The article illustrates well an ever-increasing tendency in published work on psychiatric subjects to substitute for clinical observation and description a pseudo-scientific mathematical or statistical approach. I suppose it all began with the idea that "intelligence" can be accurately measured and expressed in mathematical terms.

W. J. STANLEY.

15 St. John Street,
Manchester.

DEAR SIR,

Professor Oswald's first point concerns our method of data collection. He complains that the data were "subjectively-determined scores" and that "an element of indecision" was always present. If this is meant as a criticism, it is a criticism that embraces nearly all attempts at ordinary clinical diagnosis in psychiatry, which is inevitably based on subjectively-determined (as opposed to physically measurable) data. This does not at all invalidate the use of statistical methods, provided that the errors of judgment are randomly distributed. It would indeed be unfortunate, and erroneous, if the "indecisiveness" of clinical judgment came to be regarded as *ipso facto* invalidating the use of statistical methods in psychiatry, since the opposite is true. Statistics can be especially useful under just these circumstances, precisely because unreliability always *reduces* correlations and thus statistically significant conclusions cannot be produced by unreliability of data. Any correlations found are despite and not because of the unreliability of the individual clinical ratings. This is also the answer to one of Dr. Stanley's criticisms.

Professor Oswald's next point concerns *bias*, which is not the same as unreliability. He suggests that bias may have been operating because (i) the data could have been recorded in such a way as to fit in with our preconceived notions of neurotic and endogenous types of depression; and (ii) the recorded state of follow-up could have been influenced by knowledge of the previous history.

With regard to (i), it is really very difficult to understand just how this could have brought about the results obtained. Of course we believed that neurotic and endogenous depression were distinct entities, since previous work in the department had led us to this belief. The 35 items selected for analysis were indeed chosen precisely because of their supposed discriminating function. Yet the factor loadings on the bipolar factor did not fall into two discrete groups; on the contrary, there were nine items in the intermediate range, with loadings less than $\pm 0 \cdot 300$. Among these was the item "worse in morning", which was certainly expected to have a high correlation with diagnosis, but in fact did not (correlation $+0 \cdot 143$). One wonders just at what point the rater's bias began to operate.

Secondly, the correlation of each item with diagnosis agreed remarkably well with its factor loading on the bipolar factor. To arrive at such a close fit by bias alone would have demanded extraordinary mathematical insight on the part of the rater; the data would have to have been recorded in such a way that the 595 correlations between the 35 features produced loadings proportional to the correlations with diagnosis. Professor Oswald suggests that the fit was obtained by "splendid statistics-manship". In fact, the loadings were derived directly from the computer without rotation or other manipulation.

Professor Oswald finds a second source of bias in the follow-up ratings. He says: "The response-to-E.C.T. argument must be rejected since the 'response' was actually a score subjectively determined by authors who, at the time, knew the history and also whether E.C.T. had been given." Actually, *all* patients had been given E.C.T., and this knowledge could not have influenced the rating. Moreover, the rater made his follow-up assessments without reference to the original diagnoses. To the extent that he might have remembered some of the patients, it would have been methodologically desirable to have employed an independent rater; considerations of manpower made this impracticable.

Professor Oswald suggests that had an assessor from another school been employed, and the same results obtained, he would have been convinced. But would he? He could argue that unconscious biasses operated, that the assessor changed his allegiance during the course of the investigation, and so on. The appropriate action on his part would be to carry out a similar investigation himself or to await the publication of the results of other investigators (for example, Sandifer, Wilson and Green, *Amer. J. Psychiat.*, in press). Additional studies are also under way at the Massachusetts Mental Health Center by Rosenthal and Klerman, and in the National Institute of Mental Health Psychopharmacology Service Center Collaborative Depression Project. We would, however, stress the need for careful definition of the prognostic features under investigation, and also suggest that the use of standardized, structured interviewing techniques should be further explored.

Dr. Stanley complains that the assessments were too rough to be treated by statistical methods. As already stated, any errors in assessment may be expected to balance out, provided that they are

random. The number of decimal places worked out is immaterial; it is not the minutiae but the general pattern that is important. Correlations are given to three places, partly for conventional reasons and to allow others to work on the raw data if they so wish.

We do not agree that many of the features cannot be said to be either present or absent; one might as well say that nobody is entirely ill or entirely well. This may be literally true, but it is permissible and necessary to define "patients" according to convenient, if arbitrary, conventions. In the same way we used criteria for deciding whether a feature was to be regarded as present or absent. We would agree with Dr. Stanley that this aspect is important; and in due course we expect to be able to publish our criteria in greater detail. Knowledge, however, does not spring out fully armed, like Athene from the head of Zeus, and this is a report of work in progress. Incidentally, our method is founded on, and not a substitute for, clinical observation.

MARTIN ROTH.
M. W. P. CARNEY.
R. F. GARSIDE.

Department of Psychological Medicine,
Queen Victoria Road, Newcastle upon Tyne.

PATTERNS IN REACTIVE AND ENDOGENOUS DEPRESSIONS

DEAR SIR,

Dr. Foulds (*Journal*, November 1965) suggests that a psychiatrist who has the impression that a patient is suffering from a reactive depression would not ask about sleep disturbance with the same persistence as he would if he thought the illness was an endogenous depression. This is not so, because most British psychiatrists consider sleep disturbance to be an important differentiating symptom. Since the introduction of antidepressant drugs there has been a tendency to over-diagnose endogenous depression because of the supposed effectiveness of these drugs. In order to support the diagnosis of endogenous depression the average psychiatrist is likely to look carefully for sleep disturbance.

Dr. Foulds makes the erroneous assumption that reactive and endogenous depressions are equivalent to his neurotic and psychotic depressions. He regards delusion as an essential feature of psychotic depression. There are many patients with endogenous depression who are not deluded and will therefore be classified by Foulds as neurotic depressives.

FRANK FISH.

Department of Psychological Medicine,
The University of Liverpool.

EFFECT OF A DEPRESSIVE ILLNESS ON M.P.I. SCORES

DEAR SIR,

In their recent paper, Coppen and Metcalfe (*Journal*, March 1965) appear to make an important methodological point about test-retest reliability studies in general and those relating to the Maudsley Personality Inventory (M.P.I.) in particular. They say: "The stability of a test is often expressed in terms of the test-retest correlation coefficient, but our results show that this can be very misleading; groups of patients can evidence a considerable change in their scores even though the test-retest correlation remains high" (p. 238). Since their study differs in a number of important ways from a true test-retest reliability study, I question the validity of their discussion on this particular point.

In the first place, the test-retest correlation could be perfect, not merely "high", and yet the mean differences could still be as large as Coppen and Metcalfe report. There is no necessary relationship between the size of the mean difference and the correlation between the scores. Secondly, their data has only an indirect link with a true test-retest reliability study. Their experimental design specifically required that treatments be interpolated between the first and second testings. A control group, not undergoing any special treatment, would properly estimate repeat-test reliability over the same period of time. If a different value for this correlation coefficient were found in the experimental group, it would suggest that the treatment had had a differential effect on patients having different initial scores. Test-retest reliability is test-retest reliability and not just *any* correlation between repeated measurements.

Thirdly, the correlation coefficients quoted in their paper may not provide appropriate summaries of the data. It is apparent from Tables I and II that there are marked heterogeneities in their group of patients. Not only do the two treatment groups (E.C.T. and Drugs) give different mean scores on the M.P.I., but so do the three diagnostic groups. These differences and possible interactions between diagnostic group and the type of treatment could well invalidate all the correlation coefficients they compute. Far from being surprised how high or how low the correlations proved to be, they should regard it as remarkable that there is any correspondence at all with the data collected by others.

Finally, it should be clear that a test such as the M.P.I. should have two properties: there should be a relatively high stability—high repeat-test reliability and stable means—when no particular change is induced; and it should be sensitive to change when