TESTING THE TEST

A Study of the Reliability and Validity of the Northern Ireland Transfer Procedure Test in Enabling the Selection of Pupils for Grammar School Places

John Gardner and Pamela Cowan

Graduate School of Education
The Queen's University of Belfast

April 2000

ACKNOWLEDGEMENTS

This study would not have been possible without the co-operation and goodwill of the schools involved. Our sincere appreciation goes to them. We would also like to acknowledge gratefully the support and input of colleagues who assisted with the work on various occasions. Finally, we would acknowledge the courtesy and co-operation – to the extent that the conditions of their responsibility for the administration of the Transfer Test allowed - of the Northern Ireland Council for Curriculum, Examinations and Assessment.

Communications:

Professor John GARDNER

Head of School
Graduate School of Education
The Queen's University of Belfast
69 University Street
BELFAST BT7 1HL
Northern Ireland

Tel: *UK - (0)28 90 335 929*
Fax: *UK - (0)28 90 239 263*
E-mail: *j.gardner@qub.ac.uk*

TESTING THE TEST

This report is in three parts:

- **SUMMARY OF KEY POINTS AND FINDINGS**
- **THE REPORT** (*for those not expert in assessment issues*)
- **TECHNICAL REPORT** (*for those interested in the more technical aspects of the research*)

Please note:

To ensure each part can be read on its own, a level of repetition is necessary. The reader's indulgence is therefore requested.

TABLE OF CONTENTS

TESTING THE TEST

A Study of the Reliability and Validity of the Northern Ireland Transfer Procedure Test in Enabling the Selection of Pupils for Grammar School Places

SUMMARY OF KEY POINTS AND FINDINGS

KEY POINTS

The Status Quo

- The Transfer Procedure Test is taken by children of around 11 years of age who wish to attend grammar schools in Northern Ireland.
- The Test is specified by the Department of Education (DE) and administered on their behalf by the Council for the Curriculum, Examinations and Assessment (CCEA).
- The Department of Education specifies the subjects and subject content to be tested, the number of marks per subject, time allowed for sitting the Test, the format of the papers, the dates of the tests and the grading system.
- The Test is known as a 'high stakes' test inasmuch as the children who take it are only allowed one attempt. Their performance in the Test can also determine their future education in a manner that is not of their choice or their parents.
- Candidates are required to sit two tests, normally Test 1 and Test 2. A Supplementary Test is available for candidates who for any approved reason would not have a mark from one of the two main tests in their final score.
- Each test comprises sections on mathematics, English and science & technology[1] with 75 marks available in the proportions: 26 for each of mathematics and English, and 23 for science.
- The scores for each subject in each test are adjusted for age and then standardized. Weightings (0.35 for mathematics and English, 0.30 for science) are applied to the scores and they are aggregated to provide a final single score.
- The single score arising from two Test sittings is used to place the candidates in a rank order, which in turn enables the candidates to be awarded grades in the range: A, B1, B2, C1, C2 and D.
- The single score suggests that the Test measures a single attribute of the candidates but no information is published on what attribute the Test is designed to measure. It is known only that questions selected for the Test are based on the Key Stage 2 programmes of study in mathematics, English, science and technology.
- No information in the form of the children's total or subject scores (i.e. in mathematics, science etc.) is made available to schools, parents or the children themselves.

[1] Subsequent references to science alone should be taken to include technology.

- No information on the children's placing in the rank order of scores is made available to schools, parents or the children themselves.
- No information on the reliability of the Test is made available to the public.
- No information on the validity of any conclusions or consequences derived from the Test scores is made available to the public.
- No information on the extent of adjustment made in relation to a candidate's age is made available to schools, parents or the children themselves.
- With no information on scores, grammar schools faced with more applicants (of the same grade) than they have places for, must apply other criteria to allocate their places.

The Study

- This report covers the largest independent study of the Northern Ireland Transfer Procedure Test ever conducted.
- Samples of Test scripts, used as practice tests by 52 primary schools, were analysed to assess whether the Test functions effectively in enabling the selection of pupils for grammar school places.
- Three tests formed the basis of the study: the 1998/99 versions of Test 1, Test 2 and the Supplementary Test.
- The pupils involved were all in their final year of primary school and about to take the 1999/2000 Test.
- Large samples were used for the analysis and these comprised 1288 candidates (Test 1), 1270 (Test 2) and 623 (Supplementary Test). Combining Test 1 and Test 2 scores is the normal procedure for the Transfer Test and the samples yielded a combined group of 926 candidates who took both tests.

KEY FINDINGS

Based on the 1998/1999 Test papers and the large samples used in this study:

- The Test does not measure a singular attribute of the candidate; it measures three: mathematics, English and science. In the same manner that it would normally not be good practice to add the marks from GCSE mathematics, English and science tests, their addition in the Transfer Procedure Test is questionable.
- Since the Test does not measure a singular attribute of candidates, it cannot be used as a proxy for any particular attribute, for example children's *ability* or their *potential to benefit from a grammar school education*. The common perception that it does provide some such measure cannot be substantiated by the research.
- The Tests would be perceived as 'easy' by many pupils since more than 65% of them answered over 70% of the questions correctly.
- The 'easiness' of the Test is a serious design flaw as children would have been awarded a D grade with 70% of the available marks. It is difficult to justify how a perceived 'fail' grade, a D, can be awarded to children who have done so well.
- With such an 'easy' test format, it is very likely that the children will know or will at least have a sense of how well they did. The consequences for their self-confidence, of being awarded a D for such high scores, has not been assessed in this study but must be considered a serious issue.

- There is evidence that the science section of the Test contributes significantly to the 'easiness'. The average science score for the three tests studied was 19 out of 23 i.e. 83% correct compared to 70% for the mathematics and English sections.
- The lower weighting and relatively high average score in science can lead to disadvantage for those who have relatively low scores in the science sections. Despite having the same overall Test score to begin with, candidates with low science scores may end up with lower final scores (after age adjustment, standardization and weighting) than candidates who score relatively more in mathematics and/or English.
- The three tests were found to be highly reliable, averaging 0.90 against a maximum possible of 1.00. However, examination of the Standard Error of Measurement for the combined sample for Test1 and Test 2 indicates that a child's true score[2], with 95% confidence, will lie somewhere between 10 marks above or below their Test score.
- The Test works effectively to identify 12% of the candidates as secure A grade candidates (scaling up to 2,053 children in the total cohort) and 18% as secure D grade candidates (3,099 children). Its capacity to allocate grades accurately to children whose scores lie between the score ranges of these groups is highly questionable.
- The boundaries between the six Test grades (A, B1, B2, C1, C2 and D) cover only 18 marks out of a total of 150. Within the 95% confidence range (10 marks above or below the Test score) the Test therefore has the potential to misclassify pupils by up to three grades above or below their given grade.
- For example, 23% of the candidates (scaling up to 4,487 children in the total cohort) have A grade scores up to 10 marks above the score at the A/B1 boundary. Their true grade could be an A or depending on how close they are to the boundary, it could be any grade down to a C1. Similarly 12% of the candidates (2,148 children) have D grade scores up to 10 marks below the score at the C2/D boundary. Their true grade could be a D or depending on how close they are to the boundary, it could be C2 or a C1. Finally, 33% of the candidates (5,818 children) have scores between and including the A and D grade boundaries. Grades in the middle of the range, e.g. C1, could be any grade up to A or down to D.
- No consistent pattern of significant differences was found in the mean scores of the groups of candidates from the different education and library board areas.
- The mean scores of the preparatory school candidates were significantly higher than any of the other groups characterized by their school management types. There was no significant difference between the mean scores of the groups of candidates from the two main primary school sectors: controlled and Catholic maintained.
- The mean score of candidates from schools with high free school meal (FSM) entitlements (51%+) was significantly lower than the mean scores of groups from the other lower FSM entitlement (<10, 11-30 and 31-50%). The mean score of the group of candidates in schools with <10% FSM entitlement was also found to be significantly higher than those of other categories for Test 1 and Test 2.
- There were no significant differences in the means of the groups of boys and girls and in the means of the groups of younger and older pupils. No significant differences were

[2] The 'true score' is the score that would be obtained if any errors inherent in a single sitting, e.g. arising from distractions, ill-health, undue stress etc., were removed through multiple sittings. It is an internationally accepted convention for determining the confidence to be placed in raw test scores.

found for the means of candidates in the groups of schools with different enrolment levels (i.e. school size).

- The published information on the Test does not meet the requirements of the international standards on educational testing, both generally in the provision of standard reliability and validity information and particularly, for example, in the validation of the Test outcomes in relation to its predictive power (e.g. 'potential to benefit from a grammar school education'), establishing norms, providing information on potential misclassification, and accommodating disability.

TESTING THE TEST

A Study of the Reliability and Validity of the Northern Ireland Transfer Procedure Test in Enabling the Selection of Pupils for Grammar School Places

THE REPORT

TESTING THE TEST

THE REPORT

INTRODUCTION

At the outset, it is very important to emphasize that this report considers the effectiveness of the current Transfer Test in assisting grammar schools to allocate their quota of enrolment places. It takes no view on the broader issues surrounding the debate on 'selection', nor does it seek in any manner to undermine the role of grammar schools in Northern Ireland's education system. What it does is explained simply:

> Given that there is a selective system, and a test to assist the selections being made, this report investigates whether the Transfer Procedure Test used does its job reliably and fairly.

The answer is also simply stated. It does not.

In the largest independent study ever of the Transfer Test, this report provides empirical evidence of major weaknesses in the Test's capacity to differentiate candidates for grammar school entry. The report is offered in three sections: a *Summary of Key Points and Findings*, *The Report* and the *Technical Report*. To enable them to stand-alone as individual papers, some repetition of the most important aspects of the context and findings is necessary in the three sections. The reader's indulgence is therefore requested.

PREAMBLE

In June 1993, a concerned parent wrote to the incoming 'Direct Rule' minister, Michael Ancram, to express concern about the proposed new model for the Transfer Test. The letter set out the theoretical position that any test covering English, mathematics and science together could not be fair and reliable in enabling the allocation of grammar school places to transferring pupils. The argument was simple. Pupils who perform well in mathematics may perform less well in English, and vice versa. Science may confound the issue further. Aside from the very able candidates who score well on all of these areas, and the very weak candidates who score poorly on all of them, most candidates will have similar scores. As their high and low performances cancel each other, the scores of most candidates settle around an average score. How then can the Test reliably split the candidates into the different grades?

The response, some months later, argued simplistically that the Test was fair because *"... no individual should ... be at any disadvantage ... since the change applies to all children equally"*. This logic proposes that any manifestly unfair and unreliable test may be rendered acceptable merely by exposing everyone to it. The new Test, the response continued, *"... reflects the curriculum that is being studied by all children"* and removes the problem whereby *"... practice for the tests took time away from the normal programme of primary school work"*. The assumption that all children are exposed to the same content and quality of teaching and learning experience, under the common curriculum, is simply untenable. Seven

years on, many would consider that the problem of test practice disrupting the last year of primary school also remains undiminished.

Information on Test Performance

Given the lack of information around the Test, common sense would suggest that more evidence is needed of the Transfer Test's ability to do its job. Therein lies another problem. Anyone attempting to exercise their right to receive a report on their child's *"... results in any assessment and public examination which he or she has undertaken during the year"* (*The Parents' Charter for Northern Ireland*, DENI, 1992) will find that the school can only tell them a letter grade. How did the child perform in the science sections of the Transfer Test? No answer. By how much did the child miss out on getting an A? No answer. The lack of answers is no criticism of schools; they are as much in the dark about the details of their pupils' performances as anyone else. Anyone else, that is, who is not involved in the development and administration of the tests. What sort of information should parents have, then?

Test Accuracy

One piece of information is the Test's accuracy in predicting candidates' performance later in secondary-level schooling. In 1989, in advance of the inception of the Northern Ireland Curriculum, Anne Sutherland at Queen's reviewed the effectiveness of the various tests that had been used in Northern Ireland since 1947 (Sutherland, 1990). The evidence she found suggested that at best 1-in-7 and at worst 1-in-5 candidates were misplaced by the tests. In an average cohort of 18,000 pupils taking today's Test this would amount to somewhere between 2,600 and 3,600 individuals. This study does not, however, attempt to update Sutherland's work by looking at the predictive accuracy of the present Test (introduced in 1993/1994). However, the principle of what she called the Test's *"... accuracy in identifying able pupils at age 11"* is addressed from another angle: its ability to rank order candidates according to the construct it measures.

Achievement vs. Ability

Note that we use here the word 'achievement' instead of 'ability'. It must be clearly understood that the Transfer Test's construction relates it directly to the curriculum taught. It is important to note though, that while a child's level of achievement is clearly linked to ability, it is also subject to what might be called 'environmental' factors. Many environmental questions may be asked. For example: *Did the child's class cover all of the necessary curriculum? Was their teacher fully trained in primary science education? Did their class have sufficient resources to cover the curriculum? Did they experience any significant disruptions in the teaching they received?*

Such questions are very important in relation to the present form of the Transfer Test because it is essentially an *achievement* test; the marks awarded are largely a measure of what the children know of the curriculum areas tested. If the children have never been 'taught' the names of the parts of a flower, for example, or if they cannot remember them, they get the answer wrong. An *ability* test, on the other hand, avoids testing memory or detailed subject content and instead tests the children's reasoning powers. Prior to the current model of the Transfer Test, general reasoning or 'intelligence' tests were used directly to gauge the children's ability. Since these tests effectively measure one thing, something which might be

called ability or intelligence, they are considered more appropriate for putting the candidates in a rank order. If a test can rank order the candidates properly, deciding who to select for any purpose will be relatively easy. This aspect of the Transfer Test is investigated in this study.

Openness

Anyone coming upon the Transfer Procedure Test for the first time will be struck by the lack of information about it in the public domain. Two questions will quickly come into focus:

- "Why is information not made available about the Transfer Test's performance in reliably grading candidates?"; and
- "Why is information not made available about an individual's performance in all aspects of the Transfer Test?"

Down through the years, the processes surrounding the various selection tests and procedures have been shrouded in secrecy, and this has been the case despite the high stakes involved. It is not clear why this has been so but doubt about the effectiveness of the various tests must at least be an element in prompting secrecy. Nevertheless it is argued that many of those who take these tests are being confronted with what to them is a momentous decision point in their lives, the choice of school for their secondary-level education. They (and their parents) are therefore entitled to know and be reassured by the provision of appropriate information, which gives them details of the Test's performance and their own performance in it.

Elsewhere in the world, parents' and candidates' rights can be defended by litigation and are enshrined in procedures for good practice. Central to all good practice is the principle of openness. The yardstick for test developers and administrators around the world is the set of standards for educational and psychological testing, which the American Educational Research Association, American Psychological Association and the (US) National Council on Measurement in Education developed in 1985 and have updated recently in 1999. Openness threads through the standards to protect test developers, test users (administering bodies) and test takers and it is important to assess the extent to which the Transfer Test procedure meets them. This aspect of the Transfer Test is investigated in this study.

THE CONTEXT OF THE TRANSFER PROCEDURE TEST

Each year in Northern Ireland there are around 26,000 pupils getting ready to leave primary schools to go on to secondary education. And each year, some 18,000 of them may be expected to take the Transfer Procedure Test. This test is designed to assist grammar schools in allocating the fixed number of places they are allowed to offer to new first year pupils. Inevitably, however, there are not enough places for everyone who wants to go to a grammar school and approximately 60-70% will not be offered a place. For many of these children, a sense of failure adds to the personal disappointment of not getting a place in the school of their choice, or the school where their brothers and sisters may already attend.

In getting to this point, which many of them will think of as 'failure', the children are allowed only one 'go' at the Test. The consequences are far-reaching and irreversible as the grade they get may simply remove them from any chance of a place in the school of their and their parents' choice. Despite the efforts of the schools that they do subsequently attend, some children may never regain their confidence or may never overcome the sense of having failed to meet their parents' or their own expectations. With its serious consequences, at the level of

the individual child and his or her family, the Test therefore belongs to a class known as *'high stakes'* tests. Indeed it may be considered one of the highest stakes tests to be conducted with children by government agencies anywhere in the UK and further afield.

Clearly, then, it is important to ensure that the Test performs properly. The central question is: *"Does the Test enable schools reliably and fairly to pick their new pupils and reject others?"* International standards in testing would normally guarantee that the information needed to answer this question would be in the public domain. Not so the Transfer Test. The whole process is shrouded in secrecy and very little information is made available to the public[3]. For example, parents and schools have no access to the scores the children get or to the scores that attract particular grades.

In the absence of any official information, the answer to the question must be found by independent research. This study therefore examines how the Transfer Procedure Test stands up to technical scrutiny of its performance in assisting the allocation of grammar school places. It is argued that the selection process should be carried out with the utmost fairness, reliability and openness.

THE TRANSFER PROCEDURE

The Transfer Procedure Test is taken by children of around 11 years of age who wish to attend grammar schools in Northern Ireland. If pupils want to go to a grammar school, or to one of the small number of non-selective schools which are permitted to take in pupils for a 'grammar stream', they must take the Transfer Procedure Test unless they can plead that they have 'special circumstances'. Schools, however, cannot exceed their quota of places and in selecting their new pupils, they must take them in the order of their Transfer Test grades i.e. A before B1, B1 before B2 etc.

The Test is specified by the Department of Education (DE) and administered on their behalf by the Council for the Curriculum, Examinations and Assessment (CCEA). The Department of Education specifies the subjects and subject content to be tested, the number of marks per subject, time allowed for sitting the Test, the format of the papers, the dates of the tests and the grading system.

The Test is known as a 'high stakes' test inasmuch as the children who take it are only allowed one attempt. Its consequences may also determine their future education in a manner that is not of their choice or their parents. Candidates are required to sit two tests, normally Test 1 and Test 2. A Supplementary Test is available for candidates who for any approved reason (e.g. absence through illness) will not have a mark from one of the two main tests included in their final score.

The Test comprises sections on mathematics, English and science & technology[4] with 75 marks available in the proportions: 26 for each of mathematics and English, and 23 for

[3] In 1996 the Department of Education did publish a bulletin on the two types of tests used from 1989 to 1995 but this did not provide inferential statistics or information about the tests' reliability or validity: DENI (1996) *Transfer Test Results 1989/90-1995/96*. Statistical Bulletin, SB1/96 Department of Education for Northern Ireland, Bangor Co. Down

[4] Subsequent references to science alone should be taken to include technology.

science. The scores for each subject in each test are adjusted for age and standardized before weightings (0.35 for mathematics and English, 0.30 for science) are applied to all six standardized scores. These are then aggregated to provide a final single score.

The single score arising from two Test sittings is used to place the candidates in a rank order, which in turn enables the candidates to be awarded grades (A, B1, B2, C1, C2 and D). The single score suggests that the Test measures a single attribute of the candidates but no information is published on what attribute the Test is designed to measure. It is known only that questions selected for the Test are based on the Key Stage 2 programmes of study in mathematics, English, science and technology.

Very little information is available to the public or to the test-takers (the children), their parents or indeed the primary schools they attend or the secondary-level schools they wish to attend. For example:

- No information in the form of the children's total or subject scores (i.e. mathematics, science etc.) is made available to schools, parents or the children themselves.
- No information on the children's placing in the rank order of scores is made available to schools, parents or the children themselves.
- No information on the reliability of the Test is made available to the public.
- No information on the validity of any conclusions or consequences derived from the Test scores is made available to the public.
- No information on the extent of adjustment made in relation to a candidate's age is made available to schools, parents or the children themselves.

With no information on scores, grammar schools faced with more applicants (of the same grade) than they have places for, must apply other criteria to allocate their places. No grade, therefore, can guarantee a place. For example, if there are less places in a school than A grade applicants, some of the applicants will have to be rejected. In such cases, the school must use 'objective' criteria (e.g. the distance the child lives from the school) to allocate their places. Clearly this process could be carried out on academic grounds if the schools were able to use the children's scores or their rank order within the grade bands. However schools are not given this information; they know only that an A has been awarded. Reasons as to why they are not given this information are not published.

How are the Grades Allocated?

The quotas for each grade A to D are pre-set by the Department of Education in the following proportions:

> Grade A is awarded to the top 25% of the entire age group eligible to sit the tests, B1 is awarded to the next 5% of the pupils, B2 to the next 5%, C1 to the next 5%, C2 to the next 5% and D to those remaining[5].

In the school year 1999/2000 figures obtainable from CCEA show that there are 25,727 pupils in Primary 7. This means that the 6,432 (25% of 25,727) highest scoring candidates in

[5] CCEA (1998) *Specification of the 1999/2000 Transfer Tests*. Northern Ireland Council for the Curriculum, Examinations and Assessment, Belfast

the Test were to be given an A grade. Similarly the next 1,286 (5% of 25,727) highest scoring candidates were to be awarded a B1. This process can be repeated for B2 (5%), C1 (5%), C2 (5%) and D (the remainder).

It is not clear why the percentage quotas are referenced to the eligible population (i.e. all those in their last year of primary school) instead of the population of Test entrants. Since the final allocation of places is governed by each school's fixed entry quota, and since the Test grade itself cannot guarantee acceptance or rejection by a school, there seems no reason artificially to create a situation in which it is only the A grades that have a realistic chance of entry to many grammar schools. Using the 25, 5, 5 etc. percentages directly to allocate grades among the Test entry population would mean that the spread of grades gaining entry to grammar schools would be increased as more B1's, B2's etc (which would previously have been A's) are allocated the places.

With 17,606 pupils actually entering for the Test in 1999/2000, these 'quotas' translate into the projected percentages of the 'entrant' population shown in bold in Table 1:

Table 1 Numbers of candidates awarded grades A to D in 1999/2000 (17,606 entrants)

	GRADE					
	A	**B1**	**B2**	**C1**	**C2**	**D**
% of Eligible Population (25,762)	25	5	5	5	5	Remainder
Projected No. with each Grade	6432	1286	1286	1286	1286	6030
Projected % of Entrants	**36.5**	**7.3**	**7.3**	**7.3**	**7.3**	**34.2**
Actual No. with each Grade[6]	6633	1416	1335	1456	1333	5433
Actual % of Entrants	37.7	8.0	7.6	8.3	7.6	30.9
Actual % of Eligible Population	25.8	5.5	5.2	5.7	5.2	Remainder

Why are the Projected and Actual Numbers for Each Grade Different?

According to the published figures, all of the actual grade allocations exceeded the projected 'quotas' (with the exception of the D grade, which was reduced as a result of more children being awarded the higher grades). For example, there were 201 more candidates (6,633) with an A grade than were projected by the 25% quota. No reason has been published for the difference between the 'fixed' grade quotas and the published figures. However, correspondence from CCEA confirms that candidates with scores at the boundary between two grades are automatically awarded the higher of the two grades, hence slightly exceeding the projected quotas.

What Does the Test Measure?

Given its role in enabling the selection of pupils for grammar schools, it is important to establish what precisely the Transfer Procedure Test measures. A conventional test, with a

[6] CCEA (2000b) *1999/00 Transfer Procedure Test Results*, News Release NR/98/00 http://www.ccea.org.uk/press/nr9800.htm

single outcome score, should measure only one thing, known in educational testing circles as the 'construct'. For one test it may be knowledge of the highway code while for another it may be knowledge of the French language. With separate mathematics, English and science sections, the Transfer Test could be measuring up to three such constructs: i.e. the application of a candidate's knowledge in each of these areas. Combining the scores for each section to give a single score may make it look like it is measuring one thing. But is it? (This question is addressed later).

Strictly speaking, however, it is not officially known what the Test measures. The Department of Education states only that the Test is designed to assist schools in allocating places. It is not claimed to be a single measure of anything; not ability, intelligence, general reasoning or anything else.

Yet a pupil's performance in the Transfer Test is *commonly perceived* to be a measure of his or her *ability*, a perception that has persisted down through the years from the inception of the first '11-plus' tests in 1947. This perception allows people to take the view that a child with an A is more likely to do well in a grammar school than a child with a B1 or B2 etc. A child with an A is therefore *commonly perceived* to be better ('smarter', 'more able' etc.) than a child with a B1 or a B2 etc. The Test does not, however, offer evidence for any such perceptions. Indeed the evidence from this study is that there is insufficient difference between an A and a B1, or any of the other grades, to justify even the grading never mind a perceived difference in ability.

The details of the study and the questions it explored are provided in the next section.

THE STUDY

The study involved the examination of Transfer Test scripts used as practice tests by Primary 7 pupils in 52 schools across Northern Ireland. The pupils were in the process of completing their preparations for the 1999/2000 sitting of the Test. The sample of schools was designed to cover the main differences in schools and pupils: the five education and library board areas, the four main school types (non-denominational controlled, Catholic maintained, preparatory and integrated), boys and girls, the proportion of pupils entitled to free school meals and school size.

How Representative of the Target Population is the School Sample?

The sample of schools is summarized in Table 2. The two main sectors are well represented and two schools each from the preparatory and integrated primary sectors are also included.

Table 2 Management type of participating schools and number of pupil scripts for each test

			Number of Pupil Scripts		
School Type	**No. of Schools**	**% of Total[7] School Type**	**Test 1**	**Test 2**	**Supp. Test**
Controlled	27	6	889	809	291
Catholic Maintained	21	5	339	404	276
Preparatory	2	-	48	46	51
Integrated	2	8	12	11	5
Totals	52		1288	1270	623

The Test Samples

The tests selected for investigation were the 1998/1999 versions of Test 1, Test 2 and the Supplementary Test. The main sample was made up of 926 pupils who took both Test 1 and Test 2. This, and the individual Test samples, is sufficiently large to enable strong inferences to be drawn about the Test's behaviour with the full cohort.

The Objectives of the Study

The investigation set out to answer a number of questions including:

- Can the Test be used to differentiate children in terms of ability?
- How do children perform in the Test?
- Does the Test grade the children successfully?
- Does the Test behave differently for different groups of children (e.g. for boys and girls, and for younger and older candidates)?
- Does the Test meet international standards for educational testing?

These questions are taken in turn below.

FINDINGS

Can the Test be Used to Differentiate Children in Terms of Ability?

The common perception of the Transfer Procedure Test is that it identifies the children who have the highest *'ability'* of their year group. The rhetoric of *'the top 25%'* makes this an understandable perception among schools, teachers and parents. Those that relate the term 'ability' to the concept of intelligence will, however, avoid the term as they will know that the Transfer Test is an 'achievement' test. Though related to reasoning ability and other aspects of intelligence, performance in an achievement test is much more susceptible to factors external to the child. Factors such as the extent of curriculum covered in class or the quality of teaching received can be expected to affect achievement. Performance in an achievement test, particularly a poor performance, may therefore have little to do with a child's inherent ability.

[7] DE (1999) *Enrolment at Schools in Northern Ireland 1998/99.* Statistical Press Release, Department of Education, Bangor, Co. Down.

Another perception, favoured perhaps by those who do not wish to make the mistake of equating the concept of ability with achievement, is that the Test gives a measure of the child's '*potential to benefit from a grammar school education'*. Again something of a 'folklore' perception, this view is however unable to accommodate the fact that 33% of secondary school students[8] achieve five or more GCSE grades A to C. It would not be unreasonable to expect that they could have done this in a grammar school, if they had had that choice. Yet it is likely that they had been deemed not to have such potential on the basis of their Transfer Test results.

Suggesting a single measurable dimension of a child that might encompass the achievement-oriented design of the Test and its three subject components, is difficult. Nevertheless, the Test procedure does give a single mark to the candidates and therefore implies that the child's performance in a single 'construct' is being measured. Necessarily complex, such a construct might be something like: *'an aptitude for recalling and applying knowledge from the three disciplines: mathematics, English and science'*. One objective of the research was to establish whether the Test measures one construct or more.

Uni-dimensionality

The results show that the Test does not measure a single construct. In all of the tests, the best fit to the data was found to be the 3-construct model (see the Technical Report for full details). The model held true for the overall sample, across genders and for both younger and older candidates (though the sample sizes for the latter make their interpretation problematic). Data for all three tests confirmed the failure of a one-construct (uni-dimensional) model. Treatment of the Test score as a single measure, combining scores in the three subject areas, instead of using them as a profile of separate scores, is therefore questionable. The analysis found no valid grounds for inferring a child's ability, or potential to benefit from any particular type of education, from their Test score or grade.

The Test does however have a high reliability (around 0.90 against a maximum possible reliability measure of 1.00) and the three constructs are strongly correlated (around 0.80 to 0.95 with 1.00 as the maximum possible correlation measure). There are therefore some technical grounds for arguing that the Test shows at least some of the behaviour expected if it was measuring a single attribute of candidates taking it. Although its use for drawing inferences about a child's 'ability' or 'potential to benefit …' must be rejected, and although it is very difficult to propose any such single measure, the high reliability endorses its capacity to rank order candidates' scores even if it is not clear what the scores mean.

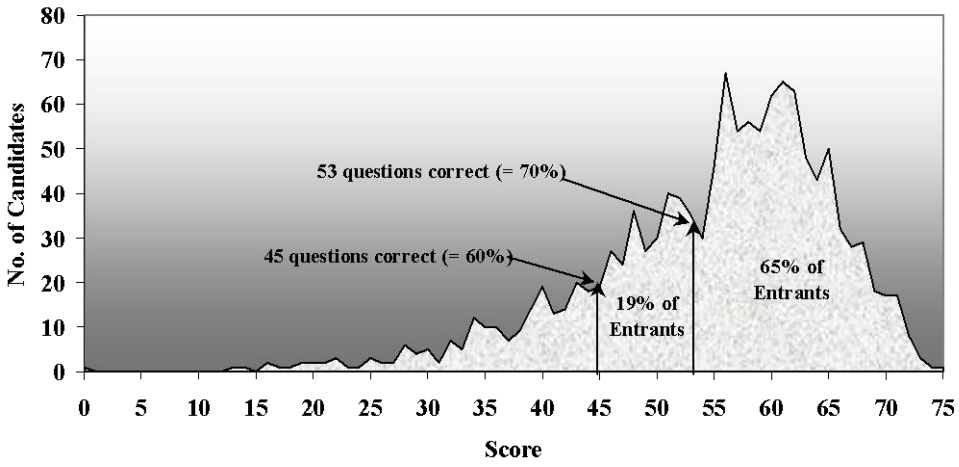
How Do Children Perform in the Test?

The results show that for the large majority of pupils in the sample, all three tests may be described as 'easy'. In Test 1 and the Supplementary Test, for example, over 65% of the candidates completed more than 70% of the items correctly (see Figure 1 for an illustration of this in relation to Test 1). Test 2 was somewhat easier with 74% of candidates getting at least 70% right. Although the comparison of the Test 2 with the Test 1 and Supplementary Test figures suggests a problem of variability between papers, the high scoring in all three tests

[8] DE (2000) *School Performance Tables 1998-99*. Department of Education, Bangor, Co. Down

provides evidence of a more worrying problem; that the score distributions are closely bunched and are at the high end. This is illustrated with data from Test 1 in Figure 1:

Figure 1: Test 1 Score Distribution



Note that 84% (19% + 65%) of the children doing this Test achieved more than 60% of the available marks. The 'easiness' of the three tests must raise questions about all such tests and not just the 1998/1999 version.

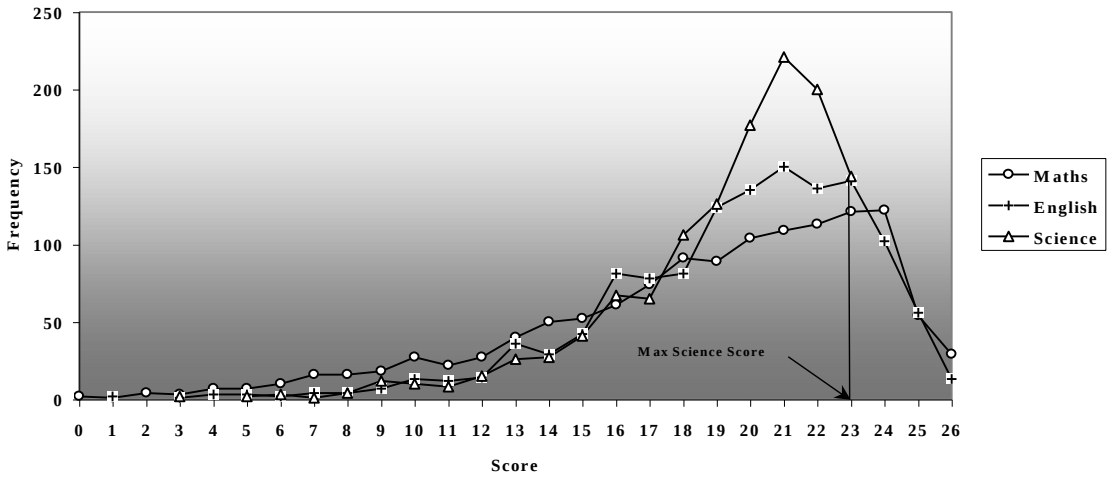
How does the 'easiness' come about?

Clearly the main reason is that overall the questions are easy. This is confirmed by item analysis, which showed that for Test 1, for example, only 3% of the items had facility values[9] less than 0.4 and more than 74% had facility values greater than 0.6. However, there is evidence to suggest that the science questions prove easiest of all the questions in each test, and do not differentiate between candidates as well as either the mathematics or English questions.

The science score distribution is therefore more markedly bunched at the high end of its scores than either mathematics or English. This is illustrated in Figure 2 for Test 2.

[9] A facility value of 0.2 means that the item is completed correctly by only 20 % of the candidates i.e. it is a 'hard' question. At the other extreme, a facility value of 0.8 means that the item is completed correctly by 80% of the candidates i.e. it is an 'easy' question.

Figure 2: Mathematics, English and Science Score Distribution for Test 2



The 'easiness' of the science sections of the three tests is borne out in Table 3, which shows that while the mean score for mathematics and English averages around 70%, that of science is between 82 and 85%.

Table 3: Mean raw scores and mean scores as a percentage of the maximum score for mathematics, English and science in each of the three tests

	Mathematics		English		Science	
Sample	**Mean**	**%**	**Mean**	**%**	**Mean**	**%**
Test 1	17.76	68	17.51	67	19.47	85
Test 2	18.61	72	19.55	75	19.29	84
Supplementary	18.35	71	17.43	67	18.78	82
Average	18.24	**70**	18.16	**70**	19.18	**83**

The lower weighting (0.3 compared with 0.35 for mathematics and English) and relatively high average (mean) scores in science can lead to disadvantage for those who have relatively low scores in the science sections. Despite having the same total Test score to begin with, candidates with low science scores may end up with lower final scores (after age adjustment, standardization and weighting) than candidates who score relatively more in mathematics and/or English. This effect is illustrated in Table 4 for three candidates each with a total Test score of 119 out of 150.

Table 4: Illustration of outcome of standardization and weighting on candidates' final scores

Candidate	DoB	Maths1	Maths2	TotMaths	Eng1	Eng2	TotEng	Sci1	Sci2	TotSci	Total Score	Final Score
1	02-May-89	21	25	**46**	20	18	**38**	18	17	**35**	119	**205**
2	11-May-89	20	22	**42**	16	17	**33**	23	21	**44**	119	**208**
3	21-May-89	21	20	**41**	15	21	**36**	19	23	**42**	119	**208**
1	02-Jul-88	21	25	**46**	20	18	**38**	18	17	**35**	119	**209**
2	02-Jul-88	20	22	**42**	16	17	**33**	23	21	**44**	119	**207**
3	02-Jul-88	21	20	**41**	15	21	**36**	19	23	**42**	119	**207**

The first half of Table 4 presents the scores of the three candidates with very similar birthdays and therefore very little difference in age adjustment. The final score for the child with the relatively low scores in the science sections of both tests (total 35) is three marks less than the scores of the other two children, whose science scores were 44 and 42 respectively. The difference arises because the science scores fall below the average (mean) score for science (around 19, see Table 3). The process of standardization, which uses the mean score, and the subsequent weighting can therefore introduce an artificial difference between the children.

The lower half of Table 4 shows what happens when three children, with the same score profiles as the children in the top half of the table, are processed on the basis of identical ages, mean scores and standard deviations in each of the Test sections. The standardization is therefore identical for all three children but the weighting introduces the opposite effect to that observed in the top half. The relatively high scores in the sections weighted by 0.35 (mathematics and English) produce a higher score for that child in comparison to the other two. These children had the same Test score but scored relatively highly in science, which is only weighted by 0.30.

What impact does the 'easiness' have on grades?

The 'feelgood' factor, which the 'easiness' of the Test is likely to give rise to, represents a serious problem when the grade allocations are considered. Based on the samples used in this study, the most striking effect of the 'easiness' relates to the perceived failing grade, D. Candidates who scored as many as 105 of their answers correct out of a maximum of 150 would have been awarded a D (see Table 5).

This means that children with 70% of the answers correct would have 'failed'. To be given a 'failing' grade with such a high proportion of correct answers is simply unheard of and is very difficult to justify. As the children will likely feel they have scored well, the potential for the award of a D to add confusion to their disappointment is all too clear.

Table 5 illustrates the spread of scores across the grades, using the data from Test 1 and 2 combined, and brings into focus other problems associated with the overall grading. Column 4 lists the percentage of the candidates associated with each grade. Note that the A grade is actually awarded to slightly more than the 36.5% projected from Table 1 as all candidates with a score of 123 (the score at the A/B1 boundary) are given an A i.e. 37.15%. The subsequent projected percentages for B1, B2 etc. therefore derive from this latter figure using the proportions projected (7.3% per grade).

Table 5: Candidates' scores and grade limits for the combined Test 1 and Test 2

Score	Score as % of Questions Correct	% of Pupils with this Score or Better	Grade Limit	Grades	Grade Range
124	83	34.88			
123	**82**	37.15	36.50	**A**	**A/B1**
122	81	39.20			
121	81	40.93			**B1**
120	80	43.30			
119	**79**	44.71	44.45	**B1/B2**	
118	79	47.73			**A to D**
117	78	50.76		**B2**	
116	**77**	53.67	52.01		**B2/C1**
115	77	56.48			
114	**76**	58.64			**C1** **18**
113	75	60.26			
112	75	61.88	60.97	**C1/C2**	
111	74	63.50			
110	73	65.44			**Marks**
109	**73**	66.74			
108	72	67.60		**C2**	
107	71	68.79			
106	71	70.19	69.18		**C2/D**
105	70	71.71			**D**

Does the Test Grade the Children Successfully?

An important point to note from Table 5 is that the grades are spread over 18 marks. This means that the six grades A to D straddle just 12% of the marks available.

In considering whether the grades awarded are to be trusted, educational testing conventions demand that the candidates' scores should be considered in the light of what is known as the *Standard Error of Measurement (s.e.m.).* Once this is calculated it is possible to identify, with 95% confidence, the range in which a candidate's true score[10] lies. This is approximately twice the s.e.m. value above or below the Test score. The s.e.m. for Test 1 and 2 combined was found to be 4.75. The true scores of candidates could therefore be 10 marks above or below their actual scores. Since 18 marks span the five grade boundaries, the potential for misclassifying a child's grade is very clear. This may be illustrated by an example.

Consider two candidates, Gary, who has a Test score of 113 and Siobhan with a Test score of 124. Grading them according to the Test score gives Gary a C1 and Siobhan an A. Yet we can only be sure, at the level of 95% confidence, that Gary's true score lies somewhere in the range 103 to 123 and that Siobhan's true score is in the range 114 to 134. Table 4 shows that Gary's true grade could be the C1 awarded or it could be a D, C2, B2, B1 or an A! Similarly,

[10] The 'true score' is the score that would be obtained if any errors inherent in a single sitting, e.g. arising from distractions, ill-health, undue stress etc., were removed through multiple sittings. It is an internationally accepted convention for determining the confidence to be placed in inferences made from raw test scores.

Siobhan's true grade could be an A as awarded or it could be a B1, a B2 or a C1! The potential misclassification of a child's grade, depending on where their score lies in the rank order, is therefore up to three grades either side of their given grade.

The number of children at risk of misclassification is summarized in Table 6.

Table 6 Predicted proportions and numbers of candidates with secure grades and with grades that are in the misclassification zone

Grade	Predicted %	Predicted Number
Secure A (with 11 or more marks above A/B1 boundary)	11.7	2,053
A (with less than 11 marks above the A/B1 boundary)	25.5	4,487
Candidates with marks between the A/B1 and C2/D boundaries	33.0	5,819
D (with less than 11 marks below C2/D boundary)	12.2	2,148
Secure D (with 11 marks or more marks below the C2/D boundary)	17.6	3,099

Note that the study suggests that only approximately 2,000 of the A-grade and 3,000 of the D-grade candidates are securely graded by the Test. As many as 4,500 A's, with scores within 10 marks of the A/B1 boundary, could however be misclassified. In the main zone of potential misclassification (between the A/B1 and C2/D grade boundaries) a further 5,800 candidates might be wrongly graded. Grade D candidates, with scores within 10 marks of the C2/D boundary (approximately 2,100 children), are also at risk of misclassification.

Does the Test Behave Differently for Different Groups of Children?

Mean score comparisons

The details of these analyses are to be found in the Technical Report, available separately.

ELB Area: While some significant differences were found between the mean scores of candidates from different ELB areas, the patterns of significance were not consistent across all three tests.

Management Type: The mean scores of the candidates from preparatory schools were significantly higher in all three tests than for any of the other school types. There was no significant difference between the mean scores of candidates from controlled and maintained schools.

Free School Meals: The only social index available to the study was each school's proportion of pupils with entitlement to free school meals (FSM). The results showed that the mean scores of candidates from schools with high proportions (greater than 51%) were significantly lower than the mean scores in any other category in Test 1 and the Supplementary Test.

The mean scores of candidates in schools with <10% FSM entitlements were significantly higher than those of the candidates in at least two of the three other FSM categories for Test 1 and Test 2. Social factors

would be generally accepted to have this type of impact on performance profiles[11].

School Size: Candidates from schools of different sizes did not score significantly differently in the tests.

Age: The age of candidates, whether considered monthly or quarterly, gave rise to no significant differences in the mean scores of the groups concerned.

Gender: There were no significant differences in the mean scores of boys and girls in all three tests. Differential item functioning analysis indicated a number of items in each test that were answered differently by boys and girls (and/or by younger and older candidates). However, few of these gave rise to there being a difference of more than 25% between the groups concerned.

Does the Test Meet International Standards for Educational Testing?

It is important to note that the UK does not have technical fidelity standards for educational testing and test developers and users do not necessarily adhere to international standards[12] used across Europe, Asia and North America. The Test and its administration does not meet the international standards set out below. There are others, which the current Test and its administration do not meet, but the selection given is self-explanatory in the context of this report. It is clear that many of the problems identified in this report, particularly in the context of openness about reliability, validity and procedure, would be addressed if international standards were to be applied. The selection of standards is set out below:

> Standard 1.1 (on Validity)
>
> A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.
>
> Standard 1.2 (on Validity)
>
> The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, *and the construct that the test is intending to assess should be clearly described.* (Our Emphasis)
>
> Standard 1.12 (on Validity)
>
> When interpretation of sub-scores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. *Where composite scores are developed, the basis and rationale for arriving at the composites should be given.* (Our Emphasis)

---

[11] DENI (1996b) *Free School Meals and Low Achievement*. Statistical Bulletin, SB2/96 Department of Education for Northern Ireland, Bangor Co. Down

[12] AERA, APA & NCME (1999) *Standards for Educational and Psychological Testing*. (American Educational Research Association, American Psychological Association and National Council on Measurement in Education) Washington DC: American Psychological Association

Standard 2.1 (on Reliability and Errors of Measurement)

For each total score, sub-score or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.

Standard 2.2 (on Reliability and Errors of Measurement)

The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.

Standard 2.4 (on Reliability and Errors of Measurement)

Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported.

Standard 10.1 (Disability)

In testing individuals with disabilities, test developers, test administrators and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement.

Standard 13.7 (on Validity)

In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the validity of the decision.

Standard 13.14 (on Reliability and Errors of Measurement)

In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores. (*Authors' note: the commentary on this standard specifically suggests the provision of information on the probability of misclassification.*)

Other standards refer to a variety of issues such as attention to the establishment of norm-related information, provision of information on score distributions etc.

CONCLUDING REMARKS

The full population data for Transfer Procedure Test entrants in any given year may not behave entirely as the samples in this study, however the problems of grade allocation and the extent of error measurement are almost certain to pervade the full cohort data just as much as the samples. Although this study demonstrates empirically that the potential for misclassification of Transfer Test candidates' grades is unacceptably high, it is also recognized that the problems of grading classification for high stakes rank ordering are difficult to solve. Indeed it is fair to say that the complete elimination of misclassification in any test is impossible to achieve. This said, it is nevertheless important that future administrations of the Test should be openly reported to assure the public of best endeavours to reduce the potential for misclassification and to provide Test takers, their parents and schools with performance information that is based on reliable methods and valid inferences.

The following suggestions for improvement and change should be considered:

- full information, including the return of marked scripts, should be provided to schools and parents;
- raw Test scores and standardized scores should be provided along with the grade;
- details of reliability and the standard error of measurement should be provided along with an explanation of the potential grade misclassification;
- science should be taken out of the Test specification or attention should be given to weighting it equally with mathematics and English, and to ensuring the items used are more discriminatory;
- the unacceptable situation of getting a high percentage of the marks and then receiving a perceived failing grade, D must be addressed;
- the reduction in the number of grade boundaries, perhaps even to one, should be considered in order to reduce the misclassification predicted by the standard error of measurement;
- the grade percentage quotas should apply directly to the Test entry population;
- the development, administration and use of the Test should conform to American Educational Research Association standards for educational testing.

TESTING THE TEST

A Study of the Reliability and Validity of the Northern Ireland Transfer Procedure Test in Enabling the Selection of Pupils for Grammar School Places

TECHNICAL REPORT

TESTING THE TEST

TECHNICAL REPORT

This section provides technical and factual details on the data sample and data analysis methods used in the study, and on the findings that emerged. A wider preamble and contextualization is offered in the section entitled *The Report*. Suffice to repeat here that this report is the culmination of the largest independent study of the Transfer Procedure Test ever carried out.

METHODS

Sample

According to CCEA (2000a), the number of pupils in Year 7 in 1999/2000 is 25,727 and the number entered for the Transfer Test was 17,606. The study involved the examination of Transfer Test scripts used as practice tests by Primary 7 pupils in 52 schools across Northern Ireland. The pupils concerned were in the process of completing their preparations for the 1999/2000 sitting of the Test and although the Test scripts could not be completed under actual Transfer Test conditions, schools do simulate these conditions for 'practice' tests. The sample of schools was designed to cover the main differences in schools and pupils: the five education and library board areas, the four main school types (non-denominational controlled, Catholic maintained, preparatory and integrated), boys and girls, the proportion of pupils entitled to free school meals and school size.

The sample of schools is summarized in Table T1:

Table T1 Management type of participating schools (Total no. of primary schools in 1998/1999 = 916[13])

			ELB Area (No. of Schools)				
School Type	**No. of Schools**	**% of Total School Type**	**BELB**	**NEELB**	**SEELB**	**SELB**	**WELB**
Totals	52	6	10	10	11	6	15
Controlled	27	6					
Catholic Maintained	21	5					
Preparatory	2	-					
Integrated	2	8					

Three 1998 tests were made available to the research team: Test 1, Test 2 and the Supplementary Test. The schools were assured of confidentiality and anonymity but were

[13] DE (1999) *Enrolment at Schools in Northern Ireland 1998/99.* Statistical Press Release, Department of Education, Bangor, Co. Down.

encouraged to include information on gender and age on the scripts. The marked scripts were collected and coded between December 1999 and mid-January 2000. The data were double-checked for coding and computer input accuracy.

Tests 1 and 2 comprised 60 questions and the Supplementary Test comprised 64 questions addressing mathematics, English and science (and technology[14]) in alternating sections (the order of which is set down by the Department of Education). The Supplementary Test is normally only used for those candidates who have missed one of the other tests for an acceptable reason.

The total mark available to the candidates in each test was 75, broken down into 26 marks for each of mathematics and English; and 23 marks for science items.

The number of scripts in each sample is presented in Table T2:

Table T2 Number of scripts in each sample (No. of Test entrants in 1999/2000 = 17,606)

TEST	TEST 1	TEST 2	SUPPLEMENTARY	MATCHED TEST 1 & TEST 2
N	1288	1270	623	926
% of Entrants	7.3	7.2	3.5	5.3

These details are presented in Table T3 in terms of the number of candidates in each category along with the sample sizes for age and gender. Aside from some small cell sizes for the Supplementary Test data, the spread of sample across the data types is good.

[14] Subsequent references to science alone should be taken to include technology.

Table T3 Number of candidates by ELB, management type of school, proportion of free school meals (FSM), school size, pupil age and gender

	NUMBER OF CANDIDATES		
	TEST 1	TEST 2	SUPP. TEST
ELB AREA			
BELB	284	273	243
NEELB	220	271	107
SEELB	348	332	147
SELB	90	49	9
WELB	346	345	117
Total	1288	1270	623
SCHOOL MANAGEMENT			
Controlled	889	809	291
Maintained	339	404	276
Integrated	12	11	5
Voluntary (Prep)	48	46	51
Total	1288	1270	623
%FSM IN SAMPLE SCHOOLS			
0-10	580	596	306
11-30	503	460	155
31-50	135	169	156
51+	70	45	6
Totals	1288	1270	623
SCHOOL SIZE			
<90 Pupils	211	211	91
90-189 Pupils	205	248	153
190+ Pupils	872	811	379
Totals	1288	1270	623
AGE (IN QUARTERS)			
Oldest 3 Months	185	207	117
Next 3 Months	171	209	98
Next 3 Months	148	170	93
Youngest 3 Months	181	217	106
Total	685	803	414
GENDER			
Male	614	597	280
Female	629	606	313
Total	1243	1203	593

Data Analysis

The data were analysed in a number of ways summarized thus:

- Simple frequency counts, standard deviation calculations, and range assessments using Microsoft Excel and SPSS;
- Confirmatory Factor Analysis (using PC-based LISREL) to test the data for uni-dimensionality;

- Item facility analyses and Differential Item Functioning analysis using Mantel-Haenszel Chi-square (χ^2) (SPSS) to examine the performance of the Test items for different candidate types; and
- Analysis of Variance (SPSS ANOVA) comparisons to examine the mean scores of candidates arranged in the ELB, management type, FSM, school size, age and gender group categories.

Differential Item Functioning (DIF) and Confirmatory Factor Analysis (CFA) are explained in more detail below.

Differential Item Functioning

Differential item functioning (DIF) is a measure of the difference in how an item functions for two sample populations. An item is functioning correctly if, for all candidates obtaining the same score on a homogeneous set of items such as the Transfer Test, the proportion of candidates answering the item correctly is the same for each sample of the population under consideration. Differences in proportions indicate a biased item.

If a number of biased items exist in a test, the technical fidelity of the test is compromised as a variety of attributes other than those that the test was designed to measure are interfering with the functioning of the test. This is of particular importance if the item bias occurs between two groups of candidates of different genders or two groups of mixed gender candidates but differing ages. In such cases the test can be viewed as measuring different things for each subgroup.

Using the Mantel-Haenszel method, where 2×2 contingency tables are created for each item, it is possible to establish occurrences of DIF. The item under analysis is frequently referred to as the *studied item*. For each studied item, a 2×2 matrix of values of the number of correct and incorrect scores against the two sub-populations is created. The sub-groups of the population are called the *reference group* (the sample serving as the basis for comparison) and the *focal group* (the sample which is the focus of the analysis). For each group, the number of correct and incorrect responses to the studied item is entered in the appropriate cell in the matrix. The structure of the contingency table is summarized below.

	Response to Item X		
	Correct	Incorrect	Total
Reference group	a	b	a+b (Total no. in group)
Focal group	c	d	c+d (Total no. in group)
Total	a+c (Total correct)	b+d (Total incorrect)	a+b+c+d

The null hypothesis for this method of DIF is that there is no difference between the proportion of correct and incorrect answers for each group, i.e.

$$\frac{a}{b} = \frac{c}{d}$$

A measure of the deviation from the null hypothesis can then be calculated from the ratio:

$$\alpha = \frac{ad}{bc}$$

When α =1, the null hypothesis is accepted. The larger that the value of α deviates from 1, the greater the difference in functioning of the studied item between the reference and focal groups.

Confirmatory Factor Analysis

Factor analysis is a popular procedure for investigating relations between a set of measured variables and their underlying latent variables. Since latent variables (factors) are theoretical constructs they are not directly observable and cannot be directly measured. Measured variables (also referred to as observed, manifest or indicator variables) are believed to represent the underlying factors of interest.

In confirmatory factor analysis (CFA) a statistical model based on prior knowledge is formulated to describe the constructs that underlie the indicator variables. The procedure then involves the testing of the model using data on all measured variables. By forcing data to fit the hypothesized model, the goodness-of-fit between the observed data and the statistical model can be determined.

CFA fits into the general structural equations modelling (SEM) approach and can be considered as a measurement sub-model of SEM. CFA defines relations between the observed and latent variables, as well as relations among the factors themselves. In SEM the most popular estimation procedure is that of maximum likelihood, in which the researcher seeks estimates of parameters most likely to have generated the measured data.

The calculations involved in a maximum likelihood solution are so complex that they are virtually impossible to handle without the use of a computer. Joreskog and his colleagues have devised computer programs to handle these computations, the most popular of which is LISREL (Linear Structural Relations: Joreskog and Sorbom, 1989).

In CFA a statistical model is postulated in advance and then the hypothesis is tested for plausibility. Unlike conventional statistical analysis where the null hypothesis is rejected, CFA uses the null hypothesis that the model provides a satisfactory fit for the observed data. This means that there should be no significant difference between the observed covariance/correlation matrix and the covariance/correlation matrix reproduced using the parameter estimates of the model. The χ^2 test can then be used to test the fit between the restricted hypothesized model and the unrestricted sample data. In CFA a small χ^2 value indicates a better fitting model.

χ^2 is a powerful test if the sample size is large while a small sample size gives rise to a high probability of accepting the hypothesis even if the model is actually a poor fit. There is no universal agreement on exactly what size a sample should be but research has shown that a sample size of at least 200 is needed for factor analytic studies (Boomsman, 1987) while Tanaka (1987) has argued that in structural equations modelling, it is the ratio of the number of subjects to the number of estimated parameters that is of concern.

As a result of this sensitivity to sample size, Joreskog and Sorbom (1989) proposed that it be used as a:

> Goodness (or badness)-of-fit measure in the sense that large χ^2-values correspond to bad fit and small χ^2-values to good fit. The degrees of freedom serve as a standard by which to judge whether χ^2 is large or small.

Wheaton, Muthen, Alwin and Summers (1977), in an early application of LISREL, suggested a χ^2/d.f. ratio of five or less as representing an adequate fit. In response to this, Carmines and McIver (1981) proposed a χ^2/d.f. of two or three as a rough indication of reasonable fit. More recently, Byrne (1989) has suggested a χ^2/d.f. ratio of less than two as representing an adequate fit. The Byrne ratio is used in this study.

No single exclusive index has been discovered for goodness of fit but a range of indices have been proposed by various researchers such as Bentler and Bonett (1980), Hoelter (1983) and Joreskog and Sorbom (1989). In addition to the χ^2/d.f. ratio, the indices used in this study include the root mean square residual (RMR) and the adjusted goodness of fit index (AGFI). It is important to emphasize that current expert opinion demands that all of the indices should be used together in scrutinizing results and reliance on only one or two is not recommended. For the purposes of this study, the authors therefore consider a model to be a good fit if the following conventional criteria are met:

- χ^2/d.f. < 2
- RMR < 0.05
- AGFI > 0.8

and if there are no significant differences between the correlation/covariance matrices.

FINDINGS

Testing for Uni-dimensionality

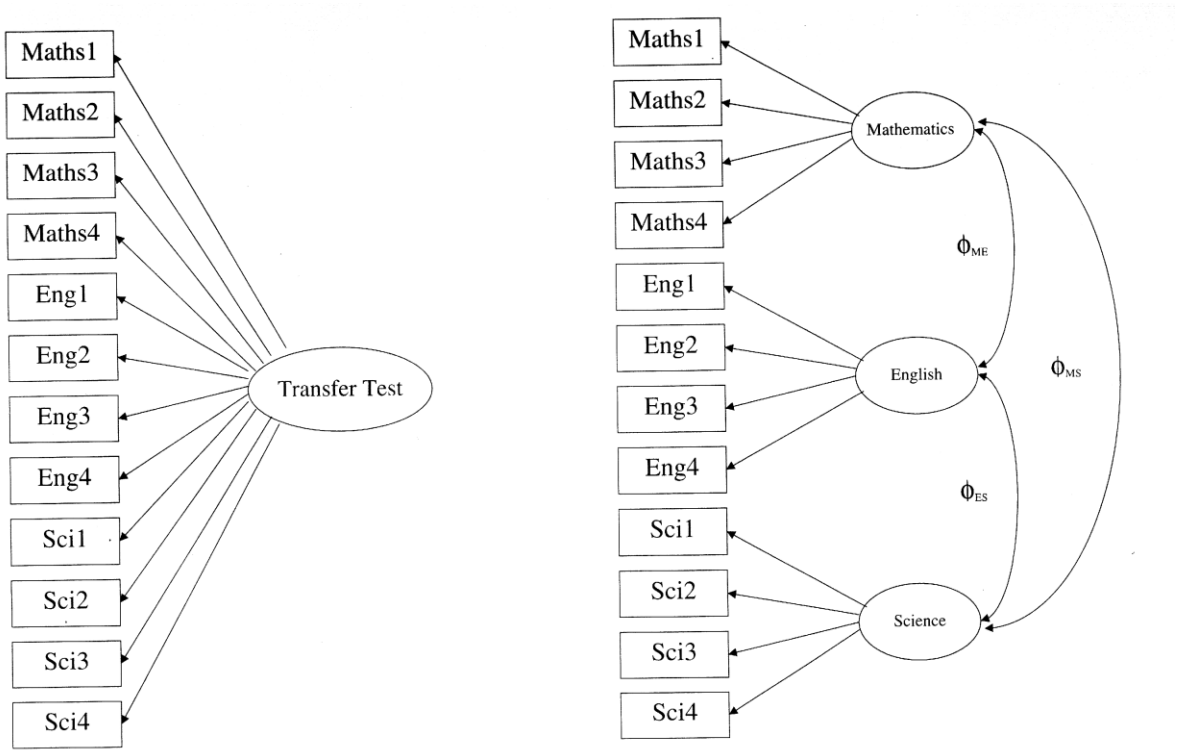
In investigating the Test for uni-dimensionality, confirmatory factor analysis was used to consider the goodness of fit of a one-construct model. The theoretical 3-construct model was also tested. The first part of the process was to group the item scores in each subject into 'bundles' of related items as follows.

The mathematics items were identified as belonging to one of the Northern Ireland Curriculum attainment targets: *Number*, *Measures*, *Shape & Space* or *Handling Data*. Using these categories the Maths1 indicator was composed of half the *Number* items, the indicator Maths2 comprised the remaining *Number* items, Maths3 contained the *Measures* and *Handling Data* items while Maths4 grouped all of the *Shape & Space* items together. Each of the bundles, Maths1 to Maths4, had the same total marks available.

A similar process was used to group the science items into four bundles: Sci1 to Sci4. This process was made easier due to the existence of the three knowledge attainment targets: *Physical Processes*, *Living Things*, *Materials* and the process attainment target *Investigating Science* - the main element of which was recognizing a fair test. In this case the four indicators were represented by equal numbers of items and recombining the items into bundles was therefore not required.

The assessment of English focused mainly on *Writing* with some aspects of the *Reading* attainment target incorporated into the items in the form of identifying the audience for a given piece of writing, selecting phrases with a similar meaning and so on. The English items were bundled in the same way as the mathematics items by combining similar sets of items into four bundles with equal totals of Test marks: Eng1 to Eng4. The two models are illustrated below:

Illustration of One- Construct and 3-Construct Models for the Transfer Procedure Test



CFA analysis was applied to data for the whole sample, for the boys and for the girls in all three tests. Analyses were also carried out for 'younger' (born between 1 March 1989 and 31 August 1989) and 'older' (born between 1 August 1988 and 30 October 1988) categories of candidates.

Using the goodness of fit criteria:

- χ^2/d.f. < 2
- RMR < 0.05
- AGFI > 0.8

the one-construct model failed to fit in all cases as shown in Table T4. Note that the girls' data from the Supplementary Test meets most of the criteria but the model failed on the basis that, for the relatively small sample, the p value (p=0.000) showed a significant difference between the correlation/covariance matrices.

Table T4 Goodness of fit criteria for the one-construct model for all three tests

Test 1					
Sample	n	χ^2/df	RMR	AGFI	Fit?
All	1288	3.456	0.027	0.963	8
Boys	614	2.204	0.031	0.951	8
Girls	629	2.431	0.033	0.951	8

Test 2					
Sample	n	χ^2/df	RMR	AGFI	Fit?
All	1270	4.339	0.029	0.953	8
Boys	597	3.109	0.035	0.931	8
Girls	606	2.599	0.032	0.944	8

Supplementary Test					
Sample	n	χ^2/df	RMR	AGFI	Fit?
All	623	3.376	0.038	0.922	8
Boys	280	2.359	0.048	0.884	8
Girls	313	1.819	0.037	0.921	8

The same indicators were used for testing the 3-construct model and the results are summarized in Table T5 for the three tests:

Table T5 Goodness of fit criteria for the 3-construct model for all three tests

Test 1					
Sample	n	χ^2/df	RMR	AGFI	Fit?
All	1288	1.406	0.016	0.986	4
Boys	614	1.197	0.023	0.975	4
Girls	629	1.574	0.025	0.970	4

Test 2					
Sample	n	χ^2/df	RMR	AGFI	Fit?
All	1270	1.733	0.018	0.983	4
Boys	597	1.739	0.026	0.964	4
Girls	606	1.438	0.023	0.970	4

Supplementary Test					
Sample	n	χ^2/df	RMR	AGFI	Fit?
All	623	1.399	0.024	0.971	4
Boys	280	1.208	0.032	0.947	4
Girls	313	1.074	0.029	0.956	4

Clearly, the hypothesis that a single construct model fits the Test is rejected and a 3-construct (mathematics, English and science) model is accepted.

To confirm the stability of the 3-construct model solution, the mathematics, English and science items was re-grouped into three bundles and two bundles respectively. The 3-construct model held firmly regardless how the items were bundled (see Table T6 as an illustration of the results for Test 1).

Table T6 Goodness of fit criteria for the 3-construct model for Test 1 (whole sample) using different arrangements of indicator measures

Number of Indicator Bundles	N	χ^2/df	RMR	AGFI	Fit?
4	1288	1.406	0.016	0.986	4
3	1288	1.306	0.013	0.990	4
2	1288	0.937	0.007	0.995	4

In all of the tests, the most parsimonious fit to the data was found to be the 3-construct model This discovery held true for the overall sample, across genders and for both the 'younger' and the 'older' candidates (though the sample sizes for the latter make their interpretation problematic).

Despite the goodness of fit of the 3-construct model and the lack of fit of the one-construct model, however, the high disattenuated correlation coefficients (ϕ) between the 3-constructs in the model (see Table T7) suggest that the Test may nevertheless behave in the manner of a one-construct model.

Table T7 Disattenuated correlation coefficients between the constructs

	Disattenuated Correlation Coefficients (ϕ)		
Test	**Maths & English**	**Maths & Science**	**English & Science**
Test 1	0.853	0.915	0.898
Test 2	0.872	0.870	0.899
Supplementary Test	0.816	0.837	0.940

The most likely explanation of this would be what is termed the Positive Manifold effect. This effect results from the pupils experiencing the same teacher, the same teaching style and the same degree of importance attached to each of the subjects: mathematics, English and science. As a result, it is possible that the candidates do not view the Test as three separate sub-tests addressing each of the subject areas, but in fact, they see it as a single test not tied to any particular area or subject. Practice and an intense focus on test-taking strategy would consolidate the perception of the Test as a unitary entity in the candidates' minds and the result is partial single-construct performance from the Test.

Item Facility

One of the characteristics of any test, which is good at creating a rank order according to the aspect of the candidates that is being measured, is that its items gather maximum information about the attainment of candidates. Items that are very difficult and items that are very easy are considered to be poor as most candidates get them incorrect and correct respectively.

Only candidates at the extremes of ability show any differences from the large majority, the very able getting the difficult items right and the very weak getting the easy items wrong. One of the characteristics used to assess the quality of an item is its 'facility' value. On a range of 0 to 1, 0 represents an item that no-one gets right and 1 represents an item that everyone gets right. The aim in any test then is to create items that have facility values around 0.5 as such items maximize the information and facilitate rank ordering The facility values of the dichotomous items[15] in the three tests were examined and are reported in Table T8:

Table T8 Facility values for Test samples

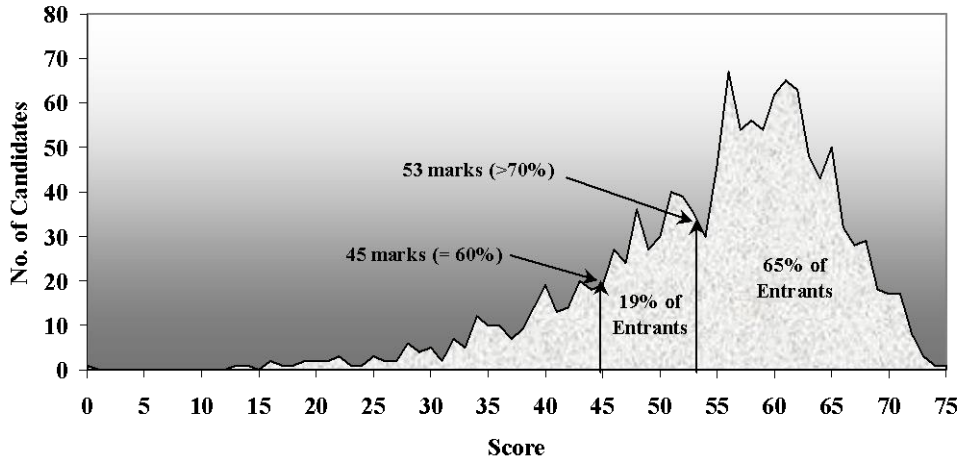
		Test1		**Test2**		**Supp Test**	
	No. of Items	65		69		70	
Facility Value	**%Candidates Correct**	**No of Items**	**% Items**	**No of Items**	**% Items**	**No of Items**	**% Items**
<0.4	0-40	2	3	0	0	2	3
0.4-0.6	41-60	15	23	9	13	13	19
>0.6	60+	48	74	60	87	55	79
>0.8	80+	31	48	30	43	30	43

Table T8 shows that Test 1 and the Supplementary Test have around 20% of their items with facility values from 0.4-0.6 while Test 2 has only 13%. These proportions indicate that there are as few as 1-in-7 and at most 1-in-5 items that support rank ordering. As can be seen from the table, this means that more than 40% of the items in each test were answered correctly by more than 80% of the children. As for 'hard' questions, Test 2 has no items that were answered by fewer than 40% of the candidates and the other two tests have only two each. Since the items attract a score of 1 or 0, then it is clear from Table T8 (3rd row) that more than 60% of the children scored more than 70% on the dichotomous items.

Frequency analysis shows that for the large majority of pupils in the sample, all three tests may be described as 'easy'. In Test 1 and the Supplementary Test, for example, over 65% of the candidates completed more than 70% of the items correctly (see Figure T1 for an illustration of this in relation to Test 1). Test 2 was somewhat easier with 74% of candidates getting at least 70% right (Figure T2). Although the comparison of the Test 2 with the Test 1 and Supplementary Test figures suggests a problem of variability between papers, the high scoring in all three tests provides evidence of a more worrying problem; that the score distributions are closely bunched and are at the high end. This is illustrated with data from Test 1 in Figure T1:

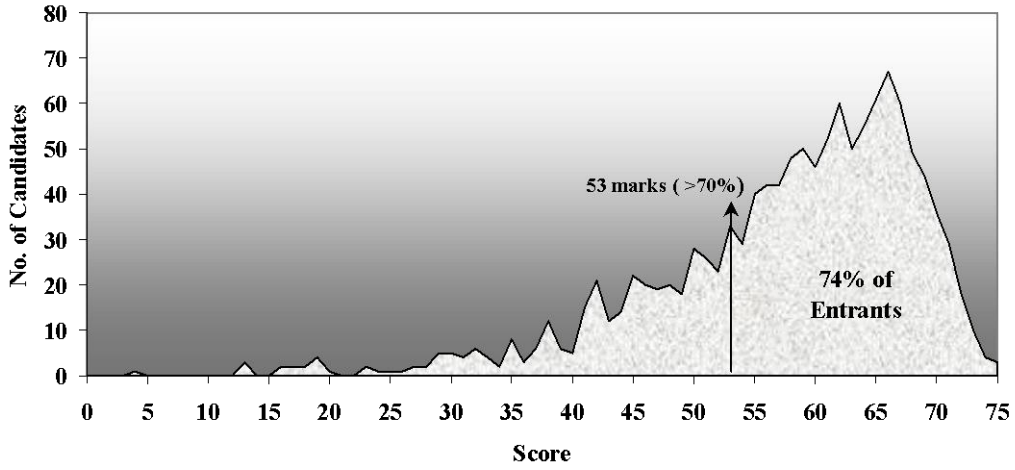
[15] Items that are marked simply as correct or incorrect

Figure T1: Test 1 Score Distribution



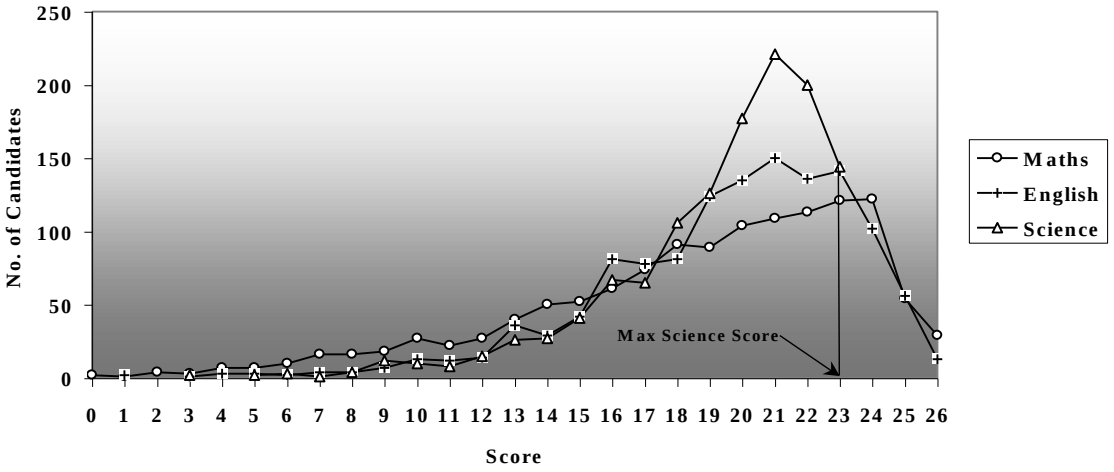
Note that 84% (19% + 65%) of the children doing this Test achieved more than 60% of the available marks. The 'easiness' of the three tests must raise questions about all such tests and not just the 1998/1999 version. We will return to this quite serious problem later in the report.

Figure T2: Test 2 Score Distribution



There is clear evidence from the results to suggest that the science questions prove easiest of all the questions in each test, and do not differentiate between candidates as well as either the mathematics or English questions. The science score distribution is therefore more markedly bunched at the high end of its scores than either mathematics or English. This is illustrated in Figure T3 for Test 2.

Figure T3: Mathematics, English and Science Score Distribution for Test 2



The 'easiness' of the science sections of the three tests is borne out in Table T9, which shows that while the mean score for mathematics and English averages around 70%, that of science is between 82% and 85%.

Table T9: Mean raw scores and mean scores as a percentage of the maximum score for mathematics, English and science in each of the three tests

	Mathematics		English		Science	
Sample	**Mean**	**%**	**Mean**	**%**	**Mean**	**%**
Test 1	17.76	68	17.51	67	19.47	85
Test 2	18.61	72	19.55	75	19.29	84
Supplementary	18.35	71	17.43	67	18.78	82
Average	18.24	**70**	18.16	**70**	19.18	**83**

The lower weighting (0.3 compared with 0.35 for mathematics and English) and relatively high average (mean) scores in science can lead to disadvantage for those who have relatively low scores in the science sections. Despite having the same total Test score to begin with, for example, candidates with low science scores may end up with lower final scores (after age adjustment, standardization and weighting) than candidates who score relatively more in mathematics and/or English. This effect is illustrated in Table T10 for three candidates each with a total Test score of 119 out of 150.

Table T10: Illustration of outcome of standardization and weighting on candidates' final scores

Candidate	**DoB**	Maths1	Maths2	**TotMaths**	Eng1	Eng2	**TotEng**	Sci1	Sci2	**TotSci**	**Total Score**	**Final Score**
1	02-May-89	21	25	**46**	20	18	**38**	18	17	**35**	119	**205**
2	11-May-89	20	22	**42**	16	17	**33**	23	21	**44**	119	**208**
3	21-May-89	21	20	**41**	15	21	**36**	19	23	**42**	119	**208**
1	02-Jul-88	21	25	**46**	20	18	**38**	18	17	**35**	119	**209**
2	02-Jul-88	20	22	**42**	16	17	**33**	23	21	**44**	119	**207**
3	02-Jul-88	21	20	**41**	15	21	**36**	19	23	**42**	119	**207**

The first half of Table T10 presents the scores of the three candidates with very similar birthdays and therefore very little difference in age adjustment. The final score for the child

with the relatively low scores in the science sections of both tests is three marks less than the scores of the other two children. The difference arises because the science scores fall below the average (mean) score for science (around 19, see Table T9). The process of standardization, which uses the mean score, and the subsequent weighting can therefore introduce an artificial difference between the children.

The lower half of Table T10 shows what happens when three children, with the same score profiles as the children in the top half of the table, are processed on the basis of identical ages, mean scores and standard deviations in each of the Test sections. The standardization is therefore identical for all three children but the weighting introduces the opposite effect to that observed in the top half. The relatively high scores in the sections weighted by 0.35 (mathematics and English) produce a higher score for that child in comparison to the other two. These children had the same Test score but scored relatively highly in science, which is only weighted by 0.30.

Differential Item Functioning

Although CFA confirmed that the tests behaved similarly for the samples of boys and girls, it was important to examine the item performance of the tests. Differential item functioning (DIF) is used to measure the difference in the way that the individual items function within a test for two groups of candidates. In this case, we were considering boys and girls with equal Test scores and comparing their performance on each of the Test items. Having the same score in the Test enabled the two groups to be considered as being matched in terms of their level of performance on the Test. Any differences, therefore, in item functioning can be attributed directly to the item itself and not to a difference in performance on the Test construct.

Using the modal score in each case, the candidates with this score were regrouped as boys or girls for the gender analysis and as 'younger' (approximately 20% of the youngest in each sample) and 'older' (approximately 20% of the oldest children in each sample) for the age analysis. Table T11 provides the details of the sub-sample sizes.

Table T11 Sub-samples of boys and girls and 'younger' and 'older' candidates

	Modal Score	Boys	Girls	Younger	Older
Test 1	56	27	38	11	10
Test 2	65	33	30	10	23
Supplementary Test	60	16	19	11	6

The Mantel-Haenszel method of DIF uses the number of correct and incorrect responses for each group for every item in the Test. Items functioning in the same way for two groups of candidates have α values equal to 1.0. The extent to which the items function differently for each group can be determined by the deviation in the α value from 1. As with the confirmatory factor analyses, the age sub-groups were relatively small and any results relating to them need to be treated with caution. The gender sample sizes were larger but a number of the contingency tests failed as the cell size in some instances was too small. These small cell sizes arose from the 'easiness' of the tests i.e. the 'incorrect' cells occasionally had values less than 5.

DIF analysis identified eight items in Test 1 that performed significantly differently for the boys and girls and nine that were answered significantly differently by younger and older candidates. Test 2, with very high success rates (60%+ candidates completing more than 84% of the items correctly) produced no gender-biased items and only three for age. The Supplementary Test was particularly prone to small sub-samples and is not reported. Table T12 provides an excerpt of the contingency testing results for Test 1 as an illustration. The difference in the proportions of boys and girls answering these eight items was significant.

Table T12 Illustration of contingency analysis results for DIF by gender and age (Test 1)

Item	Correct				Incorrect				α		Facility Value			
	Boys	Girls	Old	Young	Boys	Girls	Old	Young	Gender	Age	Boys	Girls	Old	Young
1	20	37	10	11	7	1	0	0	0.077	-	0.741	0.974	1.000	1.000
2	23	24	5	9	4	14	5	2	3.354	0.222	0.852	0.632	0.500	0.818
3	18	36	8	10	9	2	2	1	0.111	0.400	0.667	0.947	0.800	0.909
4	14	7	3	3	13	31	7	8	4.769	1.143	0.519	0.184	0.300	0.273
5	21	36	8	10	6	2	2	1	0.194	0.400	0.778	0.947	0.800	0.909
6	23	34	9	10	4	4	1	1	0.676	0.900	0.852	0.895	0.900	0.909
7	16	9	6	2	11	29	4	9	4.687	6.750	0.593	0.237	0.600	0.182
8	11	26	5	4	16	12	5	7	0.317	1.750	0.407	0.684	0.500	0.364

Note that the shaded items show particularly pronounced differences in correct answers between the boys and girls, differing as they do in facility values by more than 0.25 (25%). Items 4 and 7 were easier for the boys while 3 and 8 were easier for the girls.

Mean Score Comparisons

The sample frame covered ELB area, school management type, the proportion of free school meals (FSM), school size, pupil age and gender. It was important to assess whether candidates in the different sub- categories of these groups performed differently to any significant extent on the tests. Tables T13-T18 present the Analysis of Variance results of the various comparisons carried out (NS = not significant, p = probability, SS = small sample).

Table T13 Candidate numbers and mean scores by ELB

ELB	TEST 1 (F=5.36, p<0.001)		TEST 2 (F=3.11, NS)		SUPP. TEST (F=5.61, p<0.001)	
	N	Mean	N	Mean	N	Mean
BELB	278	53.77	270	54.14	237	56.61
NEELB	219	56.15	271	57.82	107	54.01
SEELB	342	55.53	327	58.81	147	51.75
SELB	88	57.35	49	58.84	SS	
WELB	342	53.15	339	56.00	114	54.19
Total	1269	54.73	1256	57.48	605	54.61

The results on these comparisons show no fixed pattern. Subsequent t-tests established that in Test 1 the BELB and WELB candidates' mean score is significantly less than that of the candidates in the other boards but the pattern was not repeated for the other tests.

Table T14 Candidate numbers and mean scores by MANAGEMENT TYPE

MANAGEMENT	TEST 1 (F=8.69, p<0.001)		TEST 2 (F=7.50, p<0.001)		SUPP. TEST (F=6.18, p<0.001)	
	N	Mean	N	Mean	N	Mean
Controlled	877	54.17	799	56.83	285	53.97
Maintained	333	55.15	400	57.91	273	54.08
Integrated	SS		SS		SS	
Voluntary (Prep)	47	62.06	46	64.30	51	60.82
Total	1257	54.74	1245	57.48	609	54.61

Although the numbers are relatively small for the preparatory schools (Table T14), they are nonetheless secure enough for the comparisons. In all three tests, subsequent t-tests indicated that the mean score of the preparatory school sample was significantly higher than the means of the other groups.

Table T15 Candidate numbers and mean scores by proportion of FREE SCHOOL MEALS

%FSM	TEST 1 (F=15.00, p<0.001)		TEST 2 (F=8.78, <0.001)		SUPP. TEST (F=5.53, p<0.001)	
	N	Mean	N	Mean	N	Mean
0-10	572	56.18	590	58.78	300	56.19
11-30	499	54.25	456	56.88	155	54.21
31-50	132	53.97	167	56.10	153	51.86
51+	66	47.44	43	51.37	SS	
Totals	1269	54.74	1256	57.48	608	54.61

Table T15 shows that there were significant differences among the means of the groups, depending on the proportion of free school meal (FSM) entitlements that the candidates' schools had. Subsequent t-tests indicated pupils in schools with 51%+ of the children entitled to free school meals achieved significantly lower mean scores than the other groups in Test 1 and Test 2(the 51%+ sample was too small for the Supplementary Test).

This trend, based on the proportion of FSM entitlement, was endorsed with the 11-30% group's mean score being significantly less than that of the <10% group for Test 1. For Test 2 the difference was even more pronounced with the mean score of pupils in schools with less than 10% free school meal entitlements doing significantly better than any of the other groups. For the Supplementary Test, the 31-50% group continued the trend with a significantly lower mean score than the <10% group.

Table T16 Candidate numbers and mean scores by SCHOOL SIZE

SIZE	TEST 1 (F=0.24, NS)		TEST 2 (F=0.35, NS)		SUPP. TEST (F=3.62, NS)	
	N	Mean	N	Mean	N	Mean
<90	207	54.57	210	58.02	91	56.10
90-189	199	54.33	242	57.55	153	56.07
190+	863	54.87	804	57.32	370	53.65
Totals	1269	54.74	1256	57.48	614	54.61

Table T17 Candidate numbers and mean scores by AGE

AGE	TEST 1 (F=1.67, NS)		TEST 2 (F=1.49, NS)		SUPP. TEST (F=1.37, NS)	
	N	Mean	N	Mean	N	Mean
Oldest 3 Months	182	55.57	205	58.08	115	55.96
Next 3 Months	170	56.32	209	58.17	98	55.07
Next 3 Months	147	53.79	167	56.28	92	54.09
Youngest 3 Months	188	55.37	228	58.34	106	53.08
Total	687	55.32	809	57.80	411	54.58

Table T18 Candidate numbers and mean scores by GENDER

GENDER	TEST 1 (F=0.94, NS)		TEST 2 (F=0.08, NS)		SUPP. TEST (F=0.15, NS)	
	N	Mean	N	Mean	N	Mean
Male	614	54.41	597	54.36	280	54.78
Female	629	54.97	606	57.71	313	54.73
Total	1243		1203		593	

Examination of tables T16-T18 indicates that school size, age and gender had no bearing on the performance of candidates. The 'Age' pattern is solid for all three sets of data, with all cell sizes having comfortable numbers of subjects.

Distribution of Grade Allocations

The proportion of candidates in the cohort taking the Test, in relation to the number of eligible candidates (i.e. pupils in Year 7 in primary schools), has remained more or less constant at between 68-70% since 1994/1995, as demonstrated in Table T19.

Table T19 Proportion of candidates in the cohort taking the Transfer Test in relation to the number of eligible pupils

Year	Year 7 Pupils	No. Entered	%Entered
1999/2000	25727	17606	68
1998/1999	26562	17974	68
1997/1998	26801	18229	68
1996/1997	26264	18265	70
1995/1996	26325	17995	68
1994/1995	25790	18175	70

Since 1995, the quotas for each grade A to D have been set in the following proportions:

> Grade A is awarded to the top 25% of the entire age group eligible to sit the tests, B1 is awarded to the next 5% of the pupils, B2 to the next 5%, C1 to the next 5%, C2 to the next 5% and D to those remaining (CCEA, 1998).

With 25,727 pupils in Year 7 in 1999/2000 (CCEA, 2000a), 25% represents 6,432 candidates. Therefore 6,432 of the highest scoring entrants to the Transfer Test in 1999/2000 were to be given an A. The number of candidates theoretically in each of the following bands is 5% of 25,727 i.e. 1286. Table T20 summarizes the projected numbers of candidates in each grade.

Table T20 Numbers of candidates awarded grades A to D in 1999/00 (17,606 entrants)

	GRADE					
	A	**B1**	**B2**	**C1**	**C2**	**D**
% of Eligible Population (25,762)	25	5	5	5	5	Remainder
Projected No. (from 17,606) with Each Grade	6432	1286	1286	1286	1286	6030
Projected % of Entrants	**36.5**	**7.3**	**7.3**	**7.3**	**7.3**	**34.2**
Actual No. with Each Grade	6633	1416	1335	1456	1333	5433
Actual % of Entrants	37.7	8.0	7.6	8.3	7.6	30.9
Actual % of Eligible Population	25.8	5.5	5.2	5.7	5.2	Remainder

As proportions of the cohort of 17,606 that enter the Test, the third row of Table T20 shows that these numbers represent grade quotas of 36.5, 7.3, 7.3, 7.3 and 7.3% respectively for the grades A to C2. In this manner the 45% of all possible entrants (based on the population of children in their last year of primary school) becomes approximately 66% against the cohort of Test candidates.

Having established the projected percentages in each grade[16], it is possible to carry out an analysis of the sample data to investigate how grades would have been allocated to the

[16] Arguably the sample should be subjected to the same quota proportions (25%, 5% etc) as the eligible population would be if all pupils entered the Test. For the purposes of this study, however, it was considered

candidates concerned if the sample tests had in fact been their 'real' Test. Table T21 presents the details of the allocation of grades on the basis of the raw scores.

Table T21: Candidates' scores and grade limits for the combined Test 1 and Test 2

Score	Score as % of Questions Correct	% of Pupils with this Score or Better	Grade Limit	Grades	Grade Range
124	83	34.88			
123	**82**	37.15	36.50	**A** / **A/B1**	
122	81	39.20			
121	81	40.93		**B1**	
120	80	43.30			
119	**79**	44.71	44.45	**B1/B2**	
118	79	47.73			**A to D**
117	78	50.76		**B2**	
116	**77**	53.67	52.01	**B2/C1**	
115	77	56.48			
114	**76**	58.64		**C1**	**18**
113	75	60.26			
112	75	61.88	60.97	**C1/C2**	
111	74	63.50			
110	73	65.44			**Marks**
109	**73**	66.74			
108	72	67.60		**C2**	
107	71	68.79			
106	71	70.19	69.18	**C2/D**	
105	70	71.71		**D**	

The most striking matter to note on examination of this table is the fact that candidates who scored as many as 105 of their answers correct, out of a maximum of 150, would have been awarded a D. This means that children with 70% of the answers correct would have 'failed'. To be given a 'failing' grade with such a high proportion of correct answers is simply unheard of and is very difficult to justify. As the children will likely feel they have scored well, the potential for the award of a D to add confusion to their disappointment is all too clear.

Table T21 illustrates the spread of scores across the grades, using the data from Test 1 and 2 combined, and brings into focus other problems associated with the overall grading. Column 4 lists the percentage of the candidates associated with each grade. Note that the A grade is actually awarded to slightly more than the 36.5% projected from Table T20 as all candidates with a score of 123 (the score at the A/B1 boundary) are given an A i.e. 37.15%. Once the A/B1 boundary has been established, identification of the subsequent boundaries derives from the application of the fixed quota percentages. With B1 being the threshold percentage for A+7.3% (44.45%), B2 being the threshold percentage for B1+7.3% etc (i.e. from the

best to simulate as much as possible the 'usual' circumstances of the Transfer Test i.e. with around 70% of the eligible population taking part and the percentage quotas adjusted accordingly.

table, 44.71+7.3 = 52.01%), the projected threshold percentages are as presented in Table T22:

Table T22 Projected grade threshold percentages

	Grades			
	B1	**B2**	**C1**	**C2**
Threshold %	44.45	52.01	60.97	69.18

Another important point to note from Table T21 is that the grades are spread over 18 marks. This means that the six grades A to D are separated by just 12% of the marks available.

In considering whether the grades awarded within this range are to be trusted, educational testing conventions demand that the candidates' scores should be considered in the light of what is known as the *Standard Error of Measurement (s.e.m.).* The s.e.m. gives an indication of the precision by which the observed score on the Test (i.e. the raw score) reflects the candidate's performance in the construct being measured[17]. It is calculated from the standard deviation (SD) and reliability using the formula:

$$SE_{measure} = SD \times \sqrt{1 - \text{reliability of test}}$$

The s.e.m. for each test and for the combination of Test 1 and Test 2 is given in Table T23 along with details of the standard deviation in the scores (SD) and the reliability coefficients.

Table T23 Standard deviations (SD), reliability and standard errors of measurement (s.e.m.) of the three tests individually and of the combined Test1 and Test 2

Test	SD	Reliability	s.e.m	1.96 x s.e.m.
Test 1	10.62	0.89	3.47	6.80
Test 2	10.92	0.91	3.27	6.41
Supplementary Test	10.99	0.90	3.39	6.64
Test1 & Test 2	20.26	0.95	4.75	9.31

Once the s.e.m. is calculated it is possible to identify, with 95% confidence, the range in which a candidate's true score lies i.e. it offers a measure of how valid the inferences drawn about the candidate's performance on the measured construct are. The 95% confidence range is (1.96 x s.e.m.) marks above or below the Test score. With an s.e.m. of 4.75 for Test 1 and 2 combined, the true scores of candidates therefore have the potential to be approximately 10 marks above or below their actual scores. Since 18 marks span the five grade boundaries, the potential for misclassifying a child's true grade is very clear. This may be illustrated by an example.

Consider two candidates, Gary, who has a Test score of 113 and Siobhan with a Test score of 124. Grading them according to the Test score gives Gary a C1 and Siobhan an A. Yet we

[17] The 'true score' is the score that would be obtained if any errors inherent in a single sitting, e.g. arising from distractions, ill-health, undue stress etc., were removed through multiple sittings. It is an internationally accepted convention for determining the confidence to be placed in inferences made from raw test scores.

can only be sure, at the level of 95% confidence, that Gary's true score lies somewhere in the range 103 to 123 and that Siobhan's true score is in the range 114 to 134. Table T21 shows that Gary's true grade could be the C1 awarded or it could be a D, C2, B2, B1 or an A! Similarly, Siobhan's true grade could be an A as awarded or it could be a B1, a B2 or a C1! The potential misclassification of a child's grade, depending on where their score lies in the rank order, is therefore up to three grades either side of their given grade.

The number of children at risk of misclassification is summarized in Table T23.

Table T23 Predicted proportions and numbers of candidates with secure grades and with grades that are in the misclassification zone

Grade	Predicted %	Predicted Number
Secure A (with 11 or more marks above A/B1 boundary)	11.7	2,053
A (with less than 11 marks above the A/B1 boundary)	25.5	4,487
Candidates with marks between the A/B1 and C2/D boundaries	33.0	5,819
D (with less than 11 marks below C2/D boundary)	12.2	2,148
Secure D (with 11 marks or more marks below the C2/D boundary)	17.6	3,099

Note that the study suggests that approximately 5,000 of the candidates are securely graded by the Test, i.e. that inferences drawn on the basis of their scores are reasonably safe. These are the 2,053 secure A's and 3,099 secure D's which lie outside the 95% confidence intervals that span the grades.

However, approximately 4,500 A's, whose Test scores lie up to 10 marks above the A/B1 boundary, could potentially be misclassified . In the main potential misclassification zone (between the A/B1 and C2/D grade boundaries) a further 5,800 candidates might be wrongly graded. Grade D candidates, with scores within 10 marks of the C2/D boundary (2,148 children), are also at risk of misclassification.

CONCLUDING REMARKS

One-Construct vs. 3-Construct Models for the Transfer Test

Despite the fact that confirmatory factor analysis (CFA) rejects a single construct model in almost every case, most test theorists would interpret the CFA models in this study as uni-dimensional for all practical purposes. All of the inter-construct disattenuated correlation coefficients were very high though few of the 95% confidence intervals associated with these included unity, another weakness in claiming uni-dimensionality. CFA also confirms that, for all practical purposes, the tests measure boys and girls and younger and older candidates (as defined earlier) in the same way. The fact that the same construct model fits across gender or across age indicates that the Test measures without significant bias.

The high inter-correlations of the 3-construct model are explained by what psychometricians call the Positive Manifold Effect. Ceci (1994, p. 112), writing in the context of the American curriculum, summarizes this well established effect as follows:

> In this same vein, Cronbach has remarked that the correlation between verbal and quantitative abilities may be an epiphenomenon of an individual's being jointly trained in both: *"The high correlation between verbal and numerical abilities is due in part to the fact that persons who remain in school are trained in both types of content"* (Cronbach, 1970, p. 479). One can easily imagine that a skill that is suddenly deemed important enough by the dominant culture to be included in its schooling will correlate highly with other skills taught concurrently, such as verbal and numerical skills. Thus if cooking, computing, and cartography were suddenly inserted into the school curriculum, they, too, would tend to inter-correlate.

This study's finding that a one-construct model was always rejected in favour of a 3-construct model, even though there was significant evidence of one-construct behaviour, has been replicated in a number large scale studies in the USA.

> Recently, McCardle and Horn (in press) and Loehlin (1989) present impressive evidence that differential structural models of intellectual development can be fit to the same matrix of means, standard deviations, and correlations and these various models can be quite dissimilar, despite their near equivalent fit. (Ceci, 1994, p. 112)

Item Facility Values and Differential Functioning

The Transfer Test aims to discriminate between pupils by separating them maximally on achievement. Basic psychometrics teaches that items with facilities near 0.5 enable maximum information about pupil attainment to be gathered. Very easy items (with facilities close to 1) and very difficult items (with facilities close to 0) lose information essential to the establishment of a stable rank ordering of pupils. Facility value analysis in the samples in this study revealed very few items with facilities near 0.5 and indicated that very high proportions of candidates would be expected to score highly e.g. 60%+ taking the tests completed more than 70% of the items correctly.

The very high facility values encountered in the study also militated against accurate differential item function analysis because so few items were answered incorrectly. This gives rise to small cell size problems and renders χ^2 analysis unsafe for some of the items. In general though the analysis suggested that the items, with only a few exceptions across the tests, were answered in similar proportions by the boys and girls and the younger and older candidates.

Test Reliability and Grade Allocation

The tests' internal consistency measures – those of the three individual tests being approximately equal to 0.9 with the Test 1 & Test 2 combination being 0.95 - were acceptable, although this is always a judgement call. Nuttall and Willmott (1972, p.42), writing in the context of public examinations such as the Certificate of Secondary Education (CSE) and the General Certificate of Education (GCE), posited the following "standards":

> In practice, there is likely to be a difference between subjects in the value of this upper bound [on reliability], with those having the more precise marking schemes (e.g. mathematics and the sciences) having an upper bound near 0.98, and those with the less precise marking schemes (e.g. English language essay) having a rather lower upper bound (0.8 - 0.9). … For multiple-choice examinations, reliabilities from 0.90

to 0.97 are considered good, and are typical of most well-constructed tests. Values from 0.80 to 0.90 are fair, whilst reliabilities below 0.8 are treated with some caution.

All three tests had standard errors of measurement (s.e.m.) of the order of 3.4 and the s.e.m. of the combined Test 1 & Test 2 was 4.75. This is a worrying discovery. For the composite test, the 95% confidence interval for a pupil graded C2 or B1 includes at least two grade boundaries. Classical test theory holds that scores separated by less than (4 x s.e.m.) cannot be distinguished. It follows that scores in adjacent grade categories are indistinguishable. Clearly, the potential for considerable misclassification remains when composite test scores are converted to grades.

Misclassification of Grades

No attempt is made to make candidates, parents and schools aware of the fact that all Transfer Tests misclassify pupils; that no Transfer Test can measure with accuracy greater than ± 1 grade. The reason for the misclassification is simple. In any test, which has with a non-zero standard error of measure, candidates whose scores fall short of (but are close to) a given grade boundary, can have true scores which exceed the grade boundary. This study found that pupils whose observed scores lie in one grade category can have true scores which can fall as much as three grades away. In a detailed analysis of the reliability of 16+ examinations in Britain, Willmott and Nuttall (1975) demonstrated that typically 25% of examinees are misclassified in examinations with reliabilities of the order 0.9. It must be emphasized that such misclassification is systemic and has nothing to do with the accuracy of marking.

Please (1971) used the bivariate normal distribution to establish that the percentage of misclassifications is likely to be nearer 40%. Using Please's (1971) analysis it is possible to estimate that more than 30% of the pupils who take the Transfer Test will be assigned the wrong grade. It is important to underline that these misclassifications derive from the grading framework. Errors arising from item scoring and totalling compound this error but the Transfer Test, as currently constituted, simply cannot misclassify fewer than three pupils in 10.

Please (1971) offered a solution to the problem of grade misclassification, which is worthy of note in the present context. By reporting grades in a manner that acknowledges test fallibility, misclassifications can be reduced to below 10%. Under the system proposed by Please, a Transfer Test grade would no longer be reported as B1, for example, but as: [A **B1** B2]. The grades which flank the B1 grade indicate that there is a high probability that the pupil should really be graded A or B2. Instead of grading pupils C1, for example, they would be graded: [B1 B2 **C1** C2] under this system. The inclusion of three extra grades recognizes the Test's fallibility in the B1 to C2 range for the Transfer Test grades. Pupils with observed scores in the C1 category have significantly high probabilities of having true scores in grades A, B1, B2, C2 and D.

Openness

Many of the difficulties associated with the technical aspects of the Transfer Test, to which this present work draws attention, have not come to light in other than theoretical treatments before now because access to the necessary information is prohibited. This study highlights the importance of openness in high stakes tests such as the Transfer Test. Modern validity

inquiry - which interprets all validity as construct-referenced - includes a consideration of the social consequences of testing. Given that significant adverse consequences for individuals can arise from interpreting the grade sequence C2, C1, B2, B1 as a perfect hierarchy, those responsible for designing and administering the Transfer Test have a clear responsibility to admit to the Test's frailties. This report represents a call for greater openness and accountability in respect of a test which can have a profound effect on a child's future. Clearly, no test is perfect and the Transfer Test's designers may feel, with some justification, that in order to eliminate misclassifications, they face the impossible task of reducing the standard error of measure to zero. Nevertheless this report calls for information on the Transfer Test and its weaknesses to be conveyed clearly to the public.

Standards for Test Administration

While many countries in the world have testing and assessment regimes governed by the American Educational Research Association's *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999), the British examination bodies have always avoided the publication of data bearing on the technical fidelity of their assessment instruments. At a time when transparency and accountability has been urged on a range of agencies which serve the public, no British testing agency has published a reliability or validity study of any consequence in the last decade. This is certainly the case with the Northern Ireland Transfer Test. The silence on Transfer Test information in Northern Ireland rings loud when contrasted with the approach taken in just three of the AERA *Standards*:

> Standard 1.1 (on Validity)
>
> A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.
>
> Standard 1.2 (on Validity)
>
> The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, *and the construct that the test is intending to assess should be clearly described.* (Our Emphasis)
>
> Standard 2.1 (on Reliability and Errors of Measurement)
>
> For each total score, sub-score or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.

The Transfer Test is a high stakes test and Northern Ireland pupils deserve the protection of technical fidelity standards which apply to children elsewhere in the world. This report demonstrates that the grading framework has potential for significant misclassification; the inference that a pupil with a grade B1, for example, has more 'ability' than a pupil graded B2 simply cannot be validated.

While testing agencies in countries that have adopted the AERA *Standards* can be held to account for the validity and reliability of their instruments, one could be forgiven the impression that British testing agencies are accountable only for their question setting and marking. Parents who dispute their child's Transfer Test grade have recourse to a re-mark to ensure that the correct marks were awarded for each item and that these were accurately

totalled. Parents with more fundamental concerns have no recourse except perhaps to the law. However, the secrecy that surrounds the Transfer Test leaves the courts with few options other than to assume that there is a one-to-one correspondence between the Test score and the child's 'ability'. Adoption of the *Standards* would quickly disabuse the courts of this view and would give test developers, administrators and candidates alike access to powerful evidence if the need arises for them to argue their case.

BIBLIOGRAPHY

AERA, APA, & NCME (1999). *Standards for Educational and Psychological Testing.* Washington DC: American Psychological Association

Bentler, P.M. and Bonett, D.G. (1980) Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 599-606.

Boomsman, A. (1987) The robustness of maximum likelihood estimation in structural equations models. In P. Cuttance and R, Ecobs (Eds.) *Structural Modelling by Example: Applications in Educational, Sociological and Behavioural Research*, pp 160-188. New York: Cambridge University Press.

Byrne, B.M. (1989) Multigroup comparisons and the assumption of equivalent construct validity across groups: methodological and substantive issues. *Multivariate Behavioural Research*, 24, 4, 503-523.

CCEA (1998) *Specifications of the 1999/2000 Transfer Tests*. Northern Ireland Council for the Curriculum, Examinations and Assessment, Belfast

CCEA (2000a) Number of pupils in Year 7 (boys and girls) in the years 1994/2000 and number entered for Transfer Test (personal communication from the Northern Ireland Council for the Curriculum, Examinations and Assessment, Belfast)

CCEA (2000b) *1999/00 Transfer Procedure Test Results*, News Release NR/98/00 http://www.ccea.org.uk/press/nr9800.htm

Carmines, E.G. and McIver, J.P. (1981) Analysing models with unobserved variables: analysis of covariance structures. In G.W. Bohrnstedt & E.F. Borgatta (Eds.) *Social Measurement: Current Issues*, pp.65-115. Newbury Park, CA: Sage.

Ceci, S.J. (1994). *On Intelligence*. Cambridge, MA: Harvard University Press.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Harcourt Brace Jovanovich Publishers.

Cronbach, L.J. (1970). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (pp. 443-507). Washington, DC: American Council on Education.

DENI (1992) *The Parents' Charter for Northern Ireland.* Department of Education for Northern Ireland, Bangor, Co. Down.

DENI (1996a) *Transfer Test Results 1989/90-1995/96*. Statistical Bulletin, SB1/96 Department of Education for Northern Ireland, Bangor Co. Down

DENI (1996b) *Free School Meals and Low Achievement*. Statistical Bulletin SB2/96, Department of Education for Northern Ireland, Bangor, Co. Down.

DE (1999) *Enrolment at Schools in Northern Ireland 1998/99.* Statistical Press Release, Department of Education, Bangor, Co. Down.

DE (2000) *School Performance Tables 1998/99*. Department of Education, Bangor, Co. Down

Hoelter, J.W. (1983) The analysis of covariance structures: goodness of fit indices. *Sociological Methods and Research*, 11, 325-344.

Joreskog, K. G. & Sorbom, D. (1989) *Lisrel 7: A Guide to the Program and Applications* (2nd edition). Chicago: SPSS Inc.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education.

Mulaik, S.A. (1972) *The Foundations of Factor Analysis*. New York: McGraw-Hill.

Nuttall, D. L., & Willmott, A. S. (1972). *British Examinations: Techniques of Analysis.* Slough: National Foundation for Educational Research.

Please, N. W. (1971). Estimation of the proportion of examination candidates who are wrongly graded. *British Journal of Mathematical and Statistical Psychology*, 24 (2), 230-238.

Sutherland, A. E. (1990) Selection in Northern Ireland: from 1947 Act to 1989 Order *Research Papers in Education*, 5 (1) pp. 29-48

Tanaka, J.S. (1987) How big is big enough? Sample size and goodness of fit in structural equations models with latent variables. *Child Development*, 58, 134-146

Wheaton, B., Muthen, B., Alwin, D.F. and Summers, G.F. (1977) Assessing reliability and stability in panel models. In D. R. Heise (Ed.) *Sociological Methodology*, pp. 84-136. San Francisco: Jossey-Bass.

Willmott, A. S., & Nuttall, D. L. (1975). *The Reliability of Examinations at 16+*. London: Macmillan Education.