VOLUME XXIV. JUNE, 1954 PART 2

SYMPOSIUM ON THE EFFECTS OF COACHING AND PRACTICE IN INTELLIGENCE TESTS

V.—CONCLUSIONS.

BY PHILIP E. VERNON
Professor of Educational Psychology, University of London, Institute of Education.

I.—POINTS OF AGREEMENT.

THOUGH there have been somewhat heated disputes among psychologists in the past, this symposium marks a considerable convergence of views. Indeed, there is little that I want to say which has not already been said by one or more of the other contributors. There are still discrepances in the estimates of the gains attributable to coaching children at intelligence tests ; but I hope to show that these can be explained. It is also still true that Mr. James is a strong advocate, Mr. Yates a strong opponent, of legalised coaching—in the sense of teachers giving oral explanations and advice. Yet Dr. Dempster (slightly pro) and Dr. Wiseman (slightly con) approximate fairly closely in their recommendations. Let us begin, however, by noting the much larger number of points on which there is virtually complete agreement.

(1) The problems of coaching and practice arise largely from the use of tests for competitive purposes instead of, as was originally intended, for diagnostic, survey, and experimental research purposes. Yet they are inevitable so long as educational policy requires the sharp separation of vast numbers of 10-11 year children into the ' sheep ' and the ' goats ' in a period lasting only a few weeks. They would become far less serious if the procedure became less mechanical (more in the nature of allocation than selection), if more account could be taken of teachers' estimates and records and other information collected over a longer period, also if the various types of secondary schooling became more nearly equal in their attractions.

(2) Previous practice and/or coaching do make sufficient difference to intelligence test scores to affect the fate of a proportion of children at the borderline. Even if this proportion is small, it manifestly lays the selection procedure open to criticism. Thus, we must find some way of ensuring as far as possible that particular schools, or particular children within a school, do not undergo an unfair advantage or disadvantage.

(3) Any steps that are taken should not increase, and should, if possible, reduce the strain and anxiety among primary school pupils and the distortions of their curriculum which so often result, at the moment, from undue emphasis and drilling on the selection examinations by teachers or parents. We should aim also to reduce any anxieties attributable to insufficient familiarity with the tests or examinations.

(4) The effects of any type of practice or coaching (other than learning at the actual selection test itself) are definitely limited. A few hours only produce the maximum achievable average gains ; hence any larger amount, either at home or at school, is not only undesirable but futile. Fortunately, this implies that a small amount of test experience is likely to meet the need expressed in

No. 2, by bringing previously unsophisticated children up to the level of those who have had previous experience.

5.—Regardless of the size of the average gain brought about by various forms of practice or coaching, there are always large individual differences in gains, and great irregularities in progress at successive tests. Thus with a mean rise of, say, 6 I.Q. units from initial to final test, a typical range of individual changes would be from +20 to −8 units (i.e., a S.D. of about 5·5). I welcome the fresh evidence from the National Foundation that emotional and motivational factors are involved in these irregularities,* and accept Dr. Watts's contention that the results of more than one test should be taken into account in order to improve reliability. But, from the small amount of evidence available, it appears that later tests are slightly more reliable and have slightly better predictive validity than early ones, with which children are insufficiently familiar. Reliability and validity are lowered too when some children have been practised or coached and others not. As I have shown elsewhere,[2] this effect is smaller than might have been expected; and our selection procedures are, therefore, certainly not being seriously invalidated by the present confusion over coaching, as many lay critics seem to assume. Nevertheless, it may be concluded that the best results are likely to be obtained from two or more tests, *after* all the children concerned have obtained some experience of tests.

(6) Some types of test items are much more susceptible to practice or coaching than others, depending chiefly on the complexity of the test instructions and the roundaboutness or unfamiliarity of the operations involved in answering the item. Thus creative-response opposites, and straightforward multiple-choice information items show very small gains, whereas the ingenious but over-elaborate similarities item—where one word which most closely resembles three given words has to be picked from five possibilities—shows far bigger effects. For the same reason, non-verbal test material is usually more affected than verbal. A subsidiary, but interesting, point is that the more diverse the kinds of item or sub-tests within a test, the greater is the test's improvability. It can easily be shown statistically[4] that gains in overall I.Q. depend on item-intercorrelation as well as on the improvability of the separate items. Thus a test like Otis Advanced, with numerous diverse sub-tests, is more coachable than a relatively homogeneous test. However, it certainly does not follow that the most homogeneous tests have superior validity—more likely the reverse.

Clearly there is a need for more research into tests which would be less affected by practice or coaching than many of those employed at present, which might also show better predictive value. Travers[3] has suggested that, by including some relatively coachable, other relatively non-coachable, items in the same test, it would be possible to determine which children, or which schools, had received coaching, and to correct their scores accordingly. But his own work, and that of Navathe,[4] indicate that the technique would be too unreliable for practical application.

II.—Methodological difficulties.

Turning now to discrepancies between the results and recommendations of different authors, these sometimes arise from defects or difficulties in the methods

*One of my students, Dr. D. V. Connor[1] has recently demonstrated greater variations in retest results among maladjusted than normal children, but did not find any clear tendency for variability to correlate with tests or ratings of emotional instability in the normal group.

of investigating practice or coaching. It may be pointed out first that the large Standard Deviation of gain scores (referred to in No. 5 above) leads to serious unreliability in the *mean* gains of any smallish groups of children. The Standard Error of the gain in a class of forty children is typically 0·87 units of I.Q. Thus, different classes subjected to the same practice or coaching conditions may by pure chance show gains differing by as much as 4 to 5 units. In order to explore the effects of different kinds of coaching, it is essential to have large numbers—larger than most experimental educational psychologists can readily lay their hands on. For instance, a difference of 2 units in mean gains between two groups is not likely to be statistically reliable unless the groups contain at least 100 children each ; a difference of one unit similarly requires groups of 400.

A second difficulty is that the units in which the gains are expressed are often not comparable from one test, or one experiment, to another. This is particularly true of earlier publications and student theses, dealing with practice or coaching. I have suggested that gains should always be divided by the Standard Deviation of scores of a representative group on the initial test, and thus converted to standard scores with some arbitrary S.D.—say 15—so that they would be comparable to Stanford-Binet and Moray House I.Qs. Unfortunately, this S.D. is not always easy to ascertain. It was due to a misinterpretation of the S.D. of the tests employed in Navathe's researches that I originally over-estimated the effectiveness of coaching. As most readers will know, I claimed that Navathe had found a few hours of coaching to raise the mean I.Q. of unsophisticated pupils by 15-16 I.Q. units ;[5] the correct figure was 11 units. Most of the recent studies in this country have been carried out with group tests which already yield I.Qs. with S.D. 15, hence this problem is circumvented.

There is, however, a third problem which has been ignored in several studies, namely controlling the relative difficulties of the initial and final tests. Even when tests as closely parallel as the successive Moray House ones are employed, there are slight differences in standardisation. Thus a gain quoted as, say, 5 units might have been 4 or 6 units had the final test been of precisely the same difficulty as the first. And when tests issued by other authors or organisations are investigated, the discrepancies are likely to be considerably larger. Clearly, this defect of technique does not upset comparisons between an experimental and a control group, but it does seriously affect the kind of experiment which involves comparisons of the mean scores of the same group on a series of tests. Few psychologists have taken the obvious, and not unduly troublesome, precaution of rotating the tests, so that each one is taken by a proportion of the children at each stage in the series. When comparisons are being made merely between initial and final tests, half the children should take Test A first and B last, half the reverse.

III.—POINTS OF DISAGREEMENT.

Maximum gains obtainable by coaching.—There is fair unanimity in the older literature that coaching on parallel test material leads to rises equivalent to 15 I.Q. units or over, and it is by no means true that these studies were conducted merely on American adults. Glick[6] and Gilmore[7] did work with students, Bishop[8] with high school pupils. But McIntyre[9], in Australia, tested 12-year-olds, and Chapman[10] and Johri[11] used 10-14 year-olds in English schools. On the other hand, James, Dempster, Wiseman and the National Foundation nowadays often obtain gains of 5 to 6 units only—no more, and sometimes less, than that of matched groups who receive extensive practice without coaching, and not much greater than that of control groups who take

but one practice test (the initial test). However, there is a close approximation between Navathe's finding (11 units), Watt's pilot study, Hammond's enquiry,[12] and Dempster's 1951 and 1953 results (all 8 to 9 units). In another large education authority, practice on two preliminary tests in 1952 led to a gain of about five units on the 1951 mean, while similar practice +3 hours' coaching in 1953 led to a gain of 9 units.

The discrepancies may be partly due to the problems of units and of difficulty levels of tests, already mentioned. But the main factors would undoubtedly seem to be the previous sophistication of the testees, and the ' aptness ' of the coaching. After discussions with Middlesex teachers, I certainly cannot accept Yates's contention that Middlesex children are unfamiliar with tests, nor that the effects of previous familiarisation disappear in six months.* I have summarised elsewhere the results of a number of experiments showing this generalised sophistication effect.[2] Perhaps one of the most striking is the Scottish Research Council's finding[14] of a rise of about 4 points between 1932 and 1947 on the Mental Survey test in areas where children were likely to be pretty familiar with tests in 1947, but almost no rise in more remote areas where group tests were still novel, and no gain anywhere on Binet tests.

But I agree with Wiseman that the type of coaching or practice is also extremely influential, and that coaching given by parents or teachers without doing one or more complete tests under timed examination conditions is remarkably ineffective. Navathe's studies also showed how little transfer may occur when the material, and conditions of testing, are not closely parallel. For this reason, uninstructed practice may sometimes produce bigger gains than coaching. But coaching, which includes practice, as used by Navathe, Hammond and by Dempster in 1951 and 1953, is more effective still ; and its effects are greater the more unsophisticated the testees to begin with. I would conclude then that it is the combination of these factors of unsophistication and aptness which account for the differences between the 11 or even 15 units in some experiments and 5 in others.

In most areas nowadays, we can expect rises averaging 8-9 units from ' apt ' practice and coaching. Note that, because of individual differences, this implies that some 17 per cent. of children will show gains of 14 units and over. The figure will certainly be lower for more inept coaching, and it may be higher in areas where tests are still generally unfamiliar. In any selection examination there are always likely to be a few candidates (from small rural, or from private, schools) who have no previous acquaintance with tests at all, and who are, therefore, very seriously handicapped. Moreover one cannot argue that, because the intelligence test usually receives only one third weight, the effects of coaching are diluted. For objective English, and possibly arithmetic, tests are likely to be equally susceptible to improvement derived from familiarity with the kind of test (as distinct from improvement in school English and arithmetic as such).

Differences between teachers as coaches.—I do not deny such differences, but merely point out how difficult it is to prove their existence when random variations in gains between classes are so large. Contributory evidence is provided by the observation that most small-scale experiments in which the psychologists themselves, or specially selected and trained teachers, do the coaching seem to yield larger rises than bigger investigations where more miscellaneous teachers, or all the teachers in an education area, undertake coaching.

*Professor Peel[13] has withdrawn the evidence which Yates cites as convincing.

But the crucial point, which James and Dempster have already brought out, is that class variations are likely to be larger still when some teachers coach or practice and others don't. The National Foundation's results[15] suggest two other interesting features, though the published information is insufficient to prove them : first, that class variations between practised but uncoached groups are as big as those between coached groups, and secondly, that over-coaching reduces the range of differences. If further work confirmed these suggestions, the argument that authorised coaching increases class differences could hardly be maintained.

Other doubtful points.—The results of different studies of the relation between initial ability and susceptibility to improvement differ. On the whole the evidence supports the common-sense view that dull children benefit less than bright from uninstructed practice, but that they are helped *relatively* more, even if not absolutely more, by coaching. There is little information about the effects of age, but gains among adult students and 18-year-old recruits appear generally very similar to those of 10-12 year-old pupils.

Statistically significant sex differences in gains due to coaching or practice have been claimed in favour of boys by some writers, girls by others. This needs clearing up.

IV.—Discussion and recommendations.

From the long-term viewpoint the controversies over coaching and practice, and the investigations they have stimulated, have been of considerable value to psychology in that they have brought out the need for greater caution in interpreting intelligence test results. Clearly the kind (rather than the amount) of previous test experience does make a difference, and this must be taken into account in any research work involving group comparisons, or where large test batteries are employed. It cannot be ignored, either, in educational or vocational guidance, or in other selection procedures such as those evolved by the Defence and Civil Services.

In the specific field of educational selection at 10-11 years, the situation is most distressing to psychologists, and no satisfactory solution can be expected merely from modifications of testing technique. As already pointed out, the whole problem is too much bound up with social and educational policy and with teachers' and parents' attitudes. Making more use of relatively non-improvable tests, or changing the types of test item from year to year in an effort to defeat the coacher might appreciably reduce the present unfairness, but would do nothing to lessen the pressure on the children. One can only hope, with Yates and Dempster, that the ' climate ' of selection is shifting in directions which will, eventually, eliminate the incentives to coach.

No contributor appears to have challenged James's view that illicit coaching can no longer be suppressed ; all four are willing to introduce some form of legalised practice or coaching in order to counter this. (One point which has not been brought out is that some such familiarisation is positively desirable in that it reduces children's fears about how they are going to be examined. In so far as they all understand what they are going to do, reliability and validity are likely to be improved). Yates and Wiseman provide strong arguments against authorised coaching, and the former advocates instead three or four practice tests without any instruction or knowledge of results ; the latter suggests one practice test which is marked by the pupils, but which is not accompanied by any further explanations. I agree that either procedure is likely to be as effective in raising mean scores as most of the coaching done from books, or from their

own imagination, by teachers or parents. But I also have to agree with James and Dempster that such schemes would not prevent further coaching, and that they would be less effective than a scheme, like Dempster's, which combines practice with coaching. Moreover, one can hardly expect education authorities to sanction the ' waste ' of several practice tests.

Regretfully, therefore, I come to the conclusion that in areas where coaching is widespread, competition severe, and the relations between primary teachers and the authority not too good, it is essential to give two parallel trials of the whole selection examination (not only the intelligence test) before the final one ; to mark these with the children and to allow teachers to give further guidance for a few periods after each trial, based on the scripts. But, in areas where grammar school provision is more lavish, coaching less common, and particularly where adequate use is made of primary teachers' estimates or of a second-stage entrance examination conducted by the grammar schools, I hope that a single trial run+coaching would suffice at least to iron out most of the differences between classes or individuals who have had varying amounts of previous test experience. Under both plans it is desirable that the final examination itself should be in duplicate, and the two results combined. In both, also, the teachers should be supplied with hints on effective coaching, and every opportunity should be taken to rub in, to teachers and parents, the harmfulness and futility of coaching beyond this.

V.—Summary.

1.—There is general agreement between the contributors to the symposium on a number of points, such as the dependence of coaching on the present competitive nature of grammar school selection, the need to do something about it in order to correct injustices, the desirability of reducing the pressure on the pupils, the futility of large amounts of practice or coaching, the great individual differences in ' coach-ability ' and differences between different kinds of tests in ' coach-ableness.'

2.—Many of the apparent discrepancies between the results of different investigators are attributable to methodological problems : the large Standard Error (or unreliability) of coaching gains, the expression of gains in non-comparable units, and the failure to control the difficulties of tests that are being compared.

3.—Though there is still disagreement regarding the average gain to be expected from coaching, the consensus of evidence suggests that a combination of coaching and practice will normally lead to an average rise of about 9 I.Q. units. The figure is likely to be higher among testees who are genuinely unsophisticated about tests to begin with, and it is lower when previous sophistication is considerable, and when the coaching excludes practice at taking complete tests under standard conditions (e.g., when done from published books of test materials).

4.—The contentions that the authorisation of coaching would exacerbate differences between children in different school classes, and that the application of practice tests without further coaching would eliminate differences due to differing previous test experience, cannot be accepted. When selection is carried out by competitive tests alone, a rather elaborate process of practice and coaching on all the tests is regarded as essential. But a simpler process should be adequate when variations in previous experience are less marked, and when other criteria are taken into account in selection.

VI.—REFERENCES.

[1] D. V. CONNOR : *The Effect of Temperamental Traits on the Group Intelligence Test Performance of Children*, Ph.D. Thesis, 1952, University of London Library.

[2] P. E. VERNON : " Practice and Coaching Effects in Intelligence Tests," *Educ. Forum*, XVIII, 1954.

[3] R. W. M. TRAVERS : " The Elimination of the Influence of Repetitions on the Score of a Psychological Test," *Annals of Eugen.*, VIII, 303-318, 1938.

[4] D. V. NAVATHE : *The Influence of the Form of Items on Intelligence Test Scores and its Susceptibility to Practice*, Ph.D. Thesis, 1952, University of London Library.

[5] P. E. VERNON : " Intelligence Tests," *Times Educ. Suppl.*, 25th Jan., and 1st Feb., 1952.

[6] H. N. GLICK : " Effect of Practice on Intelligence Tests," *Univ. Illinois Bull.*, No. 27, 1925.

[7] M. F. GILMORE : " Coaching for Intelligence Tests," *J. Educ. Psych.*, XVIII, 119-121, 1927.

[8] O. BISHOP : " What is Measured by Intelligence Tests ? " *J. Educ. Res.*, IX, 29-36, 1924.

[9] G. A. McINTYRE : " The Standardisation of Intelligence Tests in Australia," *A.C.E.R. Educ. Res., Ser.*, No. 54, 1938.

[10] A. E. CHAPMAN : " The Effect of School Training and Special Coaching on Intelligence Tests," *Forum of Educ.*, II, 172-183, 1924.

[11] S. R. JOHRI : *The Effect of Coaching and Practice on Intelligence Tests*, M.Ed. Thesis, 1939, University of Leeds Library.

[12] D. HAMMOND : " Preparation for I.Q. Tests," *Times Educ. Suppl.* 12th Jan, 1953.

[13] E. A. PEEL : " Footnote on ' Practice Effects Between Three Consecutive Tests of Intelligence ',' *Brit. J. Educ. Psych.*, XXIII, 126, 1953.

[14] SCOTTISH COUNCIL FOR RESEARCH IN EDUCATION : *The Trend of Scottish Intelligence* (Univ. of London Press, 1949).

[15] A. F. WATTS, D. A. PIDGEON and A. YATES : *Secondary School Entrance Examinations* (Newnes Educ. Pub. Co., 1952).