THE POLITICS OF TESTING AT PRIMARY LEVEL IN THE UNITED KINGDOM.

Maurice Galton, Linda Hargeaves and Susan Cavendish

University of Leicester.

Introduction.

The introduction into schools of the assessment procedures, devised by SEAC, for recording the performance of primary pupils on the various attainment targets in key stages one (7+) and two (11+) of the National Curriculum has been the subject of widespread criticism by teachers (Muschamp et al 1992). As a result of these complaints SEAC (Schools Examination & Assessment Council) is now proposing simpler versions with an emphasis on pencil and paper group tests wherever possible. So far, however, little empirical evidence has been forthcoming which would help test constructors decide which domains can more readily be assessed through the written mode and which cannot. Teachers find themselves in a "catch 22" situation, since having complained about the length of time it took to carry out the testing programme at key stage one they now find it difficult to oppose the governments intention to give them some relief, even if the solution offered looks to be similar to that used two decades ago for selection at 11.

The original test specifications, drawn up by SEAC when commissioning the pilot projects for key stage one were based upon the recommendations of the TEGAT report, (DES 1988). The report argued that the three main modes of assessment, currently in use in the primary school, all had certain defects which gave rise to different forms of unreliability. Written tests, by their nature, were constructed in such a way that they could not replicate the conditions in which the curriculum knowledge and curriculum processes had been acquired by the pupils. Practical tasks, while overcoming this problem of curriculum invalidity, were the subject of observer error, while the third mode suggested by TEGAT, teacher assessment required summative judgements to be recorded and these were subject to teacher bias and expectancy effects. By combining a pupil's scores on all three modes TEGAT argued that these non systematic errors would tend to cancel out.

Behind these arguments is an assumption that the three different modes are all measuring very similar traits since if this were not so there would be no justification for combining the three scores. However differences in a pupils' performance, across the three modes, could vary according to the cognitive demands generated by the task either because the skill assessed was context dependent or because there existed an interaction between the mode and the pupil's ability. The present paper is designed to explore these possibilities.

Methods

The data on which the following analysis was performed arose from a study designed to improve the quality of science teaching in the primary school. The Science Teacher Action Research Project (STAR) sought to provide teachers, initially, with a better understanding of the processes involved in science investigations. Using this improved subject knowledge teachers then engaged, with the research team acting as consultants, in a series of small scale classroom studies designed to improve the ways in which these scientific processes were taught.

Increased teacher understanding was brought about by asking teachers to mark a series of specially constructed written test items centring around a common theme, a visit to a Walled Garden (Harlen et al. 1990). In carrying out this marking exercise teachers were able to develop a shared sense of meaning when considering for example what was involved when pupils hypothesized or planned an experimental investigation. The Walled Garden test was administered at the beginning of the school year. The effectiveness of the subsequent teaching was assessed using a practical task covering the same process skills as the Walled Garden (Russell and Harlen 1991). A practical task was preferred for the post-test to ensure the maximum congruence between the teaching and the assessment. However ideally it would have been better to administer both the written test and the practical task on the second testing but teachers felt this to be too great an imposition on both themselves and the pupils. One school, with 46 pupils was, willing, however, to carry out the extra testing and it is these results which are described in this paper.

ANALYSIS OF RESULTS.

To enable comparisons to be made between pupils' scores on the Walled Garden exercise and the practical Sprinkler task each score was categorised as either 'high', 'medium' or 'low'. For the Walled Garden this was done by constructing cumulative frequencies and determining the score corresponding to the thirty third and sixty sixth percentile respectively. The same procedure could not be used for the Sprinkler since the criteria for marking this practical exercise was based upon a predicted hierarchy. Thus, for example, in order to score on interpretation of data the pupil must have either observed or measured some aspect of the sprinkling activity. In other instances, as with observation, there was very little discrimination between the scores in the different categories. Calculation of the correlation coefficient between the scores derived from the written and practical activity was therefore not possible in such cases. Table 1 shows the correlation coefficients (both Pearson r and Spearman rho) between the two sets of scores.

Process skill	Pearson r		Spearman rho	
Observing	-		-	
Recording	0.06	n.s	0.07	n.s.
Measuring	0.22	n.s	0.23	n.s.
Planning	0.35	<0.1	0.29	<0.5
Interpreting	0.44	<0.1	0.41	<0.1
Hypothesizing	0.47	<0.1	0.41	<0.1
Raising Qns	0.43	<0.1	0.34	<0.1

TABLE 1. Intercorrelations between science process skills measured through written and practical activities.

In general, the trend is for the size of the correlations and hence the common variance to increase with the cognitive demand of the tasks. A possible explanation is that more complex intellectual tasks, such as Hypothesizing and raising questions, demand a more general 'science reasoning ability rather than a specific skill. This general science factor is not so affected by the mode of presentation (i.e. written or practical). Further, on some tasks, there exists an ability-mode interaction effect such that only the more able pupils achieve 'high' or medium scores on the more complex activities but on tasks such as observation, however, those same pupils score well on the written exercise but not on the practical one.

This hypothesis can be explored by examining the correlations between the scores on a particular science process skill and the combined score on the remaining skills (i.e. a persons total score less the persons particular skill score). The higher the correlation the greater the likelihood of an individual with a high score on a particular science process skill also achieving a relatively high score on the combined totals of all the skills. Such a result would argue for the existence of a common science ability factor which extends across the range of process skills measured rather than a series of clearly identifiable traits, each related to a distinct scientific process. Table 2 shows these correlations for both the Walled Garden and the Sprinkler exercises.

Process Skill	Walled Gdn (spearman rho)	Sprinkler (pearson r)	N
observation	-	0.45	46
measure	0.78	0.34	46
crit reflection	-	0.60	46
ä plan	0.51	0.80	46
interpret	0.85	0.72	46
hypothesize	0.74	0.62	46
raise questions	0.47	0.63	46

TABLE 2. Inter-correlations between individual process skill scores and the combined skills scores for Walled garden (spearman rho) and Sprinkler (pearson r)

With the relatively small sample size all the values are statistically significant ($p<.01$ one tailed test). More importantly the trend is for the size of the correlations in the sprinkler tasks to increase as the cognitive demands become greater. The trend for the walled garden exercise is less discernible partly because opportunities for critical reflection were limited on the written exercise and because, as stated earlier, observation scores did not vary among the sample sufficiently to discriminate between high scoring and low scoring pupils. The high value for the measurement correlation on the Walled Garden is partly explained because pupils were required to carry out this activity as a preliminary to completing the questions requiring interpretation of data. During the sprinkler activity no such guidance was given and it was left to the pupils to decide whether they should take a precise measure in order to carry out the remaining parts of the task Raising questions on the walled garden activities was also an "open ended" task during which such questions arose spontaneously rather than as in the sprinkler task where they were prompted by teacher/observer. While, therefore, the evidence is not totally conclusive because of some inherent weakness in parts of the written test, there is a discernible trend in support of the hypothesis that tasks requiring higher levels of cognition tend to have a common factor irrespective of the mode in which they are presented.

The existence of an ability mode interaction on some of the process skill measures would ideally been tested using a measure of general ability as an independent variable in a two-way analysis of variance with two levels in the dependent variable (written and practical tasks). Without this general ability measure the total written test scores were ranked and assigned to the top, middle and bottom third percentile. The written scores were used as a measure of general scientific ability because in Table 2 the separate process skill scores generally correlated more highly with the total score than those for the practical tasks. Table 3 shows the average score for each process skill on the sprinkler exercise for each of the three

scientific ability levels and although the data is limited there are trends which suggest that with larger samples interaction effects could be

discerned.

	Low	SD	Med	SD	Hi	Std	f	P
Recording	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Planning	54.8	37.0	68.3	15.9	92.9	15.6	10.0	.0003
Observer	55.6	37.8	63.9	23.0	72.9	17.1	1.54	NS
Meas	13.2	12.1	17.0	16.0	21.7	15.0	1.2	NS
Cr. Ref.	16.7	14.4	25.7	16.9	32.8	16.4	3.1	.05
Interp	38.9	32.8	53.7	23.3	75.0	21.1	7.2	.002
Hyp	17.9	16.2	26.9	12.2	28.8	7.2	3.1	.054
NS?R Qs	25.0	21.3	41.7	25.7	46.9	20.2	3.3	.04

TABLE 3 Breakdown by ability - sprinkler skills

THE POLITICS OF ASSESSMENT.

The above analysis, although constrained by the limitations of the data poses interesting questions for educational policy makers currently engaged in revisions of the National Curriculum and its related standard assessment tasks. Before exploring such questions it is interesting and of value to trace the history which led to the development of the SATs. Back in the early seventies the growth of the accountability movement both in the United States and the United Kingdom led to increased demands for a system of national monitoring. In both countries these new testing procedures sought to abandon the "norm-reference" approach and to substitute banks of items which would assess a large number of different objectives, a wide range of content and also reflect different teaching approaches to the subject. Such items were said to be criterion-referenced. The distribution of item scores is bimodal consisting of a group of candidates who have and another group who have not completed the item satisfactorily.

In the United States the National Assessment of Educational Progress (NAEP), the body responsible for coordinating the design of these new test procedures was unable to develop a satisfactory analysis programme based upon the idea that scores on any sample of items drawn from the bank could be interpreted in relation to the population statistics for the particular criterion which the items purported to measure (Galton 1979). Given the enormous sums spent on the enterprise it was too damaging politically to admit this failure. The monitoring exercise was therefore put in motion and items selected for their face validity and their difficulty level using the more familiar techniques originally developed for norm reference testing. No longer able to use internal consistency coefficients to estimate reliability constraints of time prevented the use of alternatives such as test-retest correlations. These tests were said to be "objective referenced" representing a rather unsatisfactory psychometric compromise between norm referenced and criterion referenced approaches.

In the United Kingdom few lessons were learned from the American experience, although in setting up the Assessment and Performance Unit (APU) the DES sent its representatives and those from the National Foundation of Educational Research (NFER) to study and evaluate the work at that time taking place in the United States. Burstall and Kay (1978) expressed their concerns about the pressure the American testing programme exerted upon teachers, pupils and the curriculum but still took the optimistic view that the technical difficulties associated with construction of the "new style " tests would soon be solved.

Back in the United Kingdom the NFER began to develop item banks using the technique known as the Rasch Model of analysis (Willmott & Fowles 1974). Under this model the chances of a pupil making a correct response to an item depend only on the pupil's ability and the item difficulty which are both measured on the same incremental scale. The NFER used as their unit the WIT so chosen that when a person's ability exceeded the item difficulty by five WITS there was a seventy-five percent chance that the item would be answered correctly. The typical pupil was said to make two and a half WITS progress per year of schooling.

There was immediate criticism of this approach (Goldstein & Blinkhorn 1977) but the same political considerations that operated in the United States prevented any slowing down of the APU's testing programme while the appropriateness of the RASCH technique for this type of monitoring exercise could be more fully evaluated. By the time the results of the first APU surveys in science, mathematics and language were available the NFER had quietly abandoned the use of the RASCH model. Instead the proportion of the population making a correct response or offering a particular explanation were reported. None of the APU surveys contain any detailed analysis relating to the validity or reliability of the items chosen.

Since many of those who successfully bid for the SEAC contracts to develop SATS at key stages one and two were previously involved with the APU it is not surprising that the approach, particularly in science bears striking similarities. Little information has been released about the selection of items. No cross-validation studies between the three trial consortia appear to have taken place before the final selection of the key stage one testing programme. The result, following presentation of the key stage one results, is public and media consternation that one third of pupils perform badly on certain test items while a further third do extremely well although the tests were designed to reproduce this distribution.

Yet, as the results presented here show, light can be shed on some important questions relating to the modification of the existing testing programme by the use of simple correlation analysis. While for some of the higher order process skills in science it would seem to make sense to abandon the elaborate practical testing programme in favour of a topic

based series of written exercises, similar to the Walled Garden and the earlier highly successful Prismaston File (Hargreaves 1989) a case exists for retaining practical activities designed to test pupils' powers of observation and measurement. Practical tasks would appear particularly necessary for slower learning pupils.

There are obvious reasons why a body such as SEAC should wish to retain the right to supervise the critical scrutiny of its data sets. History has sharp lessons to teach those who advocate a more open approach. Nearly two decades ago the Schools Council's attempt to establish the extent of "take up" of its materials in schools provided the justification for Mrs Thatcher to close it down and the eventual establishment, under centralised control, of the National Curriculum Council and SEAC. Now however the enormous impact which the present testing programme will have on future generations of pupils and the current stress levels it has caused in teachers demands a more open policy of sharing the data with researchers so that questions raised by this admittedly small scale study may be addressed more comprehensively.

However, if earlier history both in the United States and United Kingdom is a guide nothing of this kind will happen. The N.F.E.R and the associate publishing company, Nelsons, now have strong financial interests in maintaining the momentum of the testing programme since the large proportion of SEAC development contracts have gone to this consortium. N.F.E.R's report on the pilot study of key stage one materials rules out "post-hoc" moderation procedures in favour of "communicating standards in advance" (SEAC 1991). Tasks selected must relate to a single statement of attainment. Control of information provided and the extent of teacher prompting must also be severely restricted. Few teachers had time to carry out repeated measures procedures and no data was reported. The report notes a change to more grouping but fails to indicate this involved placing pupils of perceived similar attainment level together.

The NFER report offers scant evidence to support its methods. For example it claims that the wide variation in the levels recorded for individual children across English, mathematics and science shows that teachers are applying criterion rather than norm referenced procedures (p94). On validity the report appears to contradict itself. Having argued previously

that the level of performance should be independent of task content it is then stated (p190) when the results show that different combinations of science tasks yield different estimates of the pupil's level, that "it says much for the validity of the science assessment that PC2 reports are not simply a product of 'Ourselves' the science surrogate most often found in infant classrooms". Mathematics, however, is said to be curriculum valid because different combinations of task are consistent predictors of overall performance.

There are those who see the emphasis on written tests at key stage two as a

prelude to the re-introduction of selection once the bulk of schools have become grant maintained within a competitive system where money will follow pupils. With the demise of Local Education Authorities there is an urgent need to form new alliances between higher education and clusters of schools. Such an alliance could provide the more critical detailed analysis of test results that neither the government nor those commissioned to deliver the testing programme appear willing to contemplate.

REFERENCES

Burstall C. & Kay B. (1977) Assessment, the American Experience: APU Monograph. London, Department of Education and Science.

DES (1989) National Curriculum Task Group on Assessment and Testing: A Report. London, Department of Education and Science.

Galton, M. (1979) A Constructive Response to the APU, Forum, (22) 1, 20-23.

Goldstein, H. & Blinkhorn, S. (1977) Monitoring Educational Standards, An Inappropriate Model. Bull British Psychological Society, (30) 309-311.

Hargreaves, L. (1989) Small and Large Schools in Galton, M. & Patrick, H. (eds) Curriculum Provision in the Small Primary School, London. Routledge.

Harlen, W. et al., (1990) Assessing Science in the Primary Classroom: Written Tasks. London. Paul Chapman.

Muschamp, Y. et al., (1992) "Curriculum Management in Primary School" The Curriculum Journal (3) 1, 21-40

Russell, T. & Harlen, W. (1991) Assessing Science in the Primary School: Practical Tasks. London. Paul Chapman.

SEAC (1991) Pilot Study of Standard Assessment Tasks for Key Stage 1: A Report by STAIR Consortium. London Schools Examination and Assessment Council.
ä
Willmott, A & Fowles, D. (1974) Objective Interpretation of Test Performance. Slough, National Foundation of Educational Research.