latter in terms of the ratio of root mean square errors, M_i/M_{gi}. In this expression, M_i refers to the i-th member's forecast, and M_{gi} to the corresponding group mean forecast computed as an average of all the individual forecasts that covered the same periods as those included in M_i. Only those who had made forecasts on at least five occasions in 1955–62 were included. (Altogether, forecasts were collected semiannually at sixteen dates during that period, each time for six and twelve months ahead.)

Chart 7 shows the distributions of these ratios for the level and change forecasts of GNP, classified by span. The graphs illustrate strikingly the superiority of the group mean forecasts over the individual member forecasts. Only a small part of each of the plotted distributions lies to the left of the line of the unit ratios, in the region where $M_i/M_{gi} < 1$. The distributions are skewed to the right, the best ratios

CHART 7

Forty-Seven Individual Forecasts of Levels and Changes of Gross National Product over Spans of Six and Twelve Months (Set D), Comparisons with Mean Forecasts, 1956–63



NOTE: In each panel, the vertical scale represents the number of forecasters (members of group D). The horizontal scale represents the ratios of root mean square errors for the individual forecasts and the corresponding group mean forecasts, M_i/M_{gi} (see text for further explanation).

The broken vertical line, drawn through the points of the unit ratios on the horizontal axes, separates the ratios that indicate the superiority of the individual over the mean forecasts (to the left) from the ratios that indicate the opposite (to the right).

The points labeled "*M*" on the horizontal axes locate the mean M_i/M_{gi} ratio in each case.

falling in the classes between 0.7 and 0.9, the worst ones in those between 1.6 and 2.2. The means of the ratios are identified on the horizontal axes; they vary from 1.14 to 1.37. The results for the level forecasts show that only fifteen or seventeen men, out of the forty-seven included, did better than the group means. For change forecasts, the proportions are still smaller (one to six of the forty-seven members, depending on the span).

It should be clear that all this applies strictly to the summary measures of error over time. In any particular period, many forecasters (probably about half of them) would produce more accurate forecasts than the group mean, but very few do so *consistently*. In fact, those who "beat" the mean forecast for the six-month span are not necessarily the same as those who did so for the twelve-month span (only ten out of the eighteen forecasters with ratios of less than one for either span had ratios of less than one for both spans).

The reason for the superiority of the average forecast is that it is helped by the cancellation of individual errors of opposite sign. It is easy to see that, unless all the errors (for all forecasters in the group and for each period covered) have the same sign, the absolute mean of the errors of the average forecast will be less than the absolute mean of all the individual forecast errors. (In the special—and unlikely—case where all errors have the same sign, the two means will be equal.) This must be so simply because the absolute value of a sum is smaller than (or at most equal to) the sum of absolute values. A similar argument applies to variances and therefore also to the mean square error measures.

The forecasters who did better than the group mean not only underestimated GNP less than others but also were more consistent in the sense of having smaller error variances, as required by the above reasoning. A few performed about as well as the best of the forecasts reviewed in the earlier parts of this report, which indicates the high degree of dispersion that may often be concealed by averages for large groups of forecasters. The accompanying tabulation may serve as an illustration, although the comparisons it provides are inevitably crude: [2]

[2] The numbers of observations per span are as follows: Forecaster one-14; two-15; three-7; mean forecast, group D-16. To compare the results for other forecast sets, see Table 18.

	Six-Month Level Forecasts		*Twelve-Month Level Forecasts*	
	R_1	R_2*	R_1	R_2*
Forecaster one, set D	.515	.900	.407	.924
Forecaster two, set D	.456	.821	.498	1.095
Forecaster three, set D	.421	.775	.475	.860
Mean group forecast D	.549	.983	.527	1.136

As suggested by Chart 7 and the *R* ratios for the mean forecast D, most of the individual forecasts in set D were able to pass the test of N1 but not that of N2*. Computations matching strictly the periods covered by each forecast confirm this inference. Similar results have been obtained for the forecasts of industrial production made by the individual members of group D.

RANKING THE FORECAST SETS

Definite conclusions on the relative performance of different forecasters presuppose a comparability that is seldom found in practice. Some variables are more difficult to predict than others and some periods present greater difficulties than others; hence, strict comparability would require forecasts for the same variable and for the same period. Late forecasts can take advantage of recent information not available for earlier forecasts: hence, the predictions, to be comparable, should also be issued at about the same dates. And there are still other factors that affect the answer to the seemingly simple and always intriguing question "Who forecasts best?"

It is not part of the purpose of this study to determine who the best forecasters are. Nevertheless, the question is inevitably asked and cannot well be ignored. Answering it with fairness is usually difficult and may even often prove impossible, but the process of finding this out can itself be instructive. It is in this spirit that some comparisons are made here between forecasts from different sources. The latter are not identified and this exercise is intended to be primarily of analytical interest.

The first step was to rank the sets of annual forecasts of GNP according to their root mean square errors (*M*), using the maximum number of comparisons that could be made over identical periods of time. Difficulties are at once encountered, since the rankings are not the same in the different periods covered, and the period common to all the forecasts is short. However, the rankings do show some consistency, at least at the extremes. Thus, sets G and F, which cover all the years of comparisons (1953–63), have the lowest M values in all subperiods in which they were matched with the other forecasts.[3] Set E has the lowest rank in each case, and set D the second lowest on most occasions. The range of these summary error measures is very large, e.g., in 1953–63, the smallest M equalled \$8.2 billion and the largest \$17.2 billion.

The intermediate ranks (3 to 6) are occupied by sets B, A, H, and C. Here there is little consistency, e.g., set A ranks lower than set B in 1954–63 but higher in 1956–63 and 1958–63. Some of the observed differences between the M values that underlie these ranks are small and may not be significant.

The ranks for the errors of level and change forecasts show substantial correlation (Spearman's correlation coefficient for the average ranks is .88). But there are some apparently systematic differences, notably set F is better than set G for changes, while the reverse is true for the levels.

Ranking the forecasts according to any of the M-ratios based on the extrapolative benchmark models (R_1, R_2, etc.) must, of course, yield the same results as ranking according to the M_P values proper, if the periods covered are the same.[4] If the assumption that an extrapolative model provides equally good predictions of a given variable in different periods were warranted, the use of the model should enable one to make fair comparisons between forecasts of that variable made by different persons for different time periods. Unfortunately, the assumption is not a safe one. An extrapolative technique may work significantly better in some periods than in others because of differences in the behavior of the given series in the time intervals concerned. Some such differences exist even among the periods covered by our present

[3] All eight sets were ranked for 1958–63, all but one (C) for 1956–63, and all but two (C and D) for 1954–63. Rankings of the sets G, F, B, and E were also obtained for 1953–63. See Table 16 for the longest periods used in these computations and the corresponding M values.

[4] The ratios M_P/M_N have a common denominator for all the forecasts in a given period (it is a certain specific value of M_{N1} for the ratios R_1, of M_{N2} for the ratios R_2, etc.). A given extrapolative model provides a specific standard of performance with which to compare the various forecasts proper; such a standard is clearly not needed to appraise the performance of the forecasts relative to each other, where they all refer to the same series.

investigation, though they all relate to one short phase of recent economic history and partially overlap.

The individual company forecasts (G, F, B) come out well in the comparisons for GNP. The large group forecasts D and E trail, despite the advantages of aggregation described in the previous section. (But a few members of group D did about as well as the best item on our list, and the same may apply to group E, though it could not be tested for lack of individual data.) It is tempting to infer that predictions made by individuals or small teams are, in general, better than those by a large group of forecasters. Rationalizations for such an inference are easy to find. First, the errors of an individual are visible, at least potentially, since his forecast is issued separately; whereas those of a group are not, or hardly at all, since the forecast is submerged in an anonymous poll. As a result, individual forecasts are likely to be more disciplined than those contributed to a poll. Second, some participants in group forecasting, particularly in large polls, may simply be expressing relatively uninformed opinions or guesses; the average level of competence in a large group may be quite low.[5]

Comparisons of industrial production forecasts are inconclusive because the summary measures of error show small differences. Thus the six *M* values for 1954–63 fall into the range of 4.9 to 6.0 index points; six of the seven for 1958–63, in the range of 5.1 to 5.9. Also, there is still less rank consistency here than in the case of GNP. However, the large-poll median forecast E again ranks last in 1958–63 and next to last in 1954–63. Forecast D, produced by a smaller and more select group, is also relatively weak (it ranks fourth in 1954–63 and sixth, i.e., next to last, in 1958–63). All this is at least not inconsistent with the preceding argument.

There is one mitigating factor, however, in the case against these group forecasts—one that qualifies all the above comparisons—and this is the time when the forecasts are made. In our collection of year-end forecasts, D, E, and B were made in October; A and (sometimes) C in November–December; F, G, and often C in January; and H includes predictions issued at various dates scattered between October and January.

[5] This position was taken by Gordon McKinley in his discussion of my preliminary report on the NBER forecasting study at the 1964 Annual Meeting of the American Statistical Association in Chicago.

Since 1950, the first estimates of GNP for a given quarter are provided by the Council of Economic Advisers one month after the close of the quarter. Thus, early in October all that a forecaster may know is the latest GNP figure recorded for the second quarter; in November or December, his knowledge is likely to extend to the estimate for the third quarter; and late in January he may have the first CEA estimates for the fourth quarter. Not only that, but other important data that are demonstrably helpful in forecasting, such as the McGraw-Hill Survey of Investment Intentions, are not available before November.[6] It is clear, therefore, that the relative earliness of sets D and E and the associated deficit of informational inputs may have much to do with the low ranks of these forecasts; though there is reason to believe that this is not the entire explanation and that the other explanations advanced above may still be valid.[7]

Also, for industrial production forecasts lateness is probably of considerably less help than for GNP forecasts, since the production data are available monthly. (The reduced importance of the variation in forecast dates may be one reason why the M values differ so much less for industrial production than for GNP.) But here, too, the ranks of sets D and E are low, while those of sets F and G are high.

In addition to their late timing, the GNP forecasts in set G have one advantage of a special kind. They were given originally in base-year prices (with several shifts of the base). For comparisons with other forecasts, we had to convert them to current dollars; but in so doing we imputed perfect price predictions. This is probably responsible in some part for the high rank of G, since the other forecasts (made in current dollars) undoubtedly include a component of error due to imperfect price predictions.[8]

[6] On the predictive value of investment anticipations, see above pp. 96 to 97.

[7] Thus, for GNP, forecasts D and E are also markedly inferior to B, which was also quite early. (It was usually sent to the printer during the last week of October.)

[8] It should be noted that set G has lower R ratios (i.e., comes out better relative to the benchmark models) when both forecasts and extrapolations are expressed in current dollars than when both are given in the base-year prices, though the differences are not large. This is consistent with the argument in the text.

However, errors of price-change predictions could conceivably be such as to reduce, rather than add to, the total errors. This would be so in the case of offsetting errors, when, e.g., the price-change errors were positive and the errors in predicting the change in constant dollar values were negative. This shows that the situation envisaged in the text, where the advantage is on the side of the constant dollar forecasts, need not always apply, though it is presumably more likely than others.

It should be clear by now that, because of their typical diversity in several aspects, it is extremely difficult to give the forecasts any meaningful over-all grades. Yet the catalogue of problems is still longer. Our evidence, although wide as forecast studies go, covers short periods; hence the samples are still small. To extend the evidence, it is desirable to add the comparisons of chain forecasts to those of annual forecasts, but this entails further complications. The multiperiod forecasts are even less comparable than the annual ones, since they are issued at different times and with different frequencies. Only partial adjustments for these divergencies can be made, and they considerably reduce the number of observations available for these comparisons. Finally, the results based on chain forecasts do not necessarily agree with those obtained for the annual data.[9] One probable source of such differences lies in the role of turning-point errors, which is greater for the chain forecasts than for the annual forecasts. It is not difficult to see that the turning-point criterion may yield different results from the average-errors criterion. While the latter has the advantage of using more information, the former places a greater premium on genuine contributions of forecasters as distinguished from mere extrapolative techniques. Efficient trend projections can produce relatively small average errors for series such as GNP, but not a good turning-point record.

Moreover, differences among averages that cover periods of several years have little meaning if the variation of the underlying values from year to year is very great. It is, therefore, desirable to examine the relative positions of forecasters according to accuracy in each successive year. This will be done presently with a view to establishing how stable or variable these positions are.

CONSISTENCY OF FORECASTERS' PERFORMANCE

A measure of the degree of over-all agreement among several rankings is provided by the coefficient of concordance W. In the case of perfect consistency, each forecast set would preserve the same rank throughout, which yields $W = 1$. In the case of least (zero) consistency, the sum of ranks earned in all periods covered would be the same for each forecast set, namely, equal to the mean value of such sums; here the rankings

[9] For example, the former are much less favorable to set G than the latter are, as suggested by Tables 16 and 18.

are random or uncorrelated, and $W = 0$. The coefficient of concordance is closely associated with the average of all the rank correlation coefficients that can be computed from the same ordered data.[10]

In Table 1, six forecast sets (all except C and D) cover the ten years 1954–63; we can, therefore, construct for them ten rankings according to the absolute values of their errors in each year as listed therein. For these data, the value of W is 0.123 and the average rank correlation coefficient is 0.138. According to an approximate test based on Fisher's Z-distribution, these values are not significant at the 5 per cent level.[11] Even lower W coefficients were obtained for another combination of forecasts and periods ($W = .076$ for the eight forecast sets in 1958–63). And the results for industrial production point in the same direction; here $W = .031$ for six sets of predictions of annual changes for 1954–63.

When the individual ranks for each set are averaged over the years 1954–63, the results do agree broadly with the rankings according to the M_P values, which were discussed in the preceding section: for example, the individual company forecasts are again found at the top of the list for GNP. But it is clear that the averages conceal a great deal of variation in the ranks: the relative positions of the forecasters according to accuracy show many shifts from one year to another. This puts in doubt the evidence of the averages. The less stable the individual ranks, the less meaningful it is to compare the average ranks; and if the former show no significant consistency, then the significance of differences among the latter is also questionable.

The heterogeneity of the forecast sets with respect to the degree of aggregation and dates of issue is troublesome in this context as in others. Some work has been done with the individual forecasts by

[10] Let m be the number of rankings and n that of items that are being ranked (in the present case, m would be the number of periods and n the number of forecast sets). The grand total of ranks, divided by n, gives the mean value of the sum of ranks per item, $\frac{1}{2}\, m\, (n + 1)$. Let S be the sum of squared deviations of the observed sums of ranks from that mean value. Then $W = \frac{12S}{m^2(n^3 - n)}$. The mean of the corresponding rank correlation coefficients (Spearman) can be written as $\frac{mW - 1}{m - 1}$. On the theory underlying the measurement of concordance, see Maurice G. Kendall, *Rank Correlation Methods*, London, 1948, Chapter 6.

[11] This is based on a table given in Milton Friedman, "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," *Annuals of Mathematical Statistics*, March 1940, pp. 86 ff. (reprinted in Kendall, *Rank Correlation Methods*, Appendix Table 6).

members of group D, which has the advantage of avoiding these problems. For GNP, nine members made forecasts on all occasions, that is, sixteen times at regular semiannual intervals in 1955–62. An analysis of the ranks of their twelve-month forecasts yields $W = 0.111$. This again is not significant at the 5 per cent level. For industrial production, ten members have made forecasts on all occasions in 1958–62, that is, nine times at semiannual intervals. In this case, the coefficient W for twelve-month forecasts is 0.106, which is likewise not significant at the conventional levels.[12]

According to the concordance analysis, then, the variation among the ranks of errors appears to be governed largely by chance. Nevertheless, an inspection of the ranks suggests that some of their shifts can be related to such "systematic" factors as the forecaster's bias and the variation in size of the observed changes in the predicted series. Thus one of the forecast sets ranked first (had the smallest errors) in years of high rates of economic growth—1955, 1959, and 1963—but ranked last in years in which the advance slackened—1960 and 1962. This can be linked with the fact that this forecast set exhibited a strong tendency to overestimate GNP increases in recent years (see Table 2, line 9). On the other hand, one of the sets that had a strong underestimation bias ranked at or near the bottom in most years but was first in 1954, 1960, and 1962, years that witnessed recessions or temporary slowdowns in growth (see line 7 in Tables 1 and 2).

SIZE OF PREDICTED CHANGES AND FORECASTING ERRORS

The next step, prompted by the above observations, was to rank the forecasters according to the absolute size of the changes they predicted for each period and to apply concordance analysis to these ranks. The forecasts covered are the same as those in the preceding analysis of the ranks of errors.

The coefficients of concordance W are in each case much larger for the ranks of predicted changes than for those of errors. For the six sets of GNP forecasts covering the years 1954–63, $W = 0.403$; for the

[12] The table mentioned in footnote 11 is limited to certain low values of m and n. For larger values, the F-test may be used, where $F = \frac{(m-1)W}{1-W}$ with degrees of freedom $n_1 = n - 1 - \frac{2}{m}$ and $n_2 = (m-1)\,n_1$. See Helen M. Walker and Joseph Lev, *Statistical Inference*, New York, 1953, p. 285.

TABLE 23

Relation Between Predicted Changes and Errors: Selected Individual Forecasters of GNP (1955-63) and Industrial Production (1958-63)

Forecast Period[a]	Gross National Product		Industrial Production	
	ρ^b	Recorded Change[c]	ρ^b	Recorded Change[d]
6/55 - 6/56	−0.73	23.4		
12/55 - 12/56	−0.63	19.4		
6/56 - 6/57	−0.88	23.7		
12/56 - 12/57	−0.15	14.9		
6/57 - 6/58[e]	+1.00	−11.4		
12/57 - 12/58[e]	+0.63	4.2		
6/58 - 6/59	−0.97	44.6		
10/58 - 10/59	0	30.4	−0.52	12.00
3/59 - 3/60	+0.35	22.7	−0.43	8.00
10/59 - 10/60[e]	+1.00	19.6	+0.99	2.00
3/60 - 3/61[e]	+1.00	4.4	+0.87	−1.98
10/60 - 10/61	−0.39	29.2	−0.29	13.72
4/61 - 4/62	−0.70	41.5	+0.60	15.24
10/61 - 10/62[f]	+0.48	29.0	+1.00	7.62
4/62 - 4/63[f]	−0.95	27.3	−0.86	8.38
10/62 - 10/63	−0.88	33.4	+0.22	10.97

Source: Forecasts of nine individual members of group D for GNP and forecasts of ten individual members of group D for industrial production (see text).

[a]The span of each forecast is twelve months. The first date shows when forecasts were made.

[b]Correlation between ranks based on absolute values of predicted changes and ranks based on absolute values of errors (see text).

[c]Early estimate of the actual change in GNP in billion dollars.

[d]Early estimate of the actual change in industrial production in index points, 1947-49 = 100.

[e]These periods include several months of business recession.

[f]These periods include a few months of business retardation in the second half of 1962.

eight sets that are available for 1955–63, $W = 0.442$. The six sets of industrial production forecasts relating to 1954–63 yield $W = 0.433$. All these coefficients are significant at the 1 per cent level.[18]

The data for the individual forecasters in group D produce the following values of W: for GNP, 0.383; for industrial production, 0.200. The former figure is highly significant at the 1 per cent level, the latter is just significant at the 5 per cent level.

Forecasters who predicted relatively large changes in most years, as indicated by the averages of these ranks, were more accurate than those who predicted small changes. The average ranks based on the size of anticipated changes show significant negative correlations with the ranks determined according to the over-all errors (M) of the same forecast sets. As the values were all ordered in the same direction (from the lowest to the highest), a negative coefficient means that forecasts of relatively small changes were associated, on the average, with relatively large errors. For the six sets of GNP forecasts, 1954–63, the Spearman rank correlation coefficient (adjusted for ties) is -0.657; for the industrial production forecasts in the same period, it is -0.486.

However, in times of declines or slowdowns in economic activity, the "timid" forecasters who predicted relatively small changes would come out ahead of those who foresaw larger changes (typically, increases). In such periods, therefore, the correlations between the predicted changes and the errors are likely to be positive. For the individual forecasters in group D, rank correlations between changes and errors were computed for each of the overlapping twelve-month periods covered (Table 23). It will be noted that the positive correlations are indeed, with few exceptions, related to periods of business contraction or retardation.

[18] This is strongly indicated by the F-tests defined in footnote 12. (Corrections for continuity have been applied, without affecting the results materially.)