Appendix 2: Method

Two methods, measurement of outcome and improvement, were used to compare the effects of the four different regimes, both for junior and senior children. This appendix gives details of these two methods, as briefly outlined in Chapter 3.

Clinical measures of outcome

Sainsbury argued that:

> 'Four distinct post-treatment (or care) categories can now be recognized: (1) *clinical status*, or how well the patient is after treatment; (2) *improvement* (difference), or how much better the patient is after treatment; (3) *base-free* improvement (ratio or covariance), or how much better the patient is by comparison with others who started at the same level; (4) *outcome*, or how well the patient is before and after treatment.'
>
> (Sainsbury 1975:143)

Outcome was measured by the procedure given by Sainsbury (1975). As he pointed out, the problem associated with outcome 'stems from the situation, often crucial in clinical studies, in which patients at the top (or bottom) of the scale have no room to improve (or worsen)' (Sainsbury 1975:143). He suggested the following formula to solve this problem:

$$0 = M_1 + M_2 + 2(M_2 - M_1)$$

where 0 = outcome; M_1 = initial score; M_2 = final score

Sainsbury did not indicate how this formula was derived, but it simplifies to the following:

$$0 = 3M_2 - M_1$$

Thus, in measuring outcome by Sainsbury's method, the initial score is not merely subtracted from the final score, but a differential weighting of three to one is introduced. Therefore, using this measure of outcome is much the same as carrying out an analysis of covariance when the regression co-efficient of final upon initial scores is one-third. The correlation between initial and final scores, using reliable measures (as we have done), is not usually below two-thirds and therefore, assuming standard deviation of initial and final scores to be equal, the regression co-efficient of final upon initial scores will usually be about two-thirds or more. Thus, Sainsbury's outcome method places about twice as much relative weight upon the final score as does analysis of covariance. This means that outcome is not merely a measure of change (adjusted for initial score difference): it depends both upon change and upon final state. It is, therefore, a realistic measure of clinical outcome which should take actual clinical states after treatment into account, in addition to any change which may have been brought about.

In using this method, the children's behaviour was rated by a child psychiatrist on three occasions: at base, at midline assessment (about eighteen months later) and at final follow-up (eighteen months after the midline follow-up). On each occasion the rating psychiatrist used all the available information gathered about the child. This information did not always include a complete set of data and, therefore, the numbers of children about whom outcome was available (see *Tables A3(2)* and *A3(3)* in Appendix 3) was a little greater than the numbers who had a complete set of improvement data.

The psychiatrists did not attempt to assess improvement, but rated the children on three distinct occasions. Their ratings were therefore more reliable than if they had rated improvement, which is a more difficult task. Each child was rated on a four-point scale at the midline and final follow-ups: (a) no disturbance; (b) slightly disturbed; (c) moderately disturbed; (d) markedly disturbed. The same scale was used at base, except that the 'no disturbance' rating was, of course, rarely used, and then only for junior children who were 'at risk'. The range of outcome at each follow-up therefore was minus 1 to 11. This range was then divided into three categories, corresponding to good outcome (minus 1 to 3), moderate outcome (4 and 5), and poor outcome (6 to 11). Percentages of children corresponding to these three levels of outcome were then calculated; it is these percentages that are reported in *Tables A3(2)* and *A3(3)* in Appendix 3. In the Figures in the book itself the percentages of the moderate group have been omitted, in order to give a clearer picture.

Statistical measures of improvement – covariance analysis

To compare regimes in relation to improvement we used analysis of covariance in preference to analysis of variance. Covariance analysis takes into account differences between the regimes that may affect improvement. For example, initial level of maladjustment (i.e. before treatment) is inevitably correlated with improvement and, therefore, if regimes differ initially, this should be taken into account when comparing the mean improvement of regimes. Analysis of covariance allows this to be done. In fact, the initial levels of some measures differed significantly between regimes (see Appendix 3); the use of covariance analysis was therefore justified.

As well as taking initial levels into account, we allowed for initial differences between the regimes in relation to general severity of maladjustment, non-verbal IQ, an index of social functioning within the family, and, for the junior children only, an index of family history of psychiatric illness. We did not allow for this last variable in the case of the senior children because of lack of scatter in some regimes. These particular 'covariates' were taken into account because we found that they were relatively important.

Provided that children are randomly allocated to regimes (as was the case in the present study), both analyses of covariance and variance are valid methods of testing the statistical significance of differences between regimes. However, analysis of covariance is the more sensitive procedure and, more important, it provides a greater degree of accuracy in the assessment of comparative improvement means than does analysis of variance.

Errors of measurement, as is well known, reduce the corrections to the adjusted means. Our covariates, though, were reliable, and thus under-correction was slight.

ASSUMPTIONS UNDERLYING COVARIANCE ANALYSIS

The assumptions underlying the analysis of variance – that variances are homogeneous and that distributions are normal – also apply to analysis of covariance. However, evidence from analysis of variance indicates that the analysis of covariance is robust with respect to the violation of these two assumptions. Thus, only one variable was excluded from analysis because of violation of either of these assumptions; this was the sociometric measure of rejection, which had a J-shaped distribution of scores.

Analysis of covariance also involves the further assumption that within-group regressions are homogeneous, that is, that the regression co-efficients of the variate (improvement in the present case)

upon the covariates (for example, initial score) are the same for each regime. Although there is evidence that analysis of covariance is robust with respect to this assumption (Winer 1971:772) it was thought prudent to examine this, and a test of homogeneity of regression was carried out for each variable across regimes. As a result of this a few measures were dropped from the analysis of covariance (see Appendix 3). On the remaining measures (the great majority) analysis of covariance was carried out using improvement (initial minus final score of maladjustment) as the variate, with five covariates for the junior children and four for the senior children. In general, it was found that the initial score was the most important covariate in our analyses.

INTERACTION EFFECTS

Had the differences between regimes in relation to improvement been affected by the diagnostic category of the children or by their sex, that is, had there been interaction between regime and diagnostic categories, or between regime and sex, it would have been desirable to carry out comparisons between regimes for the two diagnostic categories and for boys and girls separately. Therefore, to check this possibility, two two-way analyses of covariance (using unweighted means) were carried out first, one with regime against diagnosis and the other with regime against sex, for both senior and junior groups.

In the case of seniors only in five out of ninety-two measures (including both comparisons at midline and final follow-up) was interaction with diagnosis found to be significant. Only four interactions with sex were found to be significant. Thus, for both diagnosis and sex, significant interactions with regime occurred no more than might be expected by chance.

For juniors in only one out of sixty-six measures was there an interaction between sex and regime, and in two measures an interaction between regime and diagnosis. Thus, for juniors also there was no evidence of interaction of regime with either sex or diagnosis.

Therefore, for both seniors and juniors, our hypotheses 3 and 4 (that regimes interact with diagnostic category and with sex) were not confirmed and, moreover, it was legitimate to compare regimes (hypotheses 1 and 2) without regard to diagnosis or sex. It was also possible to compare diagnostic categories, and boys with girls, independently of regime (hypotheses 6 and 7).

Did the regimes differ from each other?

Those measures for which analysis of covariance indicated that regimes differed significantly (hypothesis 1) were investigated further to ascertain which regimes differed from which; that is, pairs of regimes (including the controls), were compared by analysis of covariance. However, when there are four regimes and their change means are arranged in order, the probability that the largest and smallest means will differ is exaggerated. This had to be taken into account when comparing pairs of means. Thus, in making pair-wise comparisons in relation to our first hypothesis that the four regimes differ, we used the Newman-Keuls technique. This is fully described in the literature (for example, Winer 1971:191 *et seq.*). The technique allows for the fact that, where more than two means are arranged in order of magnitude, the probability of two means differing is altered by putting them in order. In using this technique, we had, of course, to amend the appropriate published tables of significance levels to take into account the fact that our results were in terms of the F ratio.

Which regimes were more effective than the controls?

Hypothesis 2, which postulated that one or more of the three treatment regimes is more effective than no treatment at all, raised quite a different point from hypothesis 1, and had to be dealt with separately. We wished to compare *three* treatments with a control regime, and, accordingly, used the procedure outlined by Dunnett (1955, 1964), which takes into account the fact that the number of treatment regimes is greater than one; it is more likely that one treatment in a group of three treatments would be significantly different from a control regime, than a single treatment only. Even though hypothesis 2 was written as one-tailed (Chapter 3), we have always used a two-tailed ($p < .05$) level of significance as this is a more stringent procedure.

If the improvement mean of any treatment regime was statistically significantly better than the corresponding control mean using either the Newman-Keuls test (hypothesis 1) or Dunnett's procedure (hypothesis 2), then we regarded that treatment as superior to no treatment (in relation to the measure in question) and have reported it as such in the text.

Grouping of measures to make up aggregate measure

Principal component factor analyses were carried out to ascertain which measures should be summated. For both juniors and seniors an important general component was found underlying the fourteen Devereux items. For the juniors this accounted for 41 per cent of the variance, and for the seniors 48 per cent, having reversed items 7 and 10, i.e. comprehension and creative initiative, which are clearly both positive attributes. Thus, for both juniors and seniors it was decided to add the standardized items (reversing 7 and 10) to obtain an aggregate global Devereux score.

A further principal component analysis was carried out on the ten remaining junior measures and, again, a general component was obtained, accounting for 26 per cent of the variance, plus a bipolar component, accounting for 16 per cent of the variance. This bipolar component clearly contrasted the five neurotic measures (neurotic and psychosomatic behaviour derived from parent interview; isolation score derived from sociometry; and neurotic behaviour from the Rutter parent and teacher scales) with the five antisocial items (antisocial behaviour; activity and mood derived from parent interview; and antisocial behaviour derived from the Rutter parent and teacher scales). It was therefore decided to add the five neurotic items, the five antisocial items, and then all ten items together to generate three aggregate measures – neurotic behaviour, antisocial behaviour, and global behaviour. All items were standardized before being added.

The Barker Lunn Attitude Scale, which was used for the seniors, was next examined. It was found that although the first component was a general one, the second component (which accounted for 16 per cent of the variance) clearly distinguished the first seven items from the last three. (This agreed with Barker Lunn's (1969) own analysis.) Because of this differentiation it was decided to sum the first seven items to form an aggregate measure of attitude to school, and also to add the last three items, which measured what we have described as neurotic anxiety in relation to school and schooling.

A principal component analysis was then carried out on the remaining fourteen scores of the senior children. This excluded both the Devereux and Barker Lunn scales, which had already been analysed. Again, a general component was obtained (24 per cent of variance) and a bipolar component (15 per cent of variance). The latter contrasted nine neurotic measures (neurotic behaviour, psychosomatic behaviour, somatic disturbance, and withdrawal from the parental interview; neurotic behaviour from the Rutter parent and

teacher scales; neuroticism and introversion from the JEPI; and the isolation score from sociometry) with five antisocial measures (antisocial behaviour, activity, and mood, from the parent interview, and antisocial behaviour from both the Rutter scales). Thus, it was decided to add the standardized scores of the neurotic measures, the scores of the antisocial measures, and the scores of all fourteen measures to generate three aggregate scales: neurotic behaviour (nine measures), antisocial behaviour (five measures), and global behaviour (fourteen measures). The data relating to this principal component factor analysis are shown in *Table A2(1)*.

Patterns of behaviour and change

Hypothesis 5 stated that regimes differ in effectiveness according to patterns of behaviour, that is, that some regimes are more effective in improving neurotic behaviour, and others more effective in reducing antisocial behaviour. This hypothesis was tested by seeing whether there was any interaction between two different measures of maladjustment (i.e. the aggregate measures of neurotic behaviour and antisocial behaviour) and the four regimes. We did this using the procedure outlined by Greenhouse and Geisser (1959). However, we wanted to carry out analyses of covariance, and not merely analyses

Table A2(1) *Principal component analysis: seniors*

measure	*factor loadings*	
	general	*bipolar*
parent interview data		
neurotic behaviour	0.59	− 0.54
antisocial behaviour	0.70	0.48
psychosomatic behaviour	0.71	− 0.09
somatic disturbance	0.60	− 0.31
withdrawal	0.32	− 0.47
activity	0.65	0.39
mood	0.61	0.09
neurotic behaviour at home (Rutter A)	0.51	− 0.37
antisocial behaviour at home (Rutter A)	0.62	0.49
neurotic behaviour at school (Rutter B2)	0.11	− 0.02
antisocial behaviour at school (Rutter B2)	0.07	0.69
isolation (sociometry)	0.29	− 0.03
neuroticism (JEPI)	0.02	− 0.41
introversion (JEPI)	0.18	− 0.31
percentage of variance	24	15

Note: JEPI = Junior Eysenck Personality Inventory.

of variance as considered by Greenhouse and Geisser. Thus, we used a repeated-measures analysis of covariance method as described by Winer (1971: section 10.6). We were not, of course, using repeated measures – our measures were merely correlated. This necessitated standardizing our two measures so that the question of interaction was meaningful (see Greenhouse and Geisser 1959).

In view of the fact that the numbers in the four regimes were not the same (which does not invalidate the procedure), one of us (Ian Muir Leitch) had to write a special computer programme. The standard programmes available to us all assumed that the numbers in the groups were equal. Because of the complexity of the situation, we used only one covariant, i.e. the appropriate initial score for each measure. For both junior and senior groups we considered two measures, namely the neurotic and antisocial behaviour aggregate measures just described.