RECENT INVESTIGATIONS OF INTELLIGENCE AND ITS MEASUREMENT*

By PHILIP E. VERNON, M.A., Ph.D.

EARLY eugenic investigations of mental qualities were based, for the most part, on the investigators' own assessments of intellectual eminence or special talent, social worth or depravity, and the like. Obviously the subjectivity of these judgments was open to criticism, however fair-minded the investigator. Hence the invention of objective tests of intelligence was received with enthusiasm. They appeared to offer scope for far more scientific studies of inheritance. Thus the *Eugenics Society* has sponsored, or has undertaken the publication of, numerous and outstanding researches by psychologists such as Burt, Thomson, R. B. Cattell, Fraser Roberts and Terman, which were based on intelligence tests. It was of course realized that tests either of the Binet type or the group verbal type (e.g. Otis) were not pure measures of innate ability. Early studies like those of Gordon with canal-boat children and Freeman with American foster-children revealed a distinct influence of education and cultural background. This was confirmed by Newman, Freeman and Holzinger's work on identical twins brought up apart. However, analyses of foster-children and twin data by Burks and others indicated that environmental effects were unlikely to alter an Intelligence Quotient by more than ten points up or down ; or, more precisely, that the contribution of heredity is three or four times as great as that of environment in determining individual differences in intelligence. Thus most psychologists, in this country at least, who were interested in eugenic problems were satisfied that ordinary intelligence tests were adequate tools for genetic investigations. Moreover, when Cattell (7) employed non-verbal group tests, based on pictorial or diagrammatic material, which would presumably be less susceptible to educational influence, he obtained much the same negative correlation with family size as did workers with verbal tests.

Nevertheless, recent investigations by myself and others have forced me to the conclusion that, while intelligence tests are admirable instruments for practical purposes such as educational and occupational selection and guidance within any one cultural group, they cannot be regarded as sufficiently pure measures of innate ability to be employed in comparisons between different groups such as races or nations, nor for genetic studies. And this discouraging conclusion applies to non-verbal, performance, or other tests as much as to verbal Binet or group tests. My views have altered since 1949 (27), largely because two crucial eugenic studies—those of the Scottish Mental Survey (21) and of Cattell (8)—have given negative results. If the tests that they employed were adequate measures of inborn intelligence, then I do not see how they could fail to disclose some decline over the past fifteen years. Thomson (24) and Cattell are commendably cautious in their discussions of possible reasons for the rises in their group test means, and the lack of any drop on individual tests. But their explorations of possible genetic factors seem to lead only to blind alleys, and even Penrose's (19) ingenious picture of heterosis could hardly compensate for the relatively large fall that was anticipated. Thus they appear to incline to the view that practice or sophistication effects may have distorted the results of the group tests, particularly among girls, and that improvements in health and education may have played some part ; while Thomson admits the possibility that the negative relation between intelligence scores and family size may be partly environmental. Most of my paper to-day, therefore, will be devoted to relevant studies of practice and environmental effects, and their implications.

* A paper read before the *Eugenics Society* on May 29th, 1951.

First, however, I would like to give you the preliminary results of a large-scale survey of the abilities of some 9,000 Army recruits. When the Scottish results were published I suggested that part of the difficulty of interpretation arose from using such factorially complex tests as Moray House and Stanford-Binet, and that further data should be obtained with the aid of factor analysis to supply relatively pure measures of Spearman's *g*.

A Survey of Army Recruits

The scores of about 10,000 male National Service recruits, mostly aged 18, were obtained on six mental tests. Their height and their physical and eyesight gradings on the Pulheems system were available, together with records of the county in which they lived. During an interview with a personnel selection officer they were classified by the place they had chiefly lived as rural versus urban and as native or migrant; finally the numbers of living siblings were ascertained. Such recruits provide a fairly representative sample of the population, but unfortunately few agricultural workers were being called up at the time (the end of 1947). Also there was a wastage of 6 per cent of men who failed to supply the interview data, and these were somewhat inferior in intelligence.* The material was collected by the personnel selection staff of the War Office's Directorate of Manpower Planning, and all the essential tabulations and calculations were done from Hollerith cards by the Statistics Branch of the Admiralty. I am most grateful to those concerned and to my colleague in the Senior Psychologist's Department of the Admiralty —Mr. E. Elliott—for their assistance. A fuller account than I can give to-night will be published in due course.

Table I shows the best solution I have been able to reach so far to the analysis of the inter-correlations; it gives the loadings or saturations of the tests with four main factors or underlying types of ability: *g*— general, verbal-educational, spatial-mechanical and physical. The results are fairly logical, but one should note the overlapping between the *g* and the education factors. Not only does an arithmetic test show the highest *g*-loading but also a non-verbal test and a mechanical one have considerable educational content. Natives, you will see, are slightly (but significantly) inferior to migrants on all factors. Urban dwellers seem to be slightly inferior in all respects, except education, and particularly in physical qualities. But I am afraid that the lack of agricultural workers may have affected these figures; we shall see later that the relationship is not linear. As regards family size, it is quite clear that the negative correlation exists not so much with education as with our general factor. But as the general factor itself is somewhat mixed up with education this result does not carry us much further. It is interesting that the correlation of −·34 is higher than is usually found among children; presumably this is due to more families being complete when a cross-section is taken at 18 than at 11 years.

TABLE I

LOADINGS OF TESTS AND OTHER MEASURES ON FOUR FACTORS EXTRACTED AND ROTATED BY THE CENTROID METHOD

	g	v : ed	mech.	phys.
Arithmetic-maths.... ...	·73	·57		
Verbal ability	·61	·65		
Clerical	·71	·56	·07	
Education standard ...	·60	·71		·06
Non-verbal intelligence ...	·72	·40	·20	
Mechanical problems ...	·54	·50	·47	·12
Mechanical assembly ...	·31		·45	·16
P (physical)	·08			·37
EE (eyesight)	·01	−·23		·32
Height	·09	·33	·13	·20
Variance, per cent ...	26·8	21·4	4·8	3·2
Native	−·04	−·10	−·07	−·04
Urban	−·04	·01	−·03	−·11
Siblings	−·34	−·17	−·09	−·04

I had hoped to calculate independent factor scores, but the labour proved too great, and instead I merely combined the results of certain tests which were most highly loaded with each factor. For ease of comprehension I converted these combined scores into standard scores with a mean of 100 and a

* There were 603 such cases, whose mean intelligence "quotient," on the scale used below, was 96·5. A number of others had failed to take one or more tests, leaving a total of 9,183.

standard deviation of 16, so that they should all be comparable to I.Q.s.* Table II shows the mean standard scores for *g*, education, mechanical and physical tests for each family size. Note, however, that because of the method of derivation the education and mechanical scores are considerably impregnated with *g*. You will see that the decline

TABLE II

MEAN STANDARD SCORES OF RECRUITS WITH GIVEN NUMBERS OF SIBLINGS ON TESTS OF DIFFERENT TYPES

No. of Siblings	Freq. %	Mean Scores *g*	ed.	mec.	phys.	Scottish Mental Survey
0	13·3	106·6	107·2	104·6	102·3	105·3
1	22·0	105·8	105·8	104·3	101·6	105·1
2	18·8	101·8	101·7	101·7	100·8	101·6
3	13·8	98·8	98·5	99·5	99·8	98·6
4	10·4	94·9	94·7	96·7	98·7	95·8
5	7·8	93·2	92·9	95·2	97·7	94·2
6	5·2	92·4	92·6	93·5	96·4	92·8
7-8	5·5	88·9	91·6	92·9	96·5	91·8
9-11	2·8	87·9	88·2	90·6	96·2	90·1
12-17	0·4	87·2	86·2	91·6	95·8	86·5

is almost identical on the more *g*-loaded and the more educational tests, and that both closely parallel the results of the Scottish Mental Survey. On more mechanical tests the decline is less marked, and it is quite small on physical measures; indeed there is scarcely any drop beyond six siblings.

Turning to geographical differences: the counties were grouped into twelve major regions, which were so chosen that no region contained less than 5 per cent of the total, and each would be fairly homogeneous in respect of industrialization. Table III shows that variations in *g* scores in different regions are remarkably large, ranging from an average of 104 in the Home Counties to ninety-four in South-West Scotland. Wales, Lancashire, Stafford and Warwickshire are distinctly below average—ninety-seven to ninety-eight; but the whole of the rest of England and Scotland, including London, is around 100 to 102. There were, of course, variations between counties within regions; but numbers were too small for these to be highly significant. Surprisingly the regional differences on educational tests are smaller than on *g* tests, while mechanical differences are largest of all. The South-West Scots and Welsh are relatively well educated, the East Anglians are significantly poorer than would be expected from their intelligence. On mechanical tests it is the less industrialized and more healthy regions that do best, though Warwick-Stafford does pull up slightly. The physical results run almost parallel to the mechanical and are incidentally very similar to those discovered by W. J. Martin (16) (in so far as our differently grouped regions can be compared.)

TABLE III

MEAN STANDARD SCORES OF RECRUITS FROM DIFFERENT REGIONS

Region of Great Britain	Nos.	*g*	ed.	mec.	phys.
Home Counties ...	1,663	104·0	103·0	104·3	103·0
East Anglia, Beds, Cambs, Northants	474	101·7	100·0	103·0	100·6
S.W. and Hants ...	563	101·5	101·0	103·7	102·5
Berks, Bucks, Oxon, Glos, Here, Worcs	540	101·4	100·9	103·2	102·7
Leics, Notts, Derby, Yorks W., Ches	1,128	101·4	101·0	100·3	99·7
London	695	100·6	99·9	100·8	99·8
Yorks E. & N., Northumb, Dur, Cumb	502	99·6	100·0	99·1	99·2
Scotland E. and N. Counties	486	99·6	99·4	97·0	98·9
Wales	466	97·9	100·0	97·0	97·2
Lancashire	1,125	97·2	97·8	96·5	97·6
Warwick, Staffs, Salop	1,046	97·1	96·9	98·4	99·5
Glasgow and S.W. Scotland	495	93·7	97·0	91·4	95·0
	%				
Natives	74·8	99·0	99·0	99·2	99·5
Migrants	25·2	102·8	102·9	102·4	101·6

	Distance from open country miles	%				
Rural ...	0	4·9	100·4	99·8	102·5	102·8
Village or small town	< 1	22·4	99·8	100·0	100·4	101·0
Middle of mod. town	1-2	18·6	100·9	100·7	100·2	100·2
Edge of large town ...	1-2	15·0	103·0	103·1	102·4	100·7
Middle of large town	3+	39·1	98·4	98·5	98·5	98·7

Migrants are uniformly superior to natives. Note that they constitute only a quarter of the population in spite of all the war-time migrations. Martin found 20 per cent in

* Such units are not normally employed by Service psychologists, who are well aware of the misinterpretations to which I.Q.s, particularly among adults, are liable.

1939. My rural-urban classification refers to the place in which the recruit has chiefly lived (not necessarily where he was born nor where living at present), and was based on the distance of that place from open country. Thus the first group represents those who say they live surrounded by open country, the second those living in small towns or villages where open country can be seen from close by their homes. You will see that the size of the town is not crucial. Those living on the outskirts of large towns are highest and those in the centre lowest on almost all psychological tests. The true country dwellers do not appear to be inferior in abilities; they are actually highest on mechanical as well as physical measures. While I have admitted a possible lack of agricultural workers in my sample, the fact that 5 per cent of the total fall in the most rural group and 22 per cent in the village-small town group suggests that the loss is slight. Mrs. Bosanquet's (2) careful survey showed, like mine, the physical superiority of rural populations. Her evidence for intellectual inferiority was by no means unanimous, and she wondered whether such inferiority as was found might not be ascribed to the content of the tests favouring urban children. I would add to this the suggestion that urban children are more likely than rural to be practised, or sophisticated, in taking tests; and we shall see below that such practice effects may be very considerable. To conclude then: there is too much doubt about the representativeness of my sample for me to contradict the common belief in the lower intelligence of the rural population. But at least it would appear that any inferiority lies more in educational level than in general or practical ability. The more important finding of my survey, from the eugenic standpoint, is that the best measure of intelligence that I have been able to extract from six mental tests behaves in just the same way as tests with obvious educational content in predicting a decline in national intelligence—a decline which has so far not been borne out by direct evidence.

Practice and Coaching Effects on Intelligence Tests

Turning then to environmental effects on test scores: these may be classified under five headings—

1. Coaching or teaching the right answers on the actual test used.
2. Practice in taking the test itself.
3. Practice on other similar tests.
4. Coaching on other similar tests.
5. More generalized educational training or cultural influences which are incidentally reflected in test scores.

The first—coaching on the test itself—is obviously possible, but is unimportant because care is normally taken nowadays to prevent leakages. Secondly—practice effects from repeating the same test; these were shown to be so large in the Army during the war that additional norms had to be provided when men were retested. The improvements ranged from 2·1 per cent to 8·6 per cent on different tests, in terms of I.Q. units with a standard deviation of fifteen. They were highest among tests with unfamiliar content or with elaborate instructions and restricted time limits, such as Progressive Matrices and a clerical test; least marked among straightforward tests such as arithmetic, where also the timing was generous. More recently Dr. Heim (6) and her collaborators have shown that repeated practice at a timed group test (having adequate ceiling) produces almost indefinite rises, though these tend to tail off after about the fifth retest. Obviously this is a quite artificial situation; practice or coaching on other similar tests is of far greater importance. It is particularly serious now that intelligence tests are used by almost all education authorities for competitive entry to grammar schools at eleven years. In many primary schools, much of the final year is spent in coaching children to pass intelligence and objective English and arithmetic tests, in spite of the efforts of the authorities to prevent it. Not only is this unfair to the schools which do refrain from coaching but also it naturally renders the original test norms valueless and it distorts the results of

any comparative investigation, such as the Scottish Mental Surveys. Thus there was virtually no rise in scores between 1932 and 1947 in Scottish districts where children are unused to tests, but a rise of 3·2 points in districts where children are relatively sophisticated.

I have surveyed some of the literature on practice and coaching effects in my book, with Dr. Parry, on personnel selection (29). Important additional results have been obtained by McIntyre (14) in Australia, Johri (11) at Leeds, Dr. Watts (30) in London and Professor Peel (18) at Birmingham; and a series of experiments is being carried out by one of my students, Mr. D. T. Navathe. He is taking the precaution, which few previous investigators have done, of mixing two parallel versions of each test so that those pupils who are initially given version A take version B as their final test, and those who are initially given version B take A finally. Only in this way can slight differences in the difficulties of the versions be controlled.

The increase in score resulting merely from the practice effect of taking one previous test ranges from about 3 per cent to 8 per cent. Peel compared 1,593 pupils who took Moray House Test 41 and Test 42 a few weeks later, also 1,367 pupils who took them in the reverse direction, and obtained an average rise of 3·14 I.Q. points. Mr. Navathe got a rise of 4 per cent after two weeks using similar tests. But both these sets of pupils may have had some previous familiarity with tests, and in another experiment at a private school where the boys were virginal the rise was 7·4 per cent. These, however, were boys of initial I.Q. around 120 and therefore perhaps more able than most to learn from practice. Dearborn and Rothney (9), Rodger (20) and others have claimed that trainability is greater at higher I.Q. levels; but one would naturally expect greater increases below the average than above as a consequence of regression. Peel has attempted to allow for this regression effect and has applied his as yet unpublished technique to Moray House retest results of 7,000 pupils. Table IV derives from a smoothed graph which I drew from his results. It shows that trainability does indeed increase with intelligence up to a certain point, say I.Q. 120, but thereafter there is a dropping off, not because of ceiling effect but because the very bright children understand the initial test so well that they have little to learn.

Navathe compared the improvement on timed tests with that when ample time was allowed; the latter amounts to 2·8 per cent, that is roughly three-quarters as great as with the timed test. Hence the practice effect is not merely a matter of learning to make efficient use of time or to read and follow the instructions quickly.

TABLE IV

SMOOTHED RESULTS (FROM PEEL) OF RETEST RISES

Initial I.Q. Level	Mean Rise on Retest
140	2·9
130	3·9
120	4·1
110	3·5
100	3·2
90	2·6
80	1·9
70	(1·2)
Overall	3·14

With further practice between the initial and final test the improvement may be considerably greater. Thus Johri obtained an additional increase averaging about 12 per cent as a result of ten periods of practice at parallel tests on successive days. All investigators, however, show that this tails off; in other words, that increased practice produces diminishing returns.

TABLE V

SUMMARY OF RECENT INVESTIGATIONS, IN STANDARD SCORE (I.Q.) UNITS

	Practice Effect	Coaching Effect
McIntyre		15-18
Johri ...	12·4	14·8
Watts ...	5·4	9·5
,, ...	3·2	8·2
Navathe ...	4·0	8-11
,, ...	7·4	16·2

Under my fourth category, coaching or instruction in how to answer group test items produces very large rises. Watts obtained an increase of 8 to 9 points from quite intensive coaching over ten weeks, but his testees were secondary modern pupils who already had some acquaintance with tests. Navathe found 8 to 11 points in similar groups. But

in a private school two small groups who received only one or three periods of coaching rose by 16 points. McIntyre's pupils had four and Johri's ten periods. It is, of course, difficult to compare these experiments since the tests and the age and previous sophistication of the children varied. But it certainly appears that an hour or two of coaching can be as effective as much more prolonged training. Notice, too, that coaching is only about twice as effective as the practice resulting from a single retest. In Johri's study ten coaching periods were slightly more effective than ten practice periods, but to a statistically insignificant extent. Thus, from the standpoint of educational selection, the solution is obvious. If all primary schools were made to give a few hours' coaching shortly before the 11-year examination, previously unsophisticated children would be brought up to the same level as those who had already been coached; and schools which coached for months beforehand would, in all probability, have been wasting their time. But the mere application of a short practice sheet or (as happens in some areas) of one preliminary test is clearly not sufficient to raise children to their effective limit.

Many other points require investigation. Different types of test differ in their susceptibility to coaching, but there is no clear evidence as to which are the most prone. Navathe found that a battery of separately timed subtests yields much the same improvement as an omnibus test of the Moray House type; also that allowing ample time reduces the coaching effects only by some 10-20 per cent. He is trying now to find how lasting are the effects, also whether coached or uncoached test scores are the more valid. On the one hand we might expect many uncoached children to fail to manifest their true ability because they often do not understand what the test requires of them. On the other hand, fully coached children may have acquired what is mainly a familiarity with the tricks of the tester and their scores may be less valid indices of the intelligence they show at school or in daily life. How far training on one type of test transfers to other types is a matter on which the evidence is conflicting. In one experiment in the Navy (29), practice obtained by taking one battery of tests produced a 3·6 per cent improvement on another quite different battery. In another recent investigation by Mr. Elliott, five groups, totalling 1,719 men, took the same set of five tests in five different orders; and quite large variations, ranging up to 5·2 I.Q. points, were found when the same test occurred in different positions. For example, the I.Q. score on an abstraction test was 97·7 if taken first or second, but 101·5 if it came later in the series.* On the other hand a straightforward mathematical test requiring creative responses appeared to be upset if it followed other, selective-response, tests. Thus when taken first the score was 102, but after other tests 99·5. These tests all differed considerably from one another in form, and it seems likely that the set of responding to one type of item interfered with the sets required for other types. Thus most of the tests showed a negative practice effect, i.e. more declines than rises with lateness in the series.

Johri trained or practised his boys on analogies, similarities, directions, reasoning and mixed sentences. He also applied, but did not train them on, opposites, best answer, always has and absurdities. He claims that, of the latter, only absurdities showed any improvement; neither the practice nor coaching transferred to the other three. Navathe gave four tests—verbal classification, verbal analogies, non-verbal classification and non-verbal analogies—and he trained four of his groups, one on each of these tests; other groups were practised or coached on all tests. His results show that transfer effects were no greater than those arising from simple practice. Actually there was more transfer to

TABLE VI

SUMMARY OF NAVATHE'S RESULTS

	Points
Practice effect alone, on all four tests	4·0
Practice effect on tests dissimilar to those coached	5·0
Practice effect on tests similar to those coached	4·1
Coaching effect on tests specifically coached...	8·4
Coaching effect when whole battery is coached	11·0

* The five groups were nearly, but not quite, identical in initial ability. However, analysis of covariance showed the variations to be statistically significant.

dissimilar than to similar tests, though the difference is insignificant. For example, those coached on verbal classification showed more improvement at non-verbal analogies than they did either at verbal analogies or at non-verbal classification. Furthermore, those trained for two periods on one test showed less improvement at this test, on the average, than those trained for the same total time on all four tests.

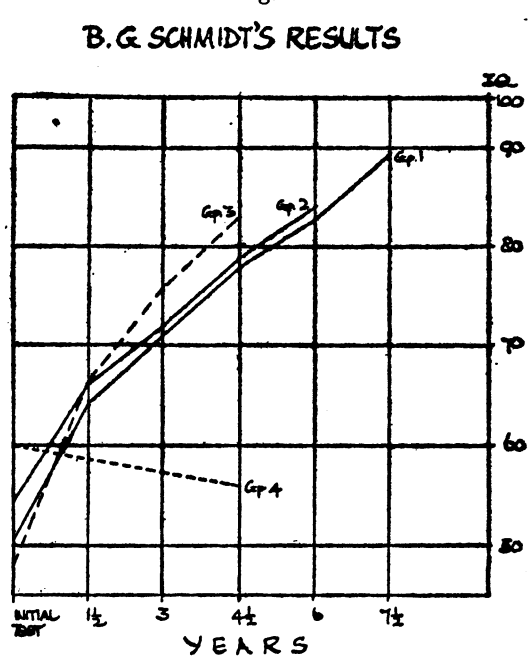
I would suggest the following conclusions, though I would like much more evidence. First, either practice or coaching has a rather general effect in familiarizing testees with intelligence tests, inducing the set of working quickly, of taking careful account of the instructions, not wasting time on difficult questions, being alert to tricky items and in improving confidence and reducing anxiety. One might call this the sophisticated attitude or even the morale of the testees. Dr. Watts suggests that this attitude may not only improve but also decline if testees are over-coached so that they get bored. Second, there is a highly specific coaching effect which helps only at the particular type or types of item coached or practised, and which may hinder facility in tackling other types.

The more general sophistication may well be stimulated to some extent by recent schooling or by the taking of any kind of examination. Adult recruits tend to show rather large practice effects, possibly because their schooldays are so remote that they do not readily settle down to writing answers to silly questions as quickly as possible. But this brings us to the fifth of my headings—the improvement in intelligence test performance attributable to environmental influences and education. I shall not attempt to resummarize the earlier literature which tended to show a definite, though quite small, environmental effect. Nor will I describe the notorious Iowa studies, which claim to show very large effects but which have been devastatingly criticized by McNemar (15). I would, however, like to discuss the startling results published in 1946 by Bernardine Schmidt (22).

B. G. Schmidt's Study of Environmental Effects

Miss Schmidt states that, during 1935-42, she took 320 Chicago children aged 12-14, who had been diagnosed as feeble minded, all with I.Q.s below 70. Three groups, totalling 252 adolescents, were educated for three years at special experimental schools where the curriculum was designed to increase emotional and social adjustment, as well as to provide appropriate training of academic and manipulative skills and study habits. Another, control, group of sixty-eight cases went to more conventional schools. The follow-up of subsequent careers continued for one and a half to four and a half years after the end of schooling. Stanford-Binet tests and certain educational tests and tests of personality adjustment were applied at eighteen-month intervals throughout. The main findings regarding intelligence are shown in Fig. 1. The controls show the com-

Fig. 1



monly observed drop over the four and a half years from an average I.Q. of 60 to 56. But the three experimental groups, starting at an average of about 52, rise to 73 in the

three years' schooling and continue to increase thereafter, Group I reaching 89 in seven and a half years. The Vineland Social Maturity scale showed even greater improvements, and the personality test results were concordant. Apart from tests, half the experimental children went on to ordinary high school courses and 27 per cent graduated from high school. When last surveyed, 83 per cent were in regular employment and nearly two-thirds of these were in clerical or skilled work. By contrast, the controls showed a very poor educational and employment record.

Naturally these results have aroused doubts and criticisms, but it has been very difficult to find flaws, apart from some minor arithmetical slips. At least one reviewer (12) seems to suspect gerrymandering, but this is unlikely since the study is backed by several highly reputable psychologists. (I myself feel grave doubts as to how children in the imbecile grade could have taken such personality questionnaires as Bernreuter's.) The initial intelligence testing was all done by trained psychologists of the Chicago schools Child Study Bureau ; thus we cannot object that some of the children were diagnosed as feeble minded because of bad testing, as occasionally occurs in Britain. Retestings were, apparently, carried out by Miss Schmidt using 1916 Stanford-Binet throughout with Groups I, II and IV, and Terman-Merrill Form L throughout with Group III. Moreover, an independent tester is said to have given a further test to some of Group I, getting virtually identical I.Q.s.

Nevertheless, certain weaknesses deserve mention. First, there are always considerable fluctuations in I.Q.s during the teens. The correlation between tests applied to a representative group four and a half to seven and a half years apart would hardly exceed ·60, and might be lower. Hence, due to regression effects alone, we would expect an apparent rise in average I.Q. of 19 points, that is from 52 to 71. The statistically untrained person finds this very difficult to grasp, but it is really quite simple. If a representative group was tested (with accurately standardized tests) at 12 years and at 16 years there would be as many children with I.Q.s below 70 on the second as on the first occasion. But they would not all be the same children. Owing to imperfect correlation, some of those scoring above 70 at 12 fall below 70 at 16, and vice versa. If, therefore, we pick out only those below 70 at 12 years (as Miss Schmidt has done) they inevitably show an average rise by 16 ; this rise has no psychological significance at all since we are neglecting the above 70s who show a fall in the same period.

A second point is that the experimental children were tested from five to eight times, the controls only twice. Schmidt denies that there is any practice effect when retests are as far apart as 18 months. I doubt this, particularly with children near to their limit of intellectual growth, and suspect that at least 5 points of the average rise might be attributed to practice.

Thirdly, you will note that much the biggest rise occurs in Group III, which was tested with Terman-Merrill. It amounts to 35 I.Q. points in four and a half years compared with 26 points in Groups I and II who were tested with Stanford-Binet. Now in spite of all Terman's care in standardization the new revision undoubtedly exaggerates adolescent I.Q.s. Every psychologist who has used Form L on children of 12 years upwards has been surprised by the number of very high I.Q.s that turn up. In 1940 I pointed out that the vocabulary test alone was considerably easier at the upper end than the Stanford-Binet vocabulary, and I carried out a rough restandardization reported in my book *The Measurement of Abilities* (26). The first two columns of Table VII show that children with a ten-year Terman-Merrill vocabulary score would score nine years only on the Stanford-Binet or according to my norms, while a twenty-year Terman-Merrill score corresponds to sixteen and a half years only. Next, over several years, my students at Glasgow University applied Terman-Merrill to small groups of children along with several other intelligence, educational and performance tests believed to be relatively accurately standardized. In this way I collected records for several hundred

children of various ages and, on graphing Terman-Merrill mental ages against other mental ages, found that Terman-Merrill was probably quite accurate around six to eight years, but then gave more and more exaggerated results. I combined these into a rather conservative restandardization summarized in the third column of Table VII. Then in 1949 the second Scottish Mental Survey appeared, showing a tremendous positive skew in the Terman-Merrill distribution of eleven-year I.Q.s. Whereas 3 per cent of the representative sample obtained I.Q.s below 70, not three but 10 per cent obtained I.Q.s above 130. It is quite simple to work out from the Scottish figures revised norms which turn this skewed distribution into a symmetrical one, and an extract from these norms appears in the last column. My three attempts at restandardization do not agree very closely, but they all show the same trend.* I do not know whether this inaccuracy in the Terman-Merrill scale arises

TABLE VII

RESTANDARDIZATIONS OF TERMAN-MERRILL MENTAL AGES

Obtained M.A.	Revised M.A.s, Based on Vocab. Test	Glasgow Tests	Scottish Survey
8	7 : 6	8 : 0	8 : 0
10	9 : 0	9 : 7	9 : 10
12	10 : 4	11 : 1	11 : 6
14	11 . 7	12 : 7	13 : 1
16	14 : 1	14 : 3	14 : 5
18	15 : 3	15 : 11	15 : 8
20	16 : 6	17 : 7	16 : 10
22	17 : 8	19 : 3	17 : 9

from Terman's failing to secure representative older groups, but I suspect that it is due more to the fact that mental growth is not even approximately linear after about ten years. Several investigations suggest that it is logarithmic, i.e. that there is a gradual deceleration. Terman makes some allowance for this, but not enough. Hence, if successive mental age years above ten represent smaller and smaller increments of mental growth the exaggeration of I.Q.s that we have found would naturally follow. Psychologists are coming more and more to the view that the M.A. and I.Q. system of scoring tests is quite inadequate above ten years or so and are resorting to standard scores or percentiles (cf. Vernon (28)).

This long digression was necessary to bring out my third criticism of Miss Schmidt. Clearly a child of average intelligence at 12 years will obtain an I.Q. greater than 100 if retested with Terman-Merrill four or five years later, merely because of this defect of scaling. I calculate the increase as about 14 points. A bright child might easily show a much larger rise of 30 points. Miss Schmidt's dull children would not be expected to increase as much, because they would still be below the level where Terman-Merrill exaggerates most, but some 5 points of the rise in her Group III may well be due to this. Whether similar exaggeration occurs with the old Stanford-Binet (as used in Groups I and II) we do not know. There is actually more evidence that it gives unduly low I.Q.s among older adolescents and adults. But everything depends on what divisor is used for calculating I.Q.s, and this Miss Schmidt fails to tell us. If it was sixteen years for those aged 16 or over, then the reported rise in average I.Q. is indeed surprising ; if fourteen, then it is much less so. But in any case Schmidt's results are open to the criticism that the I.Q. standardization of all the Binet scales is most unsatisfactory for older children and adults.* When we also take into account the natural effect of regression and the probable practice effect, we must admit that the claims for large increases in I.Q. are very dubious. Personally I am more impressed by the sociological data. One cannot dismiss the finding that, when steps are taken to improve the emotional adjustment and to provide a suitably stimulating educational environment, a large proportion of children once diagnosed as feeble minded are converted into reasonably capable adults. I am sure, however, that the investigation needs to be repeated with better controls and more adequately scaled tests.

* Quite a close approximation to the average of these figures is obtained by counting each year above 7 : 0 as 10 months. E.g. : Obtained M.A. 13 : 10 = 7 : 0 + 82 months; Corrected M.A. = 7 : 0 + 10/12 × 82 = 7 : 0 + 68 months = 12 : 8.

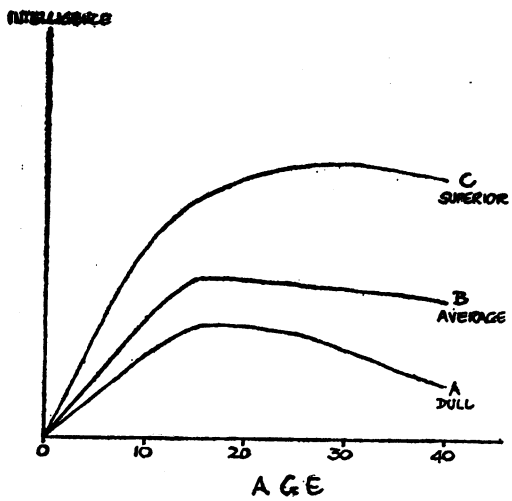
* Yet another likely reason for a spurious rise on Terman-Merrill is that the standard deviation of I.Q.s at 16-17 years is only three-quarters what it is at 12. I know of no information as to whether the same is true of Stanford-Binet.

Other Evidence of Effects of Education

Further evidence of environmental effects is forthcoming from studies of the intelligence of adults. Ever since large-scale testing began in the American army in 1917, two apparently conflicting results have been obtained. On the one hand the average M.A. of representative adult groups has been remarkably low; for example, the mean Army Alpha test score corresponded to a M.A. of 13½. On the other hand, when pupils or students in high schools or colleges are given intelligence tests of appropriate difficulty they usually continue to show rising scores from 15 up to over 20 years. Hence it seems absurd to say that the average individual reaches the limits of his intellectual growth at 13 to 14 years. But these two findings can be reconciled if we allow that growth continues as long as education or other intellectual stimulus continues and that thereafter a decline sets in. The majority of the population, including almost all the average and dull adolescents, finish their education around 14 to 16, hence their intellectual growth and decline might be represented by lines A and B in Fig. 2,

Fig. 2

CURVES OF MENTAL GROWTH AND DECLINE



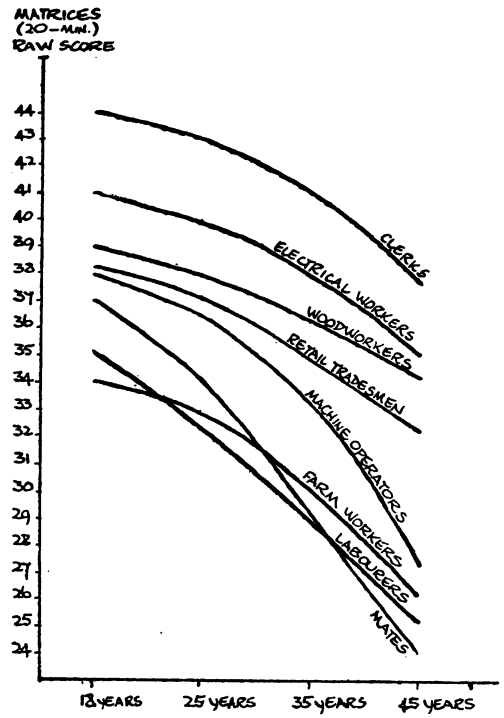
which reach their maximum height at about 15 years. But a minority receive secondary or college education to 18, 22 or later, and when this finishes they usually enter jobs where their intelligence continues to be fully exercised. They may be represented by line C, which shows an increase till 25 or 30 before a gradual decline sets in. Since, however, they are relatively few in number, the mean M.A. of all adults around 20 years is even lower than the maximum that the average pupil reached at 15.

Now what is the evidence for this hypothesis? Lorge (13) retested a number of adults at the age of 34 whose test results at 14 years were known. There was a clear effect of education beyond 14. For example, adults who had received a university education were two years superior in mental age to others who had had no further schooling but whose intelligence level had been the same at 14. Similarly Terman and Oden (23) followed up their gifted California group after 20 years and gave them an individual synonyms and analogies test. It may be deduced from their results that those who had had the best education and had entered professional careers scored best, relative to their I.Q.s as children. Might these results not be due, however, to the use of verbal tests with high educational content? Actually the same seems to hold for other types of test. Thus the mean score for 18 year recruits on an abstraction test, which has only a small verbal component, is equivalent to that of 13 year children; but brighter recruits score distinctly higher than bright 14-15 year olds and dull recruits distinctly lower than dull school-leavers. Again Raven (10) has published age norms from 5 to 65 years for his vocabulary and for his non-verbal matrices tests, both of which show the tendencies pictured in my Fig. 2, namely the fanning out of ability during adulthood and the earlier decline among lower-grade adults.

Elsewhere (29) I have given the mean scores on the twenty minutes' progressive matrices for nearly 90,000 naval recruits, ranging in age from 17 to over 40. These indicate that an average decline has begun as early as 18 years, though it is difficult to be certain of the comparability of the various

age-group samples. When the recruits are classified by civilian occupation there is a definite tendency (shown in Fig. 3) for such

Fig. 3



SMOOTHED GRAPHS SHOWING DECLINE OF INTELLIGENCE WITH AGE IN VARIOUS OCCUPATIONAL GROUPS

groups as mates and labourers, whose work demands least intelligence, to drop more rapidly than clerks, electrical and woodworkers, who are likely to make more use of their brains in their jobs and leisure pursuits.

A corollary of my hypothesis is that we would expect the relation between intelligence and education to be closer in an unselected adult population than the relation between intelligence and educational attainment among children still at school. So long as the educational stimulus is much the same for all individuals, innate factors will be the main determiners of individual differences in abilities; but after 15 years the amount of educational stimulation varies widely, and this, I would claim, affects the measured intelligence. This is borne out by our factor analysis, where the arithmetic test had as high a *g* loading as the non-verbal intelligence test, and the latter had a substantial educational loading. Hence, also, family-size differences in our measure of *g* were practically identical with differences in education.

Conclusions

Now to sum up the inferences that appear to follow from these investigations: I accept the view of Burks (3, 4), of Professor Burt (5) and others, that, within a culturally fairly homogeneous group such as the primary school population of the U.S.A. or of Great Britain, something like 75 per cent of the variance in intelligence is attributable to hereditary factors, provided that care is taken to ensure a uniform degree of familiarity with the tests employed. When the amount of practice or coaching varies between different sections of such a population, no worthwhile conclusions at all can be drawn. But I would say that this large hereditary influence is manifested only because the environmental stimulation is fairly uniform for all children. They all hear much the same language and learn to use much the same concepts; they receive a more or less standardized education; they live in a country where the same pictorial or other concrete symbols are current; they are all trained to attend to printed questions and to write down answers quickly, and so on. All these factors do of course vary between social classes, between families within classes or between children within families, though not enough to raise the environmental component of the I.Q. to more than 20-30 per cent. They begin to vary more widely when children are segregated into different types of secondary schools and after school-leaving age. Probably also they vary more widely among emotionally maladjusted children who have difficulties in absorbing adult concepts, or among Gordon's canal-boat and gipsy children, or among such cases as Miss Schmidt investigated who, through emotional handicap or intellectual weakness, had fallen by the wayside so that the education which stimulates the growth of the ordinary child became too advanced for them. Certainly the variations in concepts, habits of thought, and in attitudes to intellectual tasks, between

members of different nations or races are so large that experts like Nadel (17) and Biesheuvel (1) believe that a culturally neutral intelligence test is an impossibility and that psychologists who try to make racial comparisons are wasting their time. This applies just as much to performance or pictorial or other non-verbal tests as to verbal ones. But I am afraid that this conclusion holds also, though in lesser degree, for many of the comparisons within relatively homogeneous cultures in which eugenists are interested. The stimulus value of the environment in which the numerous children of social-problem families on the one hand, and the only children of university graduates on the other hand, are reared, may account for no more than 25 per cent of the differences in their observed intelligence; but surely this 25 per cent is too large a component for the test scores to be accepted as reliable measures of genetic differences.

I do not think that I am saying anything revolutionary about intelligence tests. It is, after all, a commonplace of genetics that qualities as such are not inherited, only the capacity to develop these qualities under favourable environmental conditions. Thus I would regard intelligence as the outcome of the interplay of innate potentiality and of such conditions as good emotional adjustment and appropriate educational stimulation. I am not sympathetic to the argument that intelligence itself is an innate quality, and that our tests are inefficient at measuring it because they are subject to environmental influences. Rather, I regard intelligence operationally as the general all-round ability that an individual manifests in his daily life adjustments, at school, in his job, or in test performances, and all these manifestations are environmentally as well as hereditarily determined. That brings me to my last point: my criticisms of tests apply only to their use in genetic studies and do not affect their application to practical problems such as educational and vocational prediction. Success at school, or in a skilled job, is helped by a superior environment or education and hindered by a poor one. Thus it is better predicted by a test which is culturally biased than it would be by that hypothetical and unattainable instrument—a pure test of inborn ability.

REFERENCES

1. Biesheuvel, S. Psychological tests and their application to non-European peoples. *1949 Yearbook of Education*. London: Evans Bros.
2. Bosanquet, B. S. The quality of the rural population. EUGEN. REV., 1950, **42**, 75-92.
3. Burks, B. S. Nature and nurture: Their influence upon achievement. *27th Yearbook, National Society for the Study of Education*. Bloomington, Ill.: Public School Publishing Company. 1928.
4. Burks, B. S. On the relative contributions of nature and nurture to average group differences in intelligence. *Proc. Nat. Acad. Sci.*, 1938, **24**, 276-82.
5. Burt, C. L. *Intelligence and Fertility*. London: *Eugenics Society* and Hamish Hamilton. 1946.
6. Cane, V. R., and Heim, A. W. The effects of repeated retesting: III. Further experiments and general conclusions. *Quart. J. Exper. Psychol.*, 1950, **2**, 182-97.
7. Cattell, R. B. Is national intelligence declining? EUGEN. REV., 1936, **28**, 181-203.
8. Cattell, R. B. The fate of national intelligence: Test of a thirteen-year prediction. EUGEN. REV., 1950, **42**, 136-48.
9. Dearborn, W. F., and Rothney, J. W. M. *Predicting the Child's Development*. Cambridge, Mass.: Sci.-Art. 1941.
10. Foulds, G. A., and Raven, J. C. Normal changes in the mental abilities of adults as age advances. *J. Ment. Sci.*, 1948, **94**, 133-42.
11. Johri, S. R. *The Effect of Coaching and Practice on Intelligence Tests*. M. Ed. Thesis, University of Leeds. 1939.
12. Kirk, S. A. An evaluation of the study of Bernardine G. Schmidt. *Psychol. Bull.*, 1948, **45**, 321-33.
13. Lorge, I. Schooling makes a difference. *Teach. Coll. Rec.*, 1945, **46**, 483-92.
14. McIntyre, G. A. *The Standardisation of Intelligence Tests in Australia*. Australian Council for Educational Research, Educ. Res. Ser., No. 54. Melbourne University Press. 1938.
15. McNemar, Q. A critical examination of the University of Iowa studies of environmental influences upon the I.Q. *Psychol. Bull.*, 1940, **37**, 63-92.
16. Martin, W. J. *The Physique of Young Adult Males*. Medical Research Council Memorandum No. 20. London: H.M. Stationery Office. 1949.
17. Nadel, S. F. The application of intelligence tests in the anthropological field. *The Study of Society* (edit. F. C. Bartlett and E. J. Lindgren). London: Kegan Paul. 1939.
18. Peel, E. A. A Note on Practice Effects in Intelligence Tests. *Brit. J. Educ. Psychol.*, 1951, **11**.
19. Penrose, L. S. The Galton Laboratory: Its work and aims. EUGEN. REV., 1949, **41**, 17-27.
20. Rodger, A. G. The application of six group intelligence tests to the same children, and the effects of practice. *Brit. J. Educ. Psychol.*, 1936, **6**, 291-305.
21. Scottish Council for Research in Education and the Population Investigation Committee. *The Trend of Scottish Intelligence*. London: University of London Press. 1949.

22. Schmidt, B. G. Changes in personal, social and intellectual behaviour of children originally classified as feebleminded. *Psychol. Monogr.*, 1946, **60**, No. 5.
23. Terman, L. M., and Oden, M. H. *The Gifted Child Grows Up.* Stanford, Calif.: Stanford University Press. 1947.
24. Thomson, G. H. Intelligence and fertility. EUGEN. REV., 1950, **41**, 163-70.
25. Vernon, P. E. Intelligence test sophistication. *Brit. J. Educ. Psychol.*, 1938, **8**, 237-44.
26. Vernon, P. E. *The Measurement of Abilities.* London: University of London Press. 1940.
27. Vernon, P. E. Psychological studies of the mental quality of the population. *Brit. J. Educ. Psychol.*, 1950, **20**, 35-42.
28. Vernon, P. E. The interpretation of intelligence test results. To be published in *Le Travail Humain*, 1951.
29. Vernon, P. E., and Parry, J. B. *Personnel Selection in the British Forces.* London: University of London Press. 1949.
30. Watts, A. F. *Coaching for the Grammar School Entrance Examination.* Paper given at the Education Section, British Psychological Society, Dec. 29th, 1950.