Issues relating to the statutory assessment of technology at key stage 3 (1989-93)

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Additional Information:

- This is a conference paper.

Metadata Record: https://dspace.lboro.ac.uk/2134/1553

Publisher: © Loughborough University

Please cite the published version.
This item was submitted to Loughborough’s Institutional Repository by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/
Issues relating to the statutory assessment of technology at key stage 3 (1989-93)

Richard Tufnell
Reader in Design and Technology, Middlesex University

Abstract
During this four year period an agency based at Middlesex University was contracted by the Schools Examination and Assessment Council to develop statutory assessment procedures for both design and technology and information technology. This paper will reflect on some of the issues faced by the development team and aspects of the research undertaken in attempting to resolve these issues. The paper will explore the context for the development work with particular reference to the constraints imposed by political decisions and the problems created in attempting to examine a National Curriculum Order whilst many schools were still grappling with its implementation. The paper will focus on design and technology and deal with issues relating to both practical tasks and written tests.

The author was Director of the agency throughout the four year period. Participants should note that certain aspects relating to this research and development work remain confidential.

1989 - The context and scale of the task
This paper reviews some aspects of the research and development of KS3 statutory assessments in design and technology. This task provided a unique set of challenges as the assessment process was unlike any that had gone before and it is unlikely that it will ever be repeated in the same format. The development of the assessment process in this subject commenced only four months after the publication of the Statutory Order, four years before the first statutory assessment. At that time design and technology was enjoying a short-lived vision of Utopia. The value of the subject had at last been recognised, it was part of the extended core, every pupil would be studying the subject from 5 to 16, finally it seemed the 'Cinderella' status of the subject was over. The development teams believed they had the opportunity to establish purposeful assessment tasks, as advocated in the report of the Task Group on Assessment and Testing, in a practical process-based subject. These beliefs became more illusory the longer the process continued.

During the four years there was a continual shift in policy from a position which put the curriculum first, to one in which it was totally dominated by assessment. Political rhetoric attempted to convince the nation that assessment had been invented by the National Curriculum and that the 'novelty' of testing would elevate educational standards over night. There was an attempt to convince society that teachers paid scant regard to assessment. However, traditionally teachers have always assessed pupils for a wide range of educational purposes: diagnostic, evaluative, predictive, selective, comparative, motivational and informative. It quickly became apparent that alongside these well accepted objectives, end of key stage statutory assessment would primarily be a quality assurance tool, an advice which would allow comparative league tables to show school performance by subject and of course, if required, teacher performance. An norm referenced instrument imposed on a criterion referenced system.

Devising a standard and reliable assessment system across the breadth of design and technology would have been sufficient challenge, without the constant redefinition of the specification. As a Director of a team, I viewed the task as a design consultant might: we had a client, there were certainly customers, albeit reluctant ones, and there was a specification, although it did not reveal the true agenda. The scale of the exercise was daunting; up to 650,000 pupils spending twelve hours on a practical task and up to one and a half hours on a written test; and approximately 30,000 teachers delivering the materials and assessing the outcomes; the practical resources involved defy quantification. Any national assessment procedure will involve this scale of operation. Key stage 4, GCSE, will be facing similar problems in 1995, although the task will be divided between the five examination groups. National Curriculum assessment attempted to be more transparent than any previous system. For all its faults, the system attempted to establish a new approach. Teachers would also act as examiners, standards would be established before the tasks or tests took place (as must be the case with any criterion referenced system) and schools were responsible for moderation and standardisation. Such a system does not allow performances to be calibrated and grade boundaries or levels set after the event. Following a lengthy selection procedure, The School Examining and Assessment Council
(SEAC), established contracts with two development agencies; both based in Institutions of Higher Education. For two years these two teams, along with four others working in other subjects, collaborated jointly on matters of principle but worked independently on specific approaches. Jointly the agencies formed a powerful cabal and could, when required, present a united front.

The guiding rule book for all the development teams was the report of the Task Group on Assessment and Testing: This report was embraced by both politicians and educationalists. Politicians were beguiled because they believed it would produce a system with performance indicators which could be monitored and improved via accountability; Education could be regulated in common with other public services and the expenditure of taxes justified or curtailed as deemed appropriate. Whilst educationalists believed that TGAT steered a middle road in relation to accountability as it talked of: ‘assessment tasks which exploit a wide range (far wider than normally envisaged for tests) of modes of presentation, operation and response, and their numerous combinations, in order to widen the range of pupils’ abilities that they reflect and so enhance educational validity.’

Every development team felt comfortable with this approach, classroom tasks used by teachers as and when appropriate which would produce both formative and summative assessment information. Both teams working in design and technology interpreted this as an opportunity to influence the curriculum, to demonstrate good practice, to translate the ideals of the Statutory Order into an educational reality in addition to developing assessment instruments.

1991 - The specification changes

As developments and trialling took place it became clear that both ministers and teachers were unhappy with what was emerging, but for different reasons. Initially the debate centred on key stage one. The burdensome attempt to assess each pupil in relation to each statement of attainment in each of the core subjects was untenable. It resulted in a level of classroom insurrection which sent immediate alarm signals to the Government. The teachers cry for a reduction in workload, allowed ministers to reassess the strategies they were employing. It was too soon to admit the unthinkable, that the subject Orders might be too complex, that admission would not come for another two years. Ministers were also too myopic to make the obvious linkage that if the Orders are complex and prescriptive then logically the assessment strategies will also fit that description. A scapegoat, or better still a herd of them, was needed and, even though the controversy was centred on KS1 and the work of some of the KS3 agencies had not officially been in the public domain, all contracts were terminated. Kenneth Clarke, the recently appointed Secretary of State stated in his first interview, with The Times in October 1991: ‘I thought ‘tasks’ was a typing error for tests. I propose to go on using the word tests.’ A new phase in the development process had been reached. Fresh contracts were placed. Ministers now indicated that the experience and expertise of the examining groups would be essential to accomplish the task. Their intentions had been sabotaged by University based research groups. However, to ensure success the specifications were also revised. These no longer referred to TGAT, they were self-contained and prescriptive. In no subject would there be more than a single agency and the formal structure within which joint issues were dealt with was dismantled. From now on the objective was clear - to develop a summative assessment system the principal purpose of which would be nomothetic (collect data to generalise, classify and rank) in nature.

The task was now very different. Despite the prescriptive specification, modifications were necessary several times in the first few months. Initially, the requirement which appeared most alien to the spirit of the Order focused on the demand for tasks related to separate materials, the literal interpretation of page 19 of the Statutory Order. In the event the necessity for tests impinged on the philosophy of the subject. On the basis of a SEAC recommendation the Department for Education laid an Assessment Order which effectively deconstructed design and technology capability. Two attainment targets, Te 1 and 4, were to be assessed via a short written test and the other two, Te 2 and Te 3, would be assessed via a task. The specification also indicated that differentiation should be achieved by task. In other subjects, such as mathematics, this approach was and is not unusual. In design and technology this approach was and is still unusual, although in an assessment context it has much to recommend it. The majority of design and technology teachers adopt a more pragmatic approach and assess by outcome; pupils essentially tackle the same task and the quality of the response is judged, frequently by comparison. I should add that during the first two years both development groups developed procedures for task negotiation so that pupil and teacher jointly agreed both on content and complexity, but this style of approach was clearly no longer on the agenda. Differentiation by task meant that teachers needed to know each pupil’s level of achievement so that they could be entered in the appropriate band; this was clearly not the case in the majority of schools. Additionally, tasks had to be
defined which focused on specific levels so that, once entered, a ceiling would be placed on the level of performance. Teachers particularly objected to the ‘ceiling of performance’ concept but without sound reasons. Each band covered four levels, each National Curriculum level defined progression over a two year period, consequently each band covered theoretically eight years of achievement and there was also a two level overlap between each band. If schools had been able to operationalise the Order and produce sound teacher assessments, entry would not have created a problem. Differentiation by task results in assessment procedures which tend to confirm achievement rather than reveal it.

The development of practical tasks - Te2 & Te3

The development of constructs for both tasks and tests which would meet both the specification and teachers’ requirement was impossible. Clearly teachers and pupils were more concerned, and rightly so, with the content but this paper is focused on the structure into which the content was placed. Firstly the practical task, there was pressure to devise different tasks for each of the bands of entry; this would have resulted in pupils in the same class tackling very different tasks. This approach was resisted by the agency. As the task effectively commenced at Te 2 pupils would have to be presented with a product design specification (PDS), the challenge would be to meet this set of requirements and produce a design specification (DS) - Te 2 and then manufacture a prototype - Te 3. Some would prefer to call this problem solving rather than designing and indeed that might be the case but it so then this is the kind of task with which many practising designers are confronted. The agency devised a system which allowed all pupils to tackle the same basic task with the product design specification increased in complexity for every band of entry, see diagram 1. Hence the behavioural criteria, the statements of attainment, could be combined with outcome goals derived from the programme of study at each level to provide assessment based both on process and product.

The major benefit from defining the complexity of the task is that the assessment criteria can be contextualised. It is interesting to note that although teachers overwhelmingly approved of contextualised assessment criteria they disapproved of prescribed tasks - two statements which are incompatible! From the earliest development in 1989 both design and technology agencies had been powerful advocates of the concept of ‘levelness’. It was during the first contract that we received a ‘yellow card’ for deconstructing the statements of attainment to produce ‘level descriptors’. These we trialled in both 1990 and 1991. In 1992 the concept was proscribed but we continued to promote the notion on two grounds, manageability and reliability. Our research had shown that teacher judgements were more reliable at the level of the attainment target than at the level of the statement of attainment. Assessment focused on the AT also removed the need to evolve a complex aggregation procedure to achieve an attainment target level. The ‘levelness’ approach in the design and technology practical task was the first example in statutory assessment where the statements of attainment were subsumed and interpreted into task-specific criteria - an approach now adopted by the proposed/revised Order. Diagram 2, below, illustrates the process involved.
in constructing the assessment criteria. Developing written tests - assessing Te1 and Te4

A construct for a short written test presented a different set of issues. The requirement for a linkage between the task and test fundamentally undermined the requirement related to manageability. Twelve different test papers and associated marking schemes presented teachers with genuine logistical problems. Increasing the number of practical tasks would have exacerbated the situation. If Te 4 had been part of the practical task it would have been purposeful and a common test could have been set for Te 1. Instead, an observation made by Jerome Bruner sums up the situation in relation to Te 4:

‘Leaving evaluation to the final stage is rather like doing military intelligence after the war is over’2

A test of two attainment targets also resulted in a superficial assessment of Te 1 and 4, one based on either 30 or 45 minutes assessment time, less than any other attainment target in any other subject. Tests had to conform to a common specification, consequently, questions had to identify the statements of attainment which they addressed. The notion of ‘levelness’, therefore, had to be defined by a different process: marks were allocated to each question (each mark point to be identified in the mark scheme); marks accumulated at a level could not be transferred between levels; at each level a mastery level would be set and the highest level at which mastery was achieved would be the level recorded. Some subjects’ developers chose to vary both the marks available at each level and the level of mastery. In design and technology an uncomplicated approach was adopted. Test papers were designed which had the same number of marks at each level and, if they could be accurately calibrated, the same mastery level could be used. This approach required controlled trialling and modification as the mastery level had to be set prior to the test taking place. National Curriculum testing did not enjoy the luxury of post test calibration and the setting of boundaries. In design and technology there was also the added complication of standardising across tests. Continuous changes to the specification and the pressure of scrutiny and approval resulted in a three month window for writing, trialling and modification and this was further complicated as pupils had first to complete the practical task before they could trial the test papers.

A vital aspect of the written tests, given that the highest level achieved would be the level recorded, was that the questions attached to the levels should represent a hierarchical progression. This assumed that the statements of attainment represented a progressive and even demand; this was clearly open to question. But then it is unlikely that the Architects of these statements appreciated the purposes to which they would be put. Two approaches for evaluating the construct validity of the tests were employed, diagram 3 illustrates the first.

A test of two attainment targets also resulted in a superficial assessment of Te 1 and 4, one based on either 30 or 45 minutes assessment time, less than any other attainment target in any other subject. Tests had to conform to a common specification, consequently, questions had to identify the statements of attainment which they addressed. The notion of ‘levelness’, therefore, had to be defined by a different process: marks were allocated to each question (each mark point to be identified in the mark scheme); marks accumulated at a level could not be transferred between levels; at each level a mastery level would be set and the highest level at which mastery was achieved would be the level recorded. Some subjects’ developers chose to vary both the marks available at each level and the level of mastery. In design and technology an uncomplicated approach was adopted. Test papers were designed which had the same number of marks at each level and, if they could be accurately calibrated, the same mastery level could be used. This approach required controlled trialling and modification as the mastery level had to be set prior to the test taking place. National Curriculum testing did not enjoy the luxury of post test calibration and the setting of boundaries. In design and technology there was also the added complication of standardising across tests. Continuous changes to the specification and the pressure of scrutiny and approval resulted in a three month window for writing, trialling and modification and this was further complicated as pupils had first to complete the practical task before they could trial the test papers.

A vital aspect of the written tests, given that the highest level achieved would be the level recorded, was that the questions attached to the levels should represent a hierarchical progression. This assumed that the statements of attainment represented a progressive and even demand; this was clearly open to question. But then it is unlikely that the Architects of these statements appreciated the purposes to which they would be put. Two approaches for evaluating the construct validity of the tests were employed, diagram 3 illustrates the first.

A test of two attainment targets also resulted in a superficial assessment of Te 1 and 4, one based on either 30 or 45 minutes assessment time, less than any other attainment target in any other subject. Tests had to conform to a common specification, consequently, questions had to identify the statements of attainment which they addressed. The notion of ‘levelness’, therefore, had to be defined by a different process: marks were allocated to each question (each mark point to be identified in the mark scheme); marks accumulated at a level could not be transferred between levels; at each level a mastery level would be set and the highest level at which mastery was achieved would be the level recorded. Some subjects’ developers chose to vary both the marks available at each level and the level of mastery. In design and technology an uncomplicated approach was adopted. Test papers were designed which had the same number of marks at each level and, if they could be accurately calibrated, the same mastery level could be used. This approach required controlled trialling and modification as the mastery level had to be set prior to the test taking place. National Curriculum testing did not enjoy the luxury of post test calibration and the setting of boundaries. In design and technology there was also the added complication of standardising across tests. Continuous changes to the specification and the pressure of scrutiny and approval resulted in a three month window for writing, trialling and modification and this was further complicated as pupils had first to complete the practical task before they could trial the test papers.

A vital aspect of the written tests, given that the highest level achieved would be the level recorded, was that the questions attached to the levels should represent a hierarchical progression. This assumed that the statements of attainment represented a progressive and even demand; this was clearly open to question. But then it is unlikely that the Architects of these statements appreciated the purposes to which they would be put. Two approaches for evaluating the construct validity of the tests were employed, diagram 3 illustrates the first.
If the questions provided a valid incline of difficulty, it would be correct to assume that level 6 questions should be answered correctly by the highest percentage and level 10 by the lowest percentage, this would result in the ranking offered as a model. The ranking level could also be correlated. For example, using Spearman and correlating with the construct model, from this data a significant level of 0.725 was obtained.

Teams of expert scrutineers were also used to assess the descriptive validity of the questions. For the statements assessed by the above test paper the ratings on a five point ordinal response (4 - excellent, 1 - poor) are shown in diagram 5 (test 2 and test 3).

<table>
<thead>
<tr>
<th>Statement</th>
<th>Test 1 Rating</th>
<th>Test 2 Rating</th>
<th>Test 3 Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Te 4.1a</td>
<td>3.4</td>
<td>2.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Te 4.2a</td>
<td>3</td>
<td>4.4</td>
<td>3</td>
</tr>
<tr>
<td>Te 4.1a</td>
<td>2.4</td>
<td>4.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Te 1.1b</td>
<td>2</td>
<td>4.6</td>
<td>4</td>
</tr>
<tr>
<td>Te 1.2b</td>
<td>2.2</td>
<td>4.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Te 1.2a</td>
<td>3</td>
<td>4.3</td>
<td>3.6</td>
</tr>
<tr>
<td>Te 4.1b</td>
<td>3.4</td>
<td>3.6</td>
<td>4</td>
</tr>
<tr>
<td>Te 4.2b</td>
<td>3.6</td>
<td>3.6</td>
<td>3.4</td>
</tr>
</tbody>
</table>

• no question was given a 0 rating by any of the scrutineers;
• only six ratings of 1 were awarded - 3% of all possible ratings;
• only one question received two ratings of 1 - Te 1.1b;
• only two questions did not receive a single rating of 4 - Te 1.7b and Te 4.10a;
• the average rating across all questions was 3.15.

Other scrutineers had to match the question to the statement it was assessing, without knowing either the attainment target or the level. Of the forty questions on the three test papers 90% of questions, on average per scrutineer, were correctly matched. The main purpose of this procedure was to evolve an assessment instrument which effectively differentiated performance. However, written tests are not the ideal means of assessing an activity defined by criteria as practical. Assessment undertaken in this fashion is bound to be a compromise as other capabilities will undoubtedly disguise or enhance design and technology capability.

Lessons for the future

In the future it is likely that statutory assessment of design and technology will once again take place. Currently, it appears that statutory teacher assessment is the favoured approach; although some believe the benefits, in relation to status, would make the anguish of end of key stage assessment worth bearing. Whatever evolves, three important lessons must be learnt. Firstly different subjects cannot be treated in the same way, an assessment procedure which is appropriate for one subject could well be the antithesis of good practice in another. Secondly within each subject assessment techniques should be selected which are appropriate to the subject and the purpose of the assessment. For example, in design and technology level descriptors, similar to those used in the long task, would provide a more efficient and reliable assessment tool. Thirdly if National Curriculum Assessment is to have integrity it should assess the curriculum as prescribed and interpreted, the rules cannot be changed once they have been established. These lessons appear obvious to educationalists, but it is apparent that others needed some convincing.

References