Thinking Skills and Admissions

A report on the Validity and Reliability

of the

TSA and MVAT/BMAT Assessments

Alan Willmott

September 2005

Contents

Section **Contents** **Page**

Contents

Section Contents Page

List of Tables

List of Tables

Overview

This report provides some background details of the work by UCLES on Thinking Skills since the late 1980s, and in so doing draws heavily on the report by Joyce Chapman (Chapman, 2005) who was closely involved in that work. Reference is made to project documents that were produced in those early days but unfortunately not all are still available now (see note about the availability of references on page vii).

The report draws together much of the past and current work and evaluates both the TSA and part of the MVAT/BMAT tests used to date and looks at the TSA as a predictor of university achievement.

Some new work has been undertaken but the aim has been to draw together the previous work on which these assessments are based and to review the TSA and BMAT assessments that have so far been used.

In essence both the TSA and BMAT (Parts 1 and 2) are found to be reasonably sound assessments. There is little predictive validity information available as yet but what there is does not offend and, indeed, could be taken to be cautiously optimistic about future developments.

There are many pointers to further work that is needed and the evidence points to the urgent need for a wider data set to be collected so that different analyses can be carried out to confirm the usefulness of the tests. A summary of these points can be found in Appendix B.

It is also clear that the current work to train new Question Writers and produce new questions is very much due as there are signs that without careful attention the development of the assessment of Thinking Skills may falter following the initial burst of activity with questions from the past.

Finally, given the growing importance of Thinking Skills and the effects that can follow from the teaching of these skills, there is a need to consider whether there is a need to produce support materials for teachers.

As with most reports of this kind, the author is grateful to others for work done and suggestions made in enabling this report to come about both by providing data and test information. Thanks are due in this regard to Robert Harding, Alec Fisher, Mark Shannon, Alf Massey and Alastair Pollitt (who, as a true researcher, held onto data from early 1990 ‘in case they were needed’). Thanks are also due to my wife, Marilena, without whom this report could not have been produced in time.

Alan Willmott
September 2005

Note on References in this Report

In preparing this report, a wide variety of documents has been drawn on including many old project documents, many of which are no longer available. The references in the text are provided as usual and each one falls into one of four categories. The list of references in the References indicates into which one of the following categories each reference falls.

Category A

Reference to an existing publication or document that may be obtained in the usual way (e.g. from libraries, bookshops, etc.).

Category B

Reference to a document held in electronic form, in either Adobe pdf or MSWord format, which may be downloaded from the Cambridge Assessment website under the Research section listing publications, conference papers and other articles.
http://www.cambridgeassessment.org.uk/research/confproceedingsetc/

Category C

Reference to a document held in electronic form, in either Adobe pdf or MSWord format, and which may be downloaded from the Cambridge Assessment website under the Research section listing older (archive) documents of interest.
http://www.cambridgeassessment.org.uk/research/historicaldocuments/

Category D

Reference to an old UCLES project document that is no longer available.

Thinking Skills and Admissions

1 Background

1.1 UCLES, Thinking Skills and MENO

The University of Cambridge Local Examinations Syndicate (UCLES) has been involved with the development of assessments of 'Thinking Skills' since the late 1980s. At that time, the numbers of people wishing to undertake Higher Education were growing rapidly and many of these people did not possess traditional entry qualifications. There was also a view that A-Levels were not necessarily the best predictors of success in Higher Education and UCLES established a research and development programme focussed on the provision of tests of academic aptitude. A detailed discussion of the work of UCLES in this period is given by Chapman (2005) and Fisher (2005) provides a discussion on how the concepts behind 'Thinking Skills' have been developed since they were first introduced.

The initial interest centred on a project to develop a test to aid in the selection of students wishing to study Law at University and although this work was not as successful as had been hoped, many useful lessons were learned (Rule, 1989). The emphasis on the development of a test to predict academic performance in a single subject, Law, was recognised as being a very specific aim and, following the evaluation of this work, discussions were started to define what should be assessed, and why, so that questions and tests could be developed that would be applicable to the prediction of success in subjects other than Law.

During this process, a number of trials were conducted using questions designed to assess a range of different skills and a more ambitious plan was then proposed with a view to designing an instrument that might be appropriate for use across a range of subjects. As an integral part of this activity, a review of past work was commissioned that would provide a basis for discussion and a paper was commissioned that would build on this review and offer suggestions for progress. The results of this work are reported in Fisher (1989, 1990a, 1990b) where a proposal is made for a general test of academic aptitude called The Higher Studies Test.

After further work, a consultation paper was produced (UCLES, undated) containing proposals for building a test and providing details of possible question types that might be used. This was followed by a change of name for the proposed test to the Academic Aptitude Profile (see Chapman, 2005) and documents were produced that outlined the proposed scheme, its rationale, some sample questions and a guide to writing and editing questions (UCLES 1992a, 1992b, 1992c, 1992d, respectively).

It had become clear that the aim should not be to produce a psychometric aptitude test but rather to identify and define those skills (Thinking Skills) that were crucial to success in Higher Education. Consideration thus moved to a Higher Education Aptitude Test (Fisher, 1992) and the production of tests that would assess the necessary skills directly. The Academic Aptitude Profile would be used for guidance where this was felt to be useful with students but the Higher Education Aptitude Test (HEAT) would be used as an aid to selection. In the end, these two separate ideas led to further discussions and came together as will be seen below.

Consultations between UCLES and a number of institutions of Higher Education took place and led to an agreement on the main focus of development. Work should concentrate on the identification and definition of skills in two main areas: Critical Thinking and Mathematical Reasoning. Five-option multiple-choice questions would be used and the language level would be that of broadsheet newspapers.

From these initial decisions, questions were developed and trial tests created for pre-testing the new ideas. From the evaluation of the results of pre-testing and subsequent refinement of questions, further pre-testing was conducted and the subsequent results further evaluated.

In the period 1989 to 1995, therefore, UCLES developed a range of assessments in the general area of Thinking Skills. New types of questions were commissioned and a Thinking Skills Service under the title of MENO came into being in the period 1993-1995 (Meno was a pupil of Socrates; see Plato's Meno dialogue). There are many publications that explain the nature of this service (see UCLES 1993a, 1993b, 1993c, 1993d), which made available a set of measures that together created a composite assessment of Thinking Skills and provided schools and universities with a means of assessing the capability of their students. Six assessments were available, assessing the kinds of skills necessary for understanding everyday arguments and logical reasoning in a variety of situations of the kind students would meet in the course of Higher Education.

Substantial work was also done to address the 'fairness' of the new questions to ensure that particular groups (defined by gender, race, ethnic origin, age and social class) would not be disadvantaged by the nature and administration of the assessment.

The rationale behind MENO suggested quite clearly that:

> "Three things are suggested by this discussion.
>
> 1. skills may be developed generally as well as in subject-specific contexts. There is no need to feel uncomfortable about this. Scientists, for example, are often highly proficient problem solvers in their own fields; but they are also able to solve problems in their everyday lives, or become business people and solve the problems involved in running a business.
>
> 2. for younger people, thinking skills are likely to be developed in the context of formal disciplines, but more mature people are more likely to demonstrate their thinking abilities in the context of everyday experience.
>
> 3. in considering the intellectual capacities of people, <u>both</u> skills and content are relevant. In many cases, it may be worth while paying more attention to skills than has happened in the past."
>
> (UCLES, 1993d)

In practice, the MENO Service was not taken up widely and only two of the six skill assessments were used to any extent. The discussion in Section 2.1 looks at this matter and presents some research evidence that relates to the nature of the six-skill model. It was not long therefore before the MENO Service came to an end although the two more popular assessments were continued by UCLES. An award entitled the Cambridge Thinking Skills Certificate (CTSC) was introduced in 1996 that built on the two most popular MENO skills and this award continued until 2000. When the new Cambridge International Examinations (CIE) was formed in 1998, the CTSC was re-named the Cambridge Award in Thinking Skills (CATS) and this ran from 2000 until 2004 until it was adapted to form part of a new modular A-Level in Thinking Skills.

1.2 A Growing Problem with Admissions

As with many universities, the proportion of students applying to the University of Cambridge who were achieving Grade A in all of their subjects had been increasing every year and by 2000 was very high indeed. In the years 1973 to 2003, for example, the changes in the pattern of applicants and admissions to the University of Cambridge were as shown in Table 1.1.

Table 1.1: Applications and Admissions for the University of Cambridge in 1973 and 2003

	Applications		Admissions	
Year	1973*	2003**	1973*	2003**
Total	3422	11518	1786	3089
Per cent AAA	26	68	41	91
Per cent AAB	17	15	21	6
Success Rate %			52	27

* Robinson (2002)
** Cambridge Reporter (2004)

Thus, for admissions to the University of Cambridge, many features changed substantially over the 30 years from 1973 to 2003, although others did not. In particular:

- The number of applicants rose more than threefold (from 3422 to 11518).

- The proportion of applicants with three Grade As rose from 26 per cent to 68 per cent.

- The proportion of applicants with two Grade As and one Grade B did not change significantly (from 17 to 15 per cent respectively).

- The proportion of applicants admitted that held three Grade As rose from 41 per cent to 91 per cent.

- The proportion of applicants admitted that held two Grade As and a Grade B fell from 21 per cent to 6 per cent.

- The success rate for applications made fell from 52 per cent to 27 per cent.

These figures show both the growth in numbers of applicants and also the increasing need to differentiate between those who are achieving maximum grades in the existing educational system (i.e. at A-Level).

1.3 A Trial

Within the University of Cambridge, the existence of the CTSC was known and it was thought that the type of questions that were being used were such that they could be used to provide useful additional information to those involved in the university admissions process in Cambridge. As a result, a small-scale trial was undertaken in one subject in 1999 (one College) and 2000 (two Colleges) using a CTSC question paper.

The trial concluded that “Thinking Skills Assessment is valuable”, that it was useful to have assessments of “both numerical and verbal reasoning” but that there were “some problems with language and cultural assumptions”. In commenting on the test used in the trial, it was considered that “the test shows promise” in that it was “consistent, independent of subject” and that while there was no specific preparation required to do well on the test, “any preparation is likely to be of general value”.

Finally it was made clear that a “large-scale validation study is needed” and that an attempt should be made to “track rejected students as well as those admitted” (Robinson, 2002).

Before the results of the trial had been fully reported, discussions were held with UCLES and it was decided to build on the work to date by developing further tests for trial to provide extra information to those administering the process of admissions. It was from these continuing developments that the TSA was born in 2001.

The numbers of students taking TSA has grown steadily since 2001 when it was taken by 289 applicants; in 2002, 472 were tested, in 2003 1551 and by 2004 this had risen to 2147 applicants from the UK and overseas.

1.4 A Further Development

Around 1997/8 there were further developments within the University that were leading to a trial of a different form of admissions test for students. In this case, there was a need to find ways of relieving the pressure on selection of applicants wishing to follow medical and veterinary courses. As some of those involved had also been involved in the work that led to MENO this new test which had three parts also contained an assessment of Thinking Skills.

Details of this test and results from analysing data from the tests used are given in Section 7 of this report.

2 The TSA Questions

2.1 The Six Skills of MENO

The six skills assessed by MENO were:

Critical Thinking
Communication
Understanding Argument
Literacy
Problem Solving
Numerical & Spatial Operations

The questions were tried out extensively (see Massey, 1994a) and the results provided much information on the nature of the six proposed scales. Analyses were conducted using data from many sub-components of the scales and questions assessing particular skills were looked at in some detail. The results showed that the analyses did not fully support the educational justification for the six skills as being independent components and so the six-skill educational model was being challenged from a measurement standpoint.

It was found that the component skills that were assessed by multiple-choice questions '… achieve modestly respectable reliability estimates but those which include examiner marking fail to do so'. Also, from work on the data set that contained the complete set of MENO components it was found that the data only really supported the presence of two '… relatively discrete thinking skills …'. These were tentatively entitled 'numerical/spatial problem solving' and 'formal reasoning' (Massey, 1994a).

Not surprisingly, it was thus also considered that a profile report of scores, as had been intended when MENO was conceived, would be difficult to support in measurement terms (Massey, 1994a). These results are relevant to the TSA and are discussed further in Section 6.5.

At the same time it became clear that although the six-component model for MENO was generally accepted as being useful educationally, the problems that were associated with the administration and marking of six different assessments were significant. Users were tending to focus on the two skills of Critical Thinking and Problem Solving as being particularly useful in assessing the capabilities of their students. Thus the practical realities of testing leading to the use of only two skills and the analytical results indicating that the other four skills did not add significantly to the information from the two central assessments sat comfortably together. A discussion of the definition and assessment of these two skills is given in Fisher (2005).

When the MENO Service came to an end, questions assessing both Critical Thinking and Problem Solving were used as the basis for a new UCLES award, the Cambridge Thinking Skills Certificate, in 1996. The Cambridge Award in Thinking Skills (CATS) introduced by Cambridge International Examinations in 2000 was also based on these two skills.

2.2 Other Developments

Since the late 1990s there has been a growing interest in introducing Critical Thinking to the curriculum in schools in the UK. As a result, and after trials with a number of question types in 1999 and 2000, there is now an AS award in Critical Thinking (OCR, 2004) that has been available since 2000. An Advanced Extension Award (AEA) has also been available since 2003 (OCR, 2003) and an A-level, which builds on the modules from the AS award, will be available from 2006 (OCR, 2004).

Cambridge International Examinations has had an AS award in Thinking Skills since 1999 (CIE, 2004), and is now developing an A-level as well. As a result of this new development, which is built on a modular structure, the CATS award has now been discontinued.

There is, however, a major difference between these two (sets of) assessments. The OCR awards are for UK students and cover Critical Thinking only while those for CIE are for the international market and cover both Critical Thinking and Problem Solving, thus continuing the concept of the awards from the CTSC and CATS.

2.3 The TSA

When it was suggested that tests could be designed to yield results that might help the University admissions process, it was considered that a test of Critical Thinking alone might prove to be too narrow a context for assessment. As a consequence, the TSA was based on both Critical Thinking (CT) and Problem Solving (PS) questions drawn from the development of the 1990s and the CTSC.

A TSA test consists of 50 questions to be answered in 90 minutes and this does provide an element of time pressure on candidates. Of the 50 questions, 25 are Critical Thinking questions and 25 are Problem Solving questions. Appendix A lists the types of questions used in the TSA to assess each of the two skills and gives the proportions of these questions in a complete test.

This report does not consider examples of question types or individual questions and those interested in the nature of the skills assessed and the types of questions that have been devised to assess these skills are referred to the documents mentioned in Section 1, to Fisher (2002) where example questions are given and to the handbook for students taking the TSA (UCLES, 2004a). The handbook provides examples of all question types currently used together with the correct answers and explanations of why those answers are correct.

The first TSA tests were used in 2001 and similar tests have been continued each year since then. For 2001 and 2002 the tests were only available on paper but in 2003 and 2004 the tests were also available online to be taken at a PC workstation. This use of technology allowed the results of those tested to be available to Colleges very quickly.

A website is available that gives details of the TSA (http://tsa.ucles.org.uk) and from which a copy of the handbook mentioned above can be downloaded. A short online test (of 10 questions) which provides feedback to the user can be taken at this website. In addition there is a full-length TSA test, of 50 questions, that can be taken on-line under ‘examination conditions’ – i.e. against the clock! This version of the test can also be downloaded and then printed for those who wish to take the test in paper-and-pencil form.

The results of an investigation into the nature of the two skills assessed by CT and PS questions are given in Section 6 of this report. It will be noted that there are indications of some differences between the measurements provided by these two types of questions and the implications of these differences will be considered.

3 Validity and Reliability

3.1 Some Properties of an Assessment

There are many criteria that can be used to judge the value of an assessment and the initial considerations might simply be the type and format of the questions on the paper. For example, will objective questions, such as multiple-choice or short answer questions, suffice for what is needed or should essay-type questions be used? What kinds of responses are being sought from the students? And for what purpose will these assessments be used?

Then there are practical matters such as the time available for administering the test and a consideration of what time should be allowed for the test. Should students be made to work against the clock or should they have plenty of time to consider their answers carefully? Here it will be important to be aware of the aims of the assessment session and to understand how these will affect the assessment produced.

In addition to these mainly practical considerations, there are then more technical, but no less important, issues such as those of validity and reliability. It is not intended here to debate all of these issues in depth but simply to indicate the importance of validity and reliability in particular. Many publications deal with the definition and measurement of these attributes and further information may be found, for example, in Anastasi (1968), Ebel (1972), Guilford (1973), Nuttall and Willmott (1972).

3.2 Validity

Validity, as might be expected, is about how valid an assessment tool is for a particular purpose but even this simple statement opens many queries about understanding just what is being defined. A test of the recognition of road signs that had been developed in the UK, for example, might be very appropriate for use in the UK, do the job well and so be regarded as a valid test. If the same test were to be given to a group in another country, however, it might well be judged not to be so appropriate and, accordingly, much less valid. The validity of a test thus depends both on being clear about the ideas that lie behind its construction (i.e. what it is intended to measure) and also depends on who takes the test, for what purpose and in what context. A general definition of test validity is that a test needs to be able to measure what the test constructor set out to measure when applied to a particular group of people in a particular context.

Here too interpretation comes into play. Is it always possible to be clear precisely what the test constructor (often a committee!) wished to measure? Is it also clear just what the test does actually measure in practice when used with a particular group and context? These questions can rarely be answered with a definitive response and so judgment generally comes into play to some degree when considering validity.

There are many forms of validity but three in particular are of relevance here. <u>Content Validity</u> is a means of considering whether a test has met the aims of its creators by measuring specific content. This form of validity is often used in conjunction with tests covering specific syllabus objectives (in a test of English, for example, has grammar been assessed to the intended degree?). Then there is the more generic <u>Face Validity</u> which asks whether the test looks as though it should measure that which it is desired to measure – does the test look OK? These two measures of validity can be considered when tests are being developed and before a test is actually used, thus focusing attention on the required outcomes.

Then there is <u>Predictive Validity</u>. This form of validity is used where the results of testing are to be used to predict some specific subsequent outcome. For example, an assessment of bodily fitness may be undertaken for an individual and then a prediction made of the time that would be taken for that individual to run 1000 m.

The supposition would be that the fitter the person (i.e. the higher the measure of fitness), the faster that person would complete the 1000 m run. Here, a direct measure of validity is obtained as the fitness assessment (however made) and the time for the run would both be known as quantitative measures. In such cases, predictive validity is measured by the correlation achieved between the results of the (fitness) test used and the subsequent results (time for the run) that were to be predicted; validity is assessed directly and is not a judgmental assessment. Judgment would, however, come into play when assessing whether the fitness test was sufficiently accurate to predict the run time; and this judgment would depend on many factors, not least whether the measure helped 'at all' in the prediction.

With predictive validity, as with correlations in general, the results (the value of the correlation achieved) can be affected by the distribution of the measures being correlated. As one of the variables becomes more skewed, with very many more high scores than low scores on one variable, for example, so the value of the correlation derived from the data will become lower than the true value. This fact is very relevant when looking at the prediction of academic success as the students whose results are being correlated are a highly selected group from the population of all students. This matter will be raised again when looking at the predictive results of the TSA.

3.3 Reliability

Reliability is a feature of assessments that reflects the accuracy of the results obtained. Would a student be likely to get a similar result when sitting a similar examination or test, for example, next time? This is a far from easy idea to pin down as people change over time, a 'similar' test is never the same as the original and conditions change but it is necessary to have some idea about the accuracy of an assessment.

The main difficulty in dealing with test reliability is finding a way to measure this important characteristic. As with validity, there are a number of different types of reliability, each having a different formulation to assess the concept that is being measured.

In order to look at the reliability of the TSA, measures of internal consistency reliability have been used. Here, the test performances of students are looked at across the test as a whole in order to judge whether the responses form a consistent pattern. Crudely, if a test is divided into two parts, to what extent do candidates get the same score on each part? By looking at all possible 'split-half' correlations and averaging them an indication is obtained of how reproducible a set of test results is in practice.

3.4 Validity and Reliability

A final point must be made to underline the fact that validity and reliability are related, one to the other. If a test gives results that are reliable then it may, or may not, be a valid test. If a test gives results that are not reliable then there is no way the test can give valid results.

As an analogy, an archer may be considered who is shooting arrows at a target, perhaps as part of a competition. A valid shot in such a circumstance would be 'an arrow in the bulls-eye' of the target (assuming that this was the intention of the competition). A reliable performance would be 'all arrows in the same place'. If the archer is reliable and all arrows fall in the bulls-eye then the archer is giving a valid performance (each shot is valid). But if the archer is not reliable, and the shots are all over the place, then there is no way that this can be considered to be a valid performance. Yes, one shot might happen to go into the bull but the next could miss the target completely; not quite what was intended. Equally well, if the archer were shooting reliably and all of the shots were falling in the same place, if that place happened to be the post holding up the target then there is also no doubt that the archer is not giving a valid performance (and no shots were valid).

There is thus a clear interplay between reliability and validity. Without reliability a test cannot provide valid measures of performance; but even the most reliable test is not much use if it is not measuring what is required in the circumstances – i.e. it is not a valid test. Of course, in practice no test can be perfectly reliable and so too no test can be perfectly valid. Further, the degree of reliability achieved will govern the degree of validity that can be achieved.

So, if a test is ‘reasonably reliable’ then it may, or may not, be ‘reasonably valid’. As with the archer, even if the TSA is found to be as reliable as might be expected from such a test, there will be no guarantee that the results will provide a good prediction of success at university.

As with many judgments, once the investigations are completed the task will be to decide if a test is reliable enough and valid enough for the purpose for which it is being used. This is never a clear cut or easy decision to make and in the end this may simply come down to a decision on whether a test has been found to be useful.

4 The Reliability of the TSA Tests

4.1 The Early Development of Thinking Skills

As has been mentioned earlier, the current work on Thinking Skills was based very much on previous work. Much of that work was not reported formally but some reports are available and these are referred to both here and in Section 6.

- An evaluation of Assessing Argument Questions (Thomson and Fisher, 1993) showed that out of 30 questions that were investigated, 20 '… performed well in the sense that subjects needed to reason in the way intended in order to get the right answer'. In addition, information was gained about why the remaining 10 questions were judged to be unsatisfactory (two had confusing wording, five had a misleading distractor and three were assessing comprehension not reasoning).

- A study of Formal Reasoning questions (Green, 1992) looked at 30 questions and found that 'Overall, most of the items functioned well' and that the evidence pointed to the fact that '…students found the items novel'. Nevertheless, there were lessons on item writing to be learned as 'Thirteen items were judged to be inappropriate items, either because students had difficulty interpreting them or because they were too difficult'.

Both of these studies provided useful information on the structure of the individual questions examined and on the development of the Thinking Skills scales generally. Being essentially small-scale, however, they did not attempt to estimate test reliability.

A much larger study involving several universities is reported by Massey (1994a). Here a substantial amount of data was collected, analysed and reported and the main findings in terms of reliability estimates are reported in Section 6. Most of the results do, however, agree with the later results reported below.

Of much more immediate use are the data collected over the more recent past.

4.2 CTSC and CATS

In all cases where TSA tests have been used, full response data have been collected from students. A record is thus available of responses to all questions in the test, including whether a question has been answered or not. These data have been used to give scores for candidates for reporting purposes and analysed to allow an evaluation of tests and questions. Evidence is thus available on how the tests as a whole performed and on any questions that can be identified that appear to be too hard (or too easy) or are causing candidates to answer them in an unusual manner.

During the development and operation of questions and tests of Thinking Skills within UCLES, there has been much pre-testing and subsequent item analysis. This process allows the necessary evaluation of tests and questions and also the estimation of the (internal consistency) measure of the reliability of the tests and pre-tests used.

Table 4.1 shows some of the item analysis data from the UCLES CTSC and the subsequent CIE CATS awards. As the number of questions in a test is not always the same, further estimates of reliability have been provided (using the Spearman-Brown prophecy formula – see Ebel, 1972), that estimate the value of the reliability that would be expected from a test with 50 questions.

Table 4.1: Reliability Estimates for CTSC and CATS Assessments – 1996-2003

Year	Award	Number of Tests	Number of Questions	Reliability	Reliability (50 questions)
1996	CTSC	2	44	0.82	0.84
1998	CTSC	2	44	0.87	0.88
1999	CTSC	1	44	0.81	0.83
1999-Pre.	CATS	3	50	0.83	0.83
2000	CATS	1	50	0.87	0.87
2000-Pre.	CATS	3	50	0.75	0.75
2001	CATS	2	50	0.77	0.77
2002	CATS	2	50	0.80	0.80
2002-Pre.	CATS	3	56	0.85	0.83
2003	CATS	2	50	0.80	0.80

It can be seen that with the data available the reliability estimates for the CTSC award ranged from 0.83 to 0.88 with an average figure of about 0.85 for a single test. With the move to the CATS award, three separate sets of pre-tests were held (see the ‘Pre.’ suffix in the Year column of the table) and these reliability estimates ranged from 0.75 to 0.83 with an average value of about 0.81 for a single pre-test. Finally, for the main CATS award, the reliability estimates range from 0.77 to 0.87 with an average of about 0.80 for a single test.

These results indicate that the CTSC was a somewhat more reliable examination than the CATS award. This is not so surprising as the CTSC had been based on substantial development work and the question types used in the tests had been subjected to considerable scrutiny during the development process. When the CATS award came into being, new questions were written and some were in a different format from those used in the CTSC. It is beyond the scope of this document to go into the changes that took place but as can be seen, the pre-testing and then the main tests were not quite so reliable as the CTSC had been.

4.3 TSA Pre-tests

When the decision was taken to introduce a trial of the TSA, the original CTSC questions were the only source of secure questions because those from the CIE CATS award were made available to candidates after the examination and so they could not be used again. These CTSC questions were thus used to provide the basis for building the first TSA tests. In later years, new questions were written but all questions were subjected to pre-testing. Table 4.2 provides the details of the reliability estimates obtained.

Table 4.2: TSA Pre-test Reliability Estimates

Year and Session	Number of pre-tests	Number of Questions	Reliability	Reliability (50 questions)
2001/1	8	50	0.88	0.88
2003/1	8	30	0.69	0.78
2003/2	16	30	0.64	0.75
2004/1	6	30	0.62	0.73
2004/2	12	30	0.64	0.75

As may be seen from the above table, the pre-tests started off in 2001 as being very reliable (0.88) but this value fell in subsequent pre-tests. This was in many ways unavoidable as in 2001 all good CTSC questions were used and the pre-test was basically a confirmation of what was known already. In later pre-testing the mix of revised questions to good old questions grew and finally a substantial number of new questions were pre-tested.

4.4 TSA Tests

The TSA tests used during the admissions process were thus built from questions with known characteristics and based very much on the CTSC. The reliability estimates of the tests used are given in Table 4.3 together with other test statistics. These statistics are based on raw test scores (the 'number correct' scores) of candidates although during admissions a scaled score was used in order to remove any differences in difficulty between tests, thus creating comparable scores.

Table 4.3: TSA Reliability Estimates

TSA Test	No. of Quest s	No. of Cands.	Mean Score (%)	Standard Deviation (%)	Reliability	Reliability (50 questions)
A	50	153	62.9	15.4	0.84	0.84
B	50	138	61.1	15.1	0.84	0.84
A-FE	50	58	58.9	16.6	0.86	0.86
C	43	235	50.1	14.1	0.76	0.79
D	46	243	50.4	16.0	0.83	0.84
E	50	522	66.6	13.5	0.81	0.81
F	50	583	71.6	14.4	0.85	0.85
G	50	89	68.8	13.4	0.81	0.81
H	50	86	59.1	12.8	0.80	0.80
J	50	83	66.9	13.5	0.82	0.82
K-FE	50	160	63.6	13.0	0.79	0.79
M	50	382	67.7	15.0	0.85	0.85
N	50	229	68.0	13.6	0.82	0.82
O	50	822	66.2	14.5	0.84	0.84
P	50	529	64.9	15.4	0.84	0.84
Q	50	185	64.2	15.8	0.86	0.86

Notes:

2001	Tests A and B
2002	Test A: this was administered in the Far East. Tests C and D; the tests actually administered had 50 questions each but were revised after the event to remove unsatisfactory questions.
2003	Tests E and F; these tests were administered using pencil and paper. Tests G, H and J; these tests were administered on-line. Test K; this test was administered in the Far East.
2004	Tests M, O and P; these tests were administered both on-line and using pencil and paper. Tests N and Q; these tests were only administered on-line.

As can be seen, there are some differences between the tests in the way that they were answered:

In 2001, the tests were very similar with mean scores in the low 60s and a standard deviation that indicates a very good spread of marks.

In 2002, the tests administered in the UK were much harder and candidates were scoring only just over 50, some 10 points lower than the previous year. The spread of marks was not the same for each test either with that for Test D being somewhat greater that that for Test C.

In 2003, however, the mean scores were again generally around the upper 60s and the standard deviations a little smaller than in 2001 indicating a somewhat narrower spread of marks. However, given that the tests were of approximately equal difficulty, it would appear that a particularly less able group of students took Test H.

In 2004, the mean scores are all around the mid 60s with a somewhat wider spread of marks than in 2003.

The reliability estimates of the TSA tests in 2001 and 2002 are around 0.84 with those for tests used in 2003 being somewhat less at around 0.80. As has been noted, this was probably due to the introduction of new questions into the later tests. As the work has progressed, questions have been developed further and the reliability figures have now risen somewhat for 2004 to about 0.84.

4.5 Other Uses of TSA

In addition to the main testing sessions, TSA tests were also used in different situations and an analysis of these data is shown in Table 4.4.

Table 4.4: Reliability Estimates for Other Uses of TSA Tests

TSA Test	No. of Quests	No. of Cands.	Mean Score (%)	Standard Deviation (%)	Reliability	Reliability (50 questions)
A1	40	365	48.9	14.8	0.77	0.81
B1	41	28	54.7	12.1	0.67	0.71
L	50	33	76.6	9.2	0.65	0.65

Notes:

2003 Test A1; a reduced version of Test A was used in an investigative project with new undergraduates in the UK.
Test B1; a reduced version of Test B was used in China.

2004 Test L; a ‘research test’ was used with an undergraduate sample in Cambridge.

As can be seen, the use of Test A/A1 with different groups of candidates gave some differences in mean score but with similar reliability estimates. The use of Test B/B1 in China was not particularly satisfactory but the number of candidates on which the analysis is based is very small. The use of Test L, however, indicated a very strong group of candidates and the smaller standard deviation indicates that the scores were bunched up towards the top of the distribution. In such a situation, it is perhaps not surprising that the reliability estimate is somewhat lower than expected.

4.6 The Reliability of TSA

It has been noted that the internal consistency reliability estimates for the TSA are of the order of 0.80 and that the initial and later tests were a little more reliable than this (about 0.84). The current test reliabilities probably reflect the benefits of recent development work and this will be of benefit to all concerned as results obtained will be more able to correlate with subsequent university achievement.

The TSA is thus a reasonably reliable assessment, a condition that must be met if it is to be used as a predictor of University achievement.

5 Issues of Validity

5.1 Some Early Findings

The early studies of the MENO tests were concerned mainly with the behaviour of the results of assessment between parts of the tests used and with the relationship of the assessments with academic results. References to the internal analyses are made in Section 6 but a comment on the tests shows their standing in relation to other measures.

Some of the MENO components were used in Singapore in 1996 and test measures were then available alongside the usual academic measures of performance. It was found that while correlations with individual academic measures were not large, those with UCAS points were '… on the high side of those normally observed for correlations between aptitude and achievement measures.' (Massey, 1997).

5.2 The Use of TSA

The TSA has been introduced to the University admissions process on a trial basis as a possible aid to selection. There is no suggestion that such a test should ever replace the existing information used during admissions (e.g. A-level grades/marks, interview results, school reports, etc.) and the emphasis behind the work is very much on providing supplementary predictive information. The likely value of the TSA tests has been based on a substantial background of work in the area of testing skills deemed to be useful for Higher Education but there is, as yet, little information available on the degree to which the TSA is able to predict success at university.

This success is generally agreed to be performance in first-year university examination results: to attempt to predict final degree performance after three years where students change considerably in their approach to many aspects of their life would be a hard task indeed.

5.3 TSA Scores as Predictors of Achievement

So far, there have been few possibilities for collecting data on the predictive powers of the TSA. The students involved in the first administration of the TSA tests in December 2001 who were selected for University entry came up the following autumn (2002) and achieved their first-year results in the summer of 2003. As only some 289 students were tested, the selection ratio being what it is means that only 48 of those were made offers, arrived in Cambridge and had results at the end of the first year.

For the second cohort of students, from the December 2002 testing, 472 students were tested, and results were available for 91 candidates in their first year. So the sample of candidates for whom first-year results were available in the summer of 2004 was about double that from the 2001 sample.

Of some 1551 students who were tested in December 2003, 465 were made conditional offers and a further 25 were made unconditional offers. Those students who took up their places are only now in their first year at University and so results on their (first year) University performance will not be available until the summer of 2005.

Accordingly, apart from the data from the initial trial described in Section 1.3, there are only the two sets of data mentioned above that are available for analysis.

Table 5.1 provides the correlations between the TSA scores and the first-year performance in the three cases where data exist and are taken from Robinson (2002) and Forster (2004).

Table 5.1: Correlations of TSA Scores with First Year Results – 2001-3 and 2002-4

Student Cohort	Number of Students	Correlation with First Year Results
Initial Trial	35	0.30
2001/2003	48	0.30
2002/2004	91	0.27

None of these correlations is particularly high but neither can they be ignored. In the context of studies in America, the US Department of Labor (1999) rates such a correlation as 'likely to be useful'

The sample numbers on which these results are based are very small and the correlations of TSA scores with University examination performance are weak but there is some indication that those for A-level scores and Interview Scores are even weaker (Robinson, 2002). The problem is that A-Level scores do not discriminate between a field of candidates all of whom have three A grades and the absolute values of Interview Scores may differ between interviewers even when they produce the same order of ranking. In this case at least, the TSA is providing a greater degree of prediction than are the other measures currently in use.

Unfortunately additional data, such as A-Level grades and Interview Scores, are not currently available in the case of the TSA candidate cohorts in 2001/2003 and 2002/2004. It is clearly important that such data are collected in future exercise in order to evaluate fully the relative benefits of the TSA and existing measures.

The validity of the TSA can be looked at from two standpoints. From the first, the test is well-founded, being based on a long development to assess the skills that are seen to be important in Higher Education (see Fisher, 2005). Here the context in which it is used is very competitive as applicants are selected directly for a three-year degree programme. In this sense the TSA is a little different from the Scholastic Aptitude Test (SAT) used in the US which is not aimed at assessing Thinking Skills. The SAT is used with applicants having a much wider range of ability and selection consists of being admitted to the first year of a four-year degree programme, with further selection being made at the end of the first year for a subsequent three years of a degree course.

There are many reasons why an assessment of Thinking Skills such as the TSA is more appropriate for use in the UK than a test such as the SAT and Fisher (2005) discusses the origins of both assessments, linking the development of the SAT to the needs of the US '… after the second World War, when admissions to Higher Education were greatly expanded'. He then compares the educational systems of the US and the UK and argues that to use a test such as the SAT in the UK to provide supplementary information for admissions would be 'a mistake' and provides a number of reasons why this would be the case.

From the second standpoint, the TSA is seen to offer something by way of predicting success. Correlations such as those above will never be high because of both the nature of what is being assessed and also the quite restricted range of ability of students involved; in the context of a highly selected group of students, however, the results are promising. In comparisons with studies in the US, for example, the correlations may not look good but when the range of ability of students is considered they appear in a much better light.

A discussion on the future investigation of validity is given in Section 8 but there are further results to help in the understanding of how Thinking Skills scores relate to University achievement.

5.4 Prediction by TSA Score Sub-scale

The TSA is made up from an equal mix of Critical Thinking (CT) and Problem Solving (PS) questions and, for the purposes of the analyses so far, the test results have been considered as a single score. If the scores are broken down, however, so that each student has a CT score and a PS score, each of these can be used as a means of predicting University performance. Table 5.2 shows the predictive power of the CT scores and the PS scores separately for the 2002-4 samples.

Table 5.2: Correlations of CT and PS Scores with First Year Results – 2002-4

Student Cohort	Score Used	Number of Students	Correlation with First Year Results
2002/2004	CT	91	0.13
2002/2004	PS	91	0.27
2002/2004	CT+PS	91	0.27

These results immediately suggest that whatever the value of the TSA as an instrument in selection, there may be a difference between the value of the CT and PS scores when it come to prediction of University achievement. Indeed, it might be tempting to say that the CT scores are adding little or nothing to the predictive ability of the PS questions.

At this point two observations may be made. First, this is a result from a single year only and one that needs to be compared with analyses of data from other cohorts. The second is that many of the students in the 2002-4 cohort were applying to read Computer Science, Economics, Natural Science or Physical Science; as such it may not be surprising that a measure of PS, based largely on numerical problems, provides a better predictor than a measure based more on logical thinking with words.

However, another user of a TSA test, outside Cambridge, also reported that the CT sub-scores did not aid in the prediction of the first-year results as did the PS sub-scores (see Section 5.6). There is thus evidence to suggest that it would be sensible to look at the sub-scales of the TSA separately and to investigate the way in which they both operate. The start of such an investigation is reported in Section 6.

5.5 The use of TSA in the Admissions Process

In 2003, following the administration of the TSA tests, a questionnaire was distributed to those involved in using the TSA during the Admissions process. The aim was to investigate the ease of use of the administration system generally. Included in the questionnaire were two questions that sought information of the perceived use of the tests. Out of 16 respondents, 15 indicated that the TSA was 'useful in the admissions process' and 13 indicated that they would 'favour trials in more subjects'. It thus seems that the TSA has received at least a basic acceptance with users.

At the same time, it is possible to follow through the test takers and to consider their results by the offers made. Table 5.3 is taken from Forster (2004) who provides a great deal of detail on the scores of various groups of candidates on the test. The table looks at scores on the two parts of the test as well as on the whole by the application decision made in December 2003.

Table 5.3: TSA Score by Application Decision – 2003 Cohort

Selection Decision	No.	Critical Thinking		Problem Solving		Overall	
		Mean	SD	Mean	SD	Mean	SD
Conditional	465	64.24	9.80	67.54	10.73	65.19	8.04
Unconditional	25	63.50	11.42	65.80	9.75	64.09	7.91
Reject	1038	56.97	9.09	60.57	9.14	58.54	7.39
N/A	23	49.03	9.22	59.42	12.18	53.54	7.98
Total	1551	59.14	10.00	62.73	10.21	60.55	8.24

It should be noted that admission result data were not available for 23 applicants. Also, the Unconditional offer group contained only 25 applicants and this must be borne in mind when considering means for this group of candidates.

On the test as a whole as well as on the sub-scales, the group with the highest means was the Conditional offer group, followed by the Unconditional offer group and then the Rejected group. Also, as might be expected, the mean scores of the two groups of candidates to whom offers were made were quite close when compared with the mean score for the group to whom no offer was made. The group for which no decision was available were also consistently the lowest scorers but little can be made of that fact.

Thus, although the TSA was not being actively used in selection, the scores by candidates consistently reflected the admissions decisions made. Clearly this result cannot be taken as evidence of the test being a useful predictor of University achievement as there will be at least some element of 'self-fulfilment' in these results. Nevertheless, for a test that is being used as a trial alongside the usual selection methods, the results are encouraging.

In the light of the lack of predictive ability of the CT sub-scores in 2002-4 when compared with that of the PS sub-scores, it is perhaps surprising to see that the differentiation in marks between the Conditional offer and Reject groups of applicants using the CT scores (7.27) was somewhat greater than that using the PS scores (6.97). It will remain to be seen whether the relative predictions found in the 2002-4 cohort are also found in the 2003-5 cohort to which the above data relate.

A similar analysis was carried out for a specific subject (Subject B), applicants. Although these data are a sub-set of the main 2003 data reported above, Table 5.4 shows how the mean scores of applicants related to the selection decisions made (UCLES, 2004b). Here, a Pool candidate is one who is does not achieve his or her first choice of College but is not rejected outright.

Table 5.4: TSA Score by Application Decision: Subject B Applicants – 2003 Cohort

Selection Decision	N	Critical Thinking		Problem Solving		Overall	
		Mean	SD	Mean	SD	Mean	SD
Offer	69	62.71	9.95	68.06	11.21	64.78	8.92
Pool	87	58.37	11.12	64.39	11.70	60.92	9.43
Reject	115	55.05	6.59	58.66	8.78	56.66	6.29

It can be seen that in the case of Subject B applicants the TSA scores and sub-scores again relate well to the selection decisions made. In the case of the Pool candidates, and especially with the CT sub-scores, there is a wider spread of scores than with either of the other two groups of applicants. This spread may well reflect the fact that some Pool candidates may not fit the profile for the particular College to which they applied but may nevertheless be good applicants while others fall short generally. As Subject B is a 'Scientific' subject, it is also hardly surprising to note that for these applicants the TSA PS sub-scores are substantially higher than the CT sub-scores and that, relative to all applicants (see table 5.3), the PS scores are higher and the CT scores are lower for candidates receiving an offer.

Again, this information is not evidence that the TSA is working as a predictor of University achievement although it is an indication that this may be the case.

As part of the work on the 2003 data, Forster (2004) reports the correlations between the CT and PS sub-scores and the total score with the 'Outcome Decision'. These results are given in Table 5.5.

Table 5.5: Correlations between TSA Sub-scores and Application Outcome

	Problem Solving	Total	Outcome

Critical Thinking	0.50	0.86	0.34
Problem Solving		0.86	0.32
Total			0.38

The inter-correlation between the two sub-scales will be picked up in Section 6.4 but the information of interest here is the difference in the correlations between the sub-scale results and the selection outcome. It can be seen that the CT sub-scores correlate at least as well with the outcome than do the PS sub-scores. This is contrary to the findings above using first-year results and underlines again the need for careful investigation when the data for the next cohort become available.

5.6 Some Comments from Users

Also of some relevance to the validity of using TSA are comments from users. Although these do not in themselves constitute evidence of predictive validity, if the users find the results useful then this is a helpful commentary on the tests used. Some general comments on TSA results being ‘useful in Admissions’ have already been mentioned but the following is a comment made by one user in the University.

> My general view is that the TSA is our best indicator of Tripos performance. It does not correlate particularly well with other indicators such as exam results and interview scores, but that just suggests that they are inaccurate. However, I would be nervous about using the TSA for pre-filtering of candidates. That would undoubtedly give rise to some injustice. It should be used as just one of several measures.

Another user of TSA, from outside the University, made the following points by way of a summary.

> The overall predictive validity of the measure, whilst not high, is within the expected range and an appropriate range of scores was obtained in the undergraduate sample. Thus, if the subgroup differences could be addressed the test could be useful in handling competition for places amongst high ability applicants.
>
> This work confirmed that many teachers and academic staff are interested in the concept of critical thinking and feel instinctively that it may well be relevant to their programmes. However, they seemed largely inexperienced in the use of the concept and were not aware of how they could get further information (e.g. website). Similarly, tutors and teachers seemed not to perceive many differences between tests. For us as admissions staff, before using a measure for live admissions we would certainly need to raise awareness of what different tests measure so that tutors could make an informed decision about what constructs best related to their programme aims and curricula.

The matter of bias in the TSA results is of some concern and an investigation of the data collected in Cambridge is reported in Section 5.7.

5.7 Bias in TSA Questions

One of the most problematic issues in testing is the writing of test questions that are unbiased. A biased question here is a question that may be answered correctly, or incorrectly, in a systematic manner by candidates with an identifiable characteristic that does not happen with all candidates. The categories that can be checked for bias are many but the most usual are gender and ethnic origin.

In the early work on Thinking Skills, Massey (1994b) reports analyses of a Problem Solving test which suggest that ‘… MENO is unbiased with respect to first language, ethnic origins and age …’ but also that ‘… the sexes do appear to have performed differentially on MENO PS …’.

In the case of the TSA, an initial study of bias has reported that:

> From all that has been found, however, it is clear that for the variables investigated there is no substantial and consistent bias in the case of most of the TSA questions used. In particular, apart from one question, no evidence of any difference between a paper-and-pencil administration and an on-line delivery has been found. (Willmott, 2004).

Despite the fact that there was little evidence of bias found in the TSA questions, this is an area where more work needs to be done. In particular, the analyses conducted only relate to a single session of TSA and need to be repeated as more test data accumulates. It is also necessary to be clear on matters of question and test bias, as opposed to differential performance, as they are not the same.

The variables included in the TSA bias analyses against which questions were checked were:

Method of Delivery	On-line and Paper-and-Pencil
Gender	Male and Female
Subject	Chosen Subject of Application
Location	Home and Overseas
School Type	Main UK School Types
Decision	Offer and Reject

One important variable that was not available for study was Ethnic Origin and this is an omission that needs to be rectified as soon as possible. Part of the difficulty here is simply the problem of collecting the data and the operation of the Data Protection Act. A further and very important part of this difficulty stems from the considerations from both UCAS and the Universities concerning the extent to which ethnic origin information should be revealed to admissions selectors. While no-one wishes to influence admissions unfairly, without collecting data on ethnicity to ensure fair assessments it will not be easy to check for any ethnic bias in questions.

Also, work could usefully be carried out on a wider range of candidates in order to have sufficient data to look properly at question bias. The report mentioned also suggests other ways in which question bias could be investigated.

Nevertheless, it is encouraging to note that in these early analyses little evidence of question bias has been found.

6 Critical Thinking and Problem Solving: One Skill or Two?

6.1 The Two Sub-scales

In the discussion of Thinking Skills and its assessment, and apart from a brief discussion in Section 5, the scores on the two types of questions used - those that assess Critical Thinking skills and those that assess Problem Solving skills - have been looked at together. The results of the assessment using these two types of questions have simply been added together to provide a single score. By so doing, any differences that exist between scores on these two sub-scales may be hidden. As the questions potentially assess different skills, some basic analyses ware carried out to see whether any such differences could be found.

To investigate the relationship between the Critical Thinking and Problem Solving scores of students, the data held for each test were split to form two sub-tests (one with CT question data and the other with PS question data). These tests were then analysed separately and the results of these analyses compared with those obtained from analysing the test as a whole.

Before presenting the results of these analyses, it is worth speculating on what differences might be expected when the results are compared if the CT and PS questions are actually assessing different aspects of Thinking Skills. Here, it is not a matter of whether the two sets of questions look different, or appear to assess different skills but whether the skills assessed by the questions are actually different.

An analysis focussed on a sub-test of questions, such as ‘all CT questions’ or ‘all PS questions’, looks at a particular question in the context of other questions that are classified as being all of a similar type. Thus the context in which the question is analysed becomes noticeably more similar to the question than is the case when looking at the same question in the context of the test as a whole. If CT questions and PS questions actually assess different skills, then a question may thus be expected to be more like all of the other questions around it for the purpose of the analysis in the sub-test as opposed to the whole test. However, if the CT and PS questions are actually measuring skills that are very similar despite their apparent differences then it is unlikely that any differences will be found.

A result of any coherence that exists would thus lead to an expectation that the students’ scores would be more ‘consistent’ across different questions than was the case in the original test in which a greater diversity of questions appeared.

This increase in coherence is certainly real in terms of the question types as this is how the two sub-tests are formed but if the coherence is also real in terms of what is being assessed, then two important effects on the statistics in the more focussed analyses are likely to be seen.

- First, the (internal consistency) reliability estimates of the sub-tests will be expected to be greater than that for the test as a whole as the tests analysed are more consistent within themselves.

- Secondly, the question discrimination indices, the correlations of scores on a question with scores on the total score on the test (full test or reduced test respectively) can be considered. With greater coherence these correlations would also be expected to be higher in the sub-test, again reflecting the increased ‘likeness’ of all questions, thus causing the question responses to be more like the total scores on the sub- test.

These two criteria are now used to consider the results of the analyses of TSA data.

6.2 Early Results

In an extensive study covering a number of Universities, Massey (1994a) found reliability estimates for CT and PS tests individually and further information is available from the use of MENO in Belgium in 1995 (Nonneman et. al, 1995). These results are summarised in Table 6.1.

Table 6.1: Estimated Reliability of TSA Sub-scores – Early Results

Sample	Section	Number of Questions	Reliability	Reliability (50 questions)
Univ. 1	CT	31	0.70	0.79
Univ. 1	PS	42	0.83	0.85
Univ. 2	PS	42	0.83	0.85
Univ. 3	PS	42	0.83	0.85
Univ. 4	PS	42	0.87	0.89
Univ. 5	CT	54	0.86	0.85
Univ. 5	PS	42	0.82	0.84

From the above results it can be seen that the reliability estimates for the CT sub-scores are 0.79 to 0.85 for a 50-question test. By way of comparison, the reliability estimates for the PS sub-scores are more consistent (at about 0.84/0.85) but with one higher value (0.89). These values are only very marginally higher than those found for the full TSA tests.

6.3 Results for TSA Analyses

The results of the sub-scale analyses of TSA data are shown in Table 6.2. In the table, each of the three columns for Critical Thinking and Problem Solving gives the number of questions in the respective sub-test, the average change of the values of the discrimination indices across the questions in the test and the estimated reliability for the sub-scale.

When a discrimination index is calculated for a question, the usual statistic used is the correlation between the responses on the question (one or zero in the case of multiple-choice questions marked as either correct or incorrect) and the total scores on the test as a whole. To the extent that the scores on the test as a whole include the score on the question, this correlation is biased as the question score cannot help but affect the total score, albeit to a ‘small degree’.

In order to be sure that in the investigation of the two sub-scales any effects due to this bias are not causing unwarranted interpretations, for the purposes of the table below unbiased correlations have been used. These correlations have been calculated between the scores on a question and the scores on the total test score without the question included. The resulting correlations will thus be unbiased and represent the best estimates of the degree to which the question scores and the total test performances are related.

Any differences in question discrimination found between a whole test or a sub-test will now be more likely to reflect any real differences between what is being assessed by the respective tests.

The differences in change in question discrimination are shown in favour of the sub-scale. For the reliability estimates, however, as each sub-test is shorter than the original (about half the length in general), the estimated value of the reliability estimate has been based on what would be expected if the test consisted of 50 questions, thus allowing sensible comparisons to be made.

Table 6.2: Summary of Sub-Scale Analyses for TSA tests

Test/	Critical Thinking			Problem Solving			Whole Test		No. of
Year	No. Qu.	Disc.	Rel. 50	No. Qu.	Disc.	Rel. 50	No. Qu.	Rel. 50	Cands.
A/2001	25	0.02	0.88	25	0.02	0.87	50	0.84	153
B/2001	25	0.02	0.89	25	0.01	0.83	50	0.84	138
A-FE/2002	25	0.03	0.91	25	0.04	0.88	50	0.86	58
C/2002	20	0.00	0.77	23	0.01	0.85	43	0.79	235
D/2002	23	-0.01	0.85	23	-0.01	0.85	46	0.84	243
E/2003	25	0.00	0.82	25	0.00	0.84	50	0.81	522
F/2003	25	0.00	0.86	25	0.01	0.88	50	0.85	583
G/2003	25	0.00	0.81	25	0.00	0.85	50	0.81	89
H/2003	25	0.00	0.82	25	0.00	0.80	50	0.80	86
J/2003	25	0.02	0.88	25	0.00	0.80	50	0.82	83
K-FE/2003	15	0.02	0.84	35	0.00	0.80	50	0.79	160
M/2004	25	0.00	0.88	25	0.00	0.86	50	0.85	382
N/2004	25	0.01	0.86	25	0.01	0.82	50	0.82	229
O/2004	25	0.01	0.86	25	0.00	0.86	50	0.84	822
P/2004	25	0.01	0.85	25	0.01	0.88	50	0.85	529
Q/2004	25	0.00	0.86	25	0.01	0.89	50	0.86	185

In general, when the two sub-scales are considered, they do tend to be somewhat more consistent than the whole test (i.e. the reliability is greater). There are a few cases where the reliability of a sub-scale is less than that of the whole test (PS in Test B, CT in Test C and PS in Test J) but these are not typical. The increases in reliability estimates are small (about 0.02 for both CT and PS questions) and do not lead to any significant conclusions.

It can also be seen that the differences in question discrimination are very small indeed. These results do not suggest that the CT and PS questions are in any way assessing skills that are substantially different.

It has to be said, however, that these analyses are not the most sensitive with which to explore the dimensionality of test data. The use of techniques such as Factor Analysis are much better placed to investigate the nature of what is being assessed by the two sets of questions.

Nevertheless, from these results, although the sub-scale reliabilities are marginally greater than those for the whole tests, there are no real indications that the CT and PS questions are assessing skills that are substantially different.

6.4 Sub-scale Inter-Correlations

The nature of the sub-scales can be investigated further by looking at the correlations between the various sub-test scores. Table 6.3 gives details of the correlations between the CT scores, the PS scores and the total scores for candidates by test.

Table 6.3: Correlations between the Sub-Scales of TSA Tests

Test/	Correlation			Rel. 50 – Test
Year	CT v Total	PS v Total	CT v PS	
A/2001	0.86	0.85	0.46	0.84
B/2001	0.89	0.83	0.50	0.84

A-FE/2002	0.86	0.82	0.42	0.86
C/2002	0.92	0.96	0.77	0.79
D/2002	0.95	0.95	0.80	0.84
A1/2003	0.88	0.85	0.51	0.81
E/2003	0.87	0.87	0.51	0.81
F/2003	0.89	0.87	0.55	0.85
G/2003	0.88	0.88	0.55	0.81
H/2003	0.89	0.87	0.54	0.80
J/2003	0.89	0.85	0.50	0.82
K-FE/2003	0.74	0.93	0.44	0.79
M/2004	0.91	0.88	0.61	0.85
N/2004	0.88	0.84	0.48	0.82
O/2004	0.88	0.89	0.56	0.84
P/2004	0.87	0.90	0.56	0.85
Q/2004	0.88	0.90	0.58	0.86

The correlation between the two sub-scales is thus seen to be around 0.5 for the TSA tests, a figure that is consistent with that reported by Forster (2004) in a report covering data from all tests used in 2003 where a figure of 0.50 is reported. Further, Massey (1994a) also reports a value of 0.47 from the use of these two components in a University context.

This value is somewhat lower than might be expected if both types of questions assessed the same skills. At the same time, the correlations between the sub-scores and the total test scores are much the same for both CT and PS questions (at about 0.88) and this does not provide any major inference that the CT questions are generally measuring in a markedly different way from the PS questions. Again, the results for Test K reflect the different balance of questions in that test, thus inflating the PS/Total correlation and depressing the CT/Total correlation when compared with other tests.

6.5 The CT and PS Sub-Scales

There are seven different types of CT questions and three different types of PS questions, details of which may be found in Appendix A. There has been no recent work to look at the ways in which these different types of question measure similar, or different, types of skills, either by sub-category or over all question types. This is something that needs to be done so that the relationships between the measures made by different question types can be explored.

There are, however, two sources of such information from past work. Massey (1994a) analysed data from different question types in a University context in the UK and, while not all Universities tried all assessments, sufficient information was collected to look at the inter-correlations of scores from different types of question.

In addition, UCLES mounted a joint exercise with two Flemish Universities to investigate the use of MENO in selection and in a report of this work correlations between the results of students on different types of questions were investigated (Nonneman, et. al., 1995).

For the CT sub-scale, Massey found that the inter-correlations of scores for different question types varied considerably and were often quite low. However, not all of the question types used were the same as those used in the TSA. The results reported by Nonneman, et. al. showed greater stability, with values of around 0.5 being typical.

For the PS sub-scale there was greater stability. Massey found correlations of about 0.48 between scores on different question types and Nonneman reports values in excess of 0.8.

Clearly these results can only be suggestive of the position with sub-scores derived from the TSA but early indications are that the PS sub-score might be more coherent, i.e. reliable, than that based on CT questions. To the extent that this is so, then this might enable the PS score to correlate more highly with subsequent achievement than the CT score.

Finally, Massey (1994a) reports the results of conducting a factor analysis of the data collected and although the number of students was not large (only just over 100), there was a clear identification of two factors, the first of which contained the CT questions and the second the PS questions.

6.6 The Implications of the Findings

There is some evidence that the CT and PS questions are measuring skills that are not completely the same. The questions generally appear to cohere better when considered separately and the scores on the two sets of questions are only moderately correlated. That this is as intended – CT and PS questions were built into the TSA in order to assess different skills - is good but the extent to which there are real differences is not clear without further analyses. In order to investigate the dimensionality of the data in a more formal manner further analyses (e.g. Factor Analyses) need to be conducted.

In terms of making predictions of academic success at University, there is some evidence that scores from CT questions and PS questions might not be assessing Thinking Skills in exactly the same way. Whether or not this is the case, a linear combination of the sub-scores could be a better predictor of University achievement than either separately or, indeed, than the simple total score.

Accordingly, an investigation into the use of a combination, linear of otherwise, as opposed to a simple addition, of the two sub-scores as a predictor of achievement would be a useful step forward.

7 The MVAT and the BMAT

7.1 The Need for an Aid to Admissions in Medical Courses

During the later stages of the work by UCLES on the development of the assessment of Thinking Skills (around 1997/98), interest was expressed by the University of Cambridge in a test to aid in the selection of students for medical and veterinary courses. Here the number of applicants substantially exceeds the number of places available and an aid to selection could be helpful during the process of selection. It was thought that the work by UCLES in the area of Thinking Skills and selection might be useful to consider and discussions were held accordingly.

As a result, a trial of a Medical and Veterinary Admissions Test (MVAT) was administered in 1999. This test consisted mainly of Problem Solving questions. Following an analysis of the data collected, the first full version of the MVAT was then administered in November 2000 as a possible aid in the selection of students who wished to enter University in October 2001.

The MVAT had three parts. Part 1 consisted of questions, multiple-choice or short answer in format, assessing Scientific Aptitude. These questions were essentially Thinking Skills questions that had a generally scientific background (indeed, some CTSC questions that were in the early TSA tests were also used in the MVAT). Part 2 of the MVAT was also made up of multiple-choice and short answer questions but these assessed Scientific Knowledge. Finally, Part 3 provided students with an opportunity to show their level of scientific understanding through tackling one compulsory question and answering one further question selected from four presented featuring Chemistry, Physics, Biology and Mathematics. In this Part, credit was given for clarity of expression, the depth of understanding shown and evidence of a wide interest in the subject and students needed to write their responses to the questions in essay form. These responses were then marked by members of the University.

The MVAT was used again within Cambridge in 2001 and 2002 but in 2003 a number of changes were made: the University of Oxford and University College, London also used the test and it was re-named the Biomedical Admissions Test (BMAT) to cover its somewhat wider use. The BMAT was used again in 2004 with the added participation of the Royal Veterinary College and the University of Bristol Veterinary School. Details of the BMAT, including Specimen Papers and Marking Schemes, may be found at http://www.bmat.org.uk.

A detailed analysis of the operation of the MVAT and BMAT is given elsewhere (see Bramley, 2001; Massey, Shannon and Dexter, 2002; Shannon, Massey and Dexter, 2003; Massey, 2004; Shannon, 2004a), so only those aspects of the testing relating to the reliability and validity of the MVAT and BMAT will be mentioned here.

7.2 The Reliability of the MVAT and BMAT Assessments

7.2.1 MVAT and BMAT Part 1

Data are available for the MVAT/BMAT from the original 1999 trial through to its use in 2004. The reliability estimates for Part 1 of these assessments are given in Table 7.1 below.

Table 7.1: Reliability Estimates for Part 1 of the MVAT/BMAT Tests – 1999 - 2004

Year	Award	Number of Questions	Reliability	Reliability (50 questions)
1999	MVAT	25	0.72	0.84
2000	MVAT	20	0.69	0.85
2001	MVAT	19	0.66	0.84
2002	MVAT	17	0.67	0.86
2003	BMAT	40	0.73	0.77
2004	BMAT	35	0.73	0.79

It can be seen that the reliability estimates for the MVAT/BMAT ranged from 0.77 to 0.86 with the later BMAT tests being somewhat less reliable than the previous MVAT tests. In each case the BMAT tests had a small number of questions that, while satisfactory to the Question Writers and Editors, did not work very well in practice, thus lowering the reliability estimates.

7.2.2 MVAT and BMAT Part 2

The reliability estimates for Part 2 of the MVAT and BMAT assessments are given in Table 7.2 below.

Table 7.2: Reliability Estimates for Part 2 of the MVAT/BMAT Tests – 1999 - 2004

Year	Award	Number of Questions	Reliability	Reliability (50 questions)
1999	MVAT	44	0.75	0.77
2000	MVAT	30	0.83	0.89
2001	MVAT	30	0.79	0.86
2002	MVAT	25	0.79	0.88
2003	BMAT	30	0.72	0.81
2004	BMAT	27	0.66	0.78

It can be seen that the reliability estimates for Part 2 of the MVAT/BMAT, when estimated for a 50-question test, ranged from 0.77 to 0.89. The low figure for 1999 is likely to reflect the trial nature of the test and its questions but the values for 2000, 2001 and 2002 show tests giving results that are quite reliable. In 2003 and 2004, however, the values are somewhat lower than would be expected. Investigating the individual questions for 2004, for example, it appears that there was one very poorly discriminating question and three more that did not discriminate very well and these effects will have depressed the overall estimate of (internal consistency) reliability.

7.2.3 MVAT and BMAT Part 3

As Part 3 of the MVAT/BMAT consists of open-ended essay questions that are individually marked, no information is available here on the reliability of this part of the assessments.

7.3 MVAT and BMAT Scores as Predictors of Performance

As with the TSA, there have been few occasions to carry out a full evaluation of the MVAT/BMAT as predictors of performance.

An overview of the level of prediction is provided by Forster (2004) where the predictive validity of the MVAT scores from 2000 to the 2002 Tripos results was 0.30. This correlation is for all students and is for the Part 1 (Thinking Skills) section of the MVAT only. Massey (2004) provides more detail and this is summarised in Table 7.3 below.

Table 7.3: Predictive Validity of MVAT 2000

Course	MVAT-1	MVAT-2	MVAT-3	Interview
Medicine	0.15	0.29	0.25	0.08
Veterinary	0.22	0.31	-0.01	0.22

It is clear from Table 7.3 that Part 2 of the MVAT is providing a better prediction than is Part 1 for the students concerned but also that the extended answer section (Part 3) and the interview are by no means consistent predictors. Massey provides a great deal of information at the level of individual courses and reports that 'the level of association observed between Interview Scores and achievement was disappointing'.

7.4 The Use of MVAT and BMAT in the Admissions Process

As with the TSA, it is possible to follow through applicants and consider the test scores achieved for the various categories of selection decisions made. Table 7.4 provides this information for the MVAT in 2000, taken from Bramley (2001). It must be remembered, however, that, alongside other measures of achievement, the assessment scores are actively being considered for the purposes of selection.

Table 7.4: MVAT Scores by Application Decision - 2000

Selection Decision	N	Part 1		Part 2		Part 3	
		Mean	SD	Mean	SD	Mean	SD
Offer	367	12.38	3.70	29.34	4.46	11.11	1.96
Pool - offer	49	11.88	3.30	28.24	3.87	10.90	1.86
Pool - reject	160	11.88	3.39	27.57	4.51	10.43	1.65
Reject	921	9.56	3.16	22.85	6.36	9.70	2.02
Total	1497	10.58	3.57	25.12	6.39	10.16	2.06

As with the TSA, it is clear that the results on all Parts of the MVAT relate appropriately to the selection decisions made. In the case of the Pool candidates, Part 1 provides no differentiation but overall, all three Parts separate well the Offer, Pool and Reject candidates.

Table 7.5, also from Bramley (2001), shows the inter-correlations between the various parts of the MVAT and the Selection decision.

Table 7.5: Correlations - MVAT Scores and Application Decisions - 2000

MVAT 2000	Part 2	Part 3	Outcome
Part 1	0.47	0.20	0.30
Part 2		0.36	0.40
Part 3			0.28

As might be expected from the above discussion, Part 2 scores are the best predictors, followed by Part 1 scores and then Part 3 scores. It is also worth noting that the correlation between Part 1 scores and Part 2 scores (0.47) is very similar to that found between CT and PS questions with the TSA (0.50).

Tables 7.6 and 7.7 provide similar information for the MVAT administered in 2001 but only cover the results from Part 1 and Part 2 of the test (taken from Massey, Shannon and Dexter, 2002).

Table 7.6: MVAT Scores by Application Decision - 2001

Selection Decision	N	Part 1		Part 2		Part 1 + Part 2	
		Mean	SD	Mean	SD	Mean	SD
Offer	321	12.13	3.25	24.64	5.86	36.77	7.99
Pool - offer	74	11.45	2.89	24.18	4.46	35.62	6.09
Pool - reject	279	10.24	3.29	21.71	5.16	31.95	6.97

Reject	1042	8.42	3.09	17.02	5.68	25.44	7.52
Total	1716	9.54	3.48	19.51	6.45	29.05	8.83

Table 7.7: Correlations - MVAT Scores and Application Decisions - 2001

MVAT 2001	Part 2	Part 1 + Part 2	Outcome
Part 1	0.54	0.79	0.39
Part 2		0.94	0.43
Part 1 + Part 2			0.46

These results are much as found in 2001, although the Part 1 (Thinking Skills) score is a much better here predictor than was the case in 2000.

Tables 7.8 and 7.9 repeat the above information for the MVAT taken in 2002 (Shannon, Massey and Dexter, 2003)

Table 7.8: MVAT Scores by Application Decision - 2002

Selection Decision	N	Part 1		Part 2		Part 1 + Part 2	
		Mean	SD	Mean	SD	Mean	SD
Offer	340	11.65	2.63	22.02	4.57	33.67	6.18
Pool – offer	56	11.29	2.65	21.30	3.52	32.59	5.25
Pool – reject	250	10.51	2.52	19.85	4.15	30.36	5.50
Reject	1051	8.43	2.83	15.74	4.91	24.18	6.65
Total	1697	9.48	3.06	17.79	5.41	27.27	7.54

Table 7.9: Correlations - MVAT Scores and Application Decisions - 2002

MVAT 2002	Part 2	Part 1 + Part 2	Outcome
Part 1	0.55	0.80	0.38
Part 2		0.94	0.42
Part 1 + Part 2			0.46

The results here are not unexpected and the correlation with Outcome now shows the Part 1 score as being very similar to the Part 2 score.

Finally, Tables 7.10 and 7.11 (Shannon, 2004a) show the information for the (then new) BMAT taken in 2003.

Table 7.10: BMAT Scores by Application Decision - 2003

Selection Decision	N	Part 1		Part 2		Part 1 + Part 2	
		Mean	SD	Mean	SD	Mean	SD
Offer	328	5.71	0.68	5.69	0.81	11.40	1.29
Pool – offer	42	5.65	0.78	5.43	0.91	11.07	1.45
Pool – reject	223	5.26	0.64	5.41	0.72	10.67	1.05
Reject	1190	4.87	0.66	4.87	0.75	9.74	1.19
Total	1783	5.09	0.74	5.10	0.83	10.19	1.37

It must be noted that the scores on the MVAT examinations were reported as raw scores and these form the basis for the reports in Tables 7.4, 7.6 and 7.8. When the BMAT started in 2003, a new system of scaled scores was introduced that led to the smaller numbers that for the basis for the reports in Table 7.10 and 7.12.

Table 7.11: Correlations - BMAT Scores and Application Decisions - 2003

MVAT 2003	Part 2	Part 1 + Part 2	Outcome
Part 1	0.52	0.86	0.42
Part 2		0.89	0.35
Part 1 + Part 2			0.44

With 2003 results looking much as before, Tables 7.4 to 7.11 indicate that Parts 1 and 2 of the MVAT/BMAT are related to selection decisions in a constant and sensible way. After the first administration in 2000, the two Parts are very similar in the way they relate to the selection decisions for applicants although in 2003 the Part 2 score correlates less well with the Outcome measure than in previous years.

These findings are useful to know as the results presented in Section 7.3 showed that Part 2 results were a far better predictor of University achievement than Part 1 scores. It will be interesting see if subsequent analyses show that this was a feature of the 2000 testing or whether the part 2 score remains the best predictor.

7.5 The MVAT and BMAT Sub-scores

The results of the sub-scale analyses of MVAT and BMAT data for Part 1 of the assessments are shown in Table 7.12 below. The format is the same as for the TSA analyses reported in Section 6.3 where each of the three columns for Critical Thinking and Problem Solving gives the number of questions in the respective sub-test, the average change of the values of the discrimination indices across the questions in the test and the estimated reliability for the sub-scale. Again, unbiased question/total correlations have been used.

Table 7.12: Summary of Sub-Scale Analyses for Part 1 of MVAT/BMAT

Test/ Year	Critical Thinking			Problem Solving			Whole Test		No. of Cands.
	No. Qu.	Disc.	Rel. 50	No. Qu.	Disc.	Rel. 50	No. Qu.	Rel. 50	
MVAT2000	6	-0.04	0.79	14	0.00	0.89	20	0.85	1506
MVAT2001	7	0.00	0.89	12	0.00	0.85	19	0.84	1733
MVAT2002	5	-0.07	0.78	12	-0.01	0.89	17	0.86	1715
BMAT2003	14	-0.02	0.78	26	0.00	0.80	40	0.77	4099
BMAT2004	10	-0.02	0.83	25	0.00	0.81	35	0.79	4321

In the case of Problem Solving questions, the above results are similar to those found for the TSA. For Critical Thinking questions, however, the average discriminations are actually worse, or in one case the same, in the sub-test than in the whole test.

The main difference here is that the number of CT questions in Part 1 of a MVAT/BMAT test is proportionally much smaller than is the case with the TSA. This means that there are some very short CT tests with total scores based on only a few questions; as such, these total scores will not be very accurate estimates of candidates' skills and so the correlations may be small for this reason.

When the two sub-scales are considered, they tend to be generally more consistent than the whole test (i.e. the reliability is greater); the exceptions to this are the CT sub-scales from MVAT tests from 2000 and 2002.

From these results, as with those from the TSA analyses, while the sub-scale reliabilities are generally marginally greater than those for the whole tests, there are no indications that the CT and PS questions are substantially different in what they are assessing.

7.6 The Sub-Scale Inter-Correlations

Table 7.13 now shows how the CT and PS sub-scores for Part 1 of the MVAT/BMAT relate to each other.

Table 7.13: Correlations between the Sub-Scales of Part 1 MVAT and BMAT

Test/ Year	Correlation			Rel. 50 – Test
	CT v Total	PS v Total	CT v PS	
MVAT2000	0.75	0.95	0.49	0.85

MVAT2001	0.79	0.90	0.45	0.84
MVAT2002	0.75	0.96	0.53	0.86
BMAT2003	0.78	0.92	0.47	0.77
BMAT2004	0.74	0.93	0.45	0.79

The correlation between the two Thinking Skills sub-scales is again seen to be of the order of 0.5, as was the case for the TSA tests. The sub-score/total correlations, low for CT when compared with PS, reflect the small numbers of CT questions used relative to the PS questions (see Table 7.12).

Finally, Shannon (2004b) reports on a Factor Analysis of the BMAT data for the 2003 test. While there is much detail of interest in the results, the first factor is clearly a PS factor and the second is clearly a CT factor. The first factor accounts for almost 11 per cent of the total variance, the second factor just under 4 per cent and the correlation between these factors is 0.49. There are individual questions that load onto other factors but the support for the findings with the analyses of both the MVAT/BMAT Part 1 assessments and also with the TSA is substantial. As discussed in Section 6.3, it is clear that further work using Factor Analytic methods would help the understanding of the skills assessed by the BMAT Part 1 and TSA tests.

First, there do appear to be identifiable skills of Critical Thinking and Problem Solving and they do appear to be correlated by about 0.5. Also, it does look as though the Problem Solving skills, as assessed by PS questions, are more general skills than those of Critical Thinking.

Only one such analysis has been done and it is clearly important to repeat similar analyses with other MVAT/BMAT (and TSA) data. Nevertheless, the results here are very much in keeping with what has gone before with the TSA.

7.7 The Implications of the Analyses

Much of what has been found from the MVAT/BMAT Part 1 analyses has already been seen from the TSA analyses. There are differences but these are, generally, quite small as, indeed, might be expected. Of course, the BMAT has an extra two Parts over and above the TSA and, indeed, even Part 1 of the BMAT concentrates on PS questions as a means of assessing the Thinking Skills of applicants.

As with the TSA tests, there is a need for a full data set with as much information as possible to be available so that different analyses can be conducted to look at question bias, models of prediction, types of questions, etc.

Despite this need, however, Parts 1 and 2 of the BMAT provide reasonably reliable assessments although care needs to be taken with some questions which are less than satisfactory.

8 TSA and BMAT in the Future

8.1 Much Needs Doing

This report has provided a brief overview of the origins from which the TSA and the MVAT/BMAT were created and has shown how these tests have worked in practice. From what has been seen, it is quite clear that both tests provide a good basis for development but also that there is much that can and should be done to improve the assessment vehicles currently being piloted. Until such work has been done, the tests are not likely to reach their full potential.

The greatest need is for further data to be collected to evaluate the predictive power of the tests. This is currently happening but, crucially, what is needed are sets of data which include not only the test scores and the first-year results of students but also AS, A-Level and AEA grades and marks, Interview Scores and other evidence used during the selection process. With these data, many different features of the tests and their efficiency could be evaluated. A linear prediction model using the CT and PS sub-scores separately in the light of the possible differences in the skills assessed by the CT and PS questions would be another area of investigation that could usefully be explored.

It has been noted that the results of using interviews, Interview Scores, indicate that these have been very poor at predicting subsequent achievement at University. Indeed, Interview Scores are notable for being unreliable and subject to unexpected variation between interviewers. This is summarised by Salvatori (2001) who, in the context of selection of students for the health professions writes ‘Controversy remains as to the value of personal interviews and written submissions as selection tools, although it is clear that training of assessors and explicit rating guidelines enhance their reliability and validity.’ This point is underlined by Lucius (2003) in a conference presentation (where the context was not academic selection but Human Resource Management) where the following points were among those made.

> Interviews are very widely used but can be very subjective as many Interviewers lack formal training.
>
> Interview scores can suffer from halo effects, lack of interview structure and agreed rating scale for questions asked and the scores can thus be unreliable (i.e. the agreement between multiple Interviewers is often low).
>
> If this lack of training is overcome, interviews have the potential to yield scores that were more reliable and could thus provide useful information.

While the use of structured questions with an agreed response rating used by of multiple interviewers is clearly somewhat optimistic for use in an academic context, it is entirely possible that with careful training Interview Scores might become more useful in selecting students who have academic potential that is the case at present.

As part of the development of TSA and BMAT, it would be extremely useful to find a partner with whom the TSA could be used on a trial basis. The University of Cambridge is one that attracts applications from students who are of very high ability and, while the range of ability is large, the ability of most applicants falls at the extreme end of the scale. This may well be why the TSA works as well as it does in relation to other measures that do not discriminate well between applicants of high ability but it would be extremely useful to expand the trial to include an institution that does not have such a selected group of applicants. Not only would this be likely to increase the reliability of the tests used but the value of using such tests might also be seen more clearly. It would be important not to lose the focus on the need for selection between applicants with high A-Level grades but many of the aspects of measurement and bias could be tackled better with data from applicants with a greater range of ability than has been the case so far.

If a composite data sets were collated and, taking account of the problems on use raised in Section 5.7, background variables on students could be included that covered issues such as ethnic background, then it would be possible to use these data for analyses looking at the various sources of question bias as well as study the analyses of data by question type.

It is also clear that the nature of the TSA needs to be reviewed. The skills assessed and the types of questions used are those that were identified in the early 90s and, although they are still likely to be relevant today, they should be re-affirmed and the types of questions used re-considered. Just as the degree courses themselves will have changed over the past years, so too the skills assessed by the TSA may need to be amended to focus on what is currently important for success in university. This would enable the reliability of the TSA to be considered, and possibly increased, and might also improve its predictive validity.

Part of this review of skills could usefully include an investigation into the dimensionality of the existing test data already collected. The, somewhat insensitive, investigation here has indicated no major differences between CT and PS questions in terms of what they assess but there are suggestions that there may be elements of these skills that are different (indeed, it would be surprising if there were not such differences). Just how large these differences are, whether they can be pinpointed and whether or not any differences are important would be useful to know as it would allow the review to be conducted against a backdrop that indicated what was being assessed at present. A view could then be formed about whether any such skill divisions needed nurturing.

No mention has been made here of prediction of success in different subjects. This has not been realistic with the current data as sample sizes have been small but as more data are obtained so too can the value of using the TSA in different subjects be investigated.

Only when all of this information is available can the value of the tests as supplementary predictors be fully evaluated.

Following up the purpose to which the TSA and BMAT results are used (i.e. helping to decide which applicants should be offered a place at university), it would also be useful to follow up those students who were not successful with their application as well as those who were admitted. This is, however, a task that would be extremely complex and would require substantial resources even though the information gained would be invaluable in evaluating the worth of these tests.

There is little doubt that in the current educational climate the problems of selection are creating issues for all concerned with access to Higher Education. The Tomlinson Report (DfES, 2004) and the subsequent discussions clearly underline concerns about the use of the existing A-level system for university selection. As time goes on, candidates are achieving better and better grades and university admissions officers are at the sharp end in the sense of having more and more applicants with better and better grades. As has been seen, the selection rate at Cambridge is currently about 1 in 4 but in other universities it may be as high as 1 in 10 or more. In such cases, supplementary predictive information that can help the task of admissions is likely to be welcomed.

University admissions was also the focus of the work of the Working Group chaired by Stephen Schwartz (Admissions to Higher Education Steering Group, 2004), where the topic of how to deal with the changing nature of admissions is considered. In particular, consideration is given to the use of a common test in the UK and 'The Group notes that US-style SATs are one test worth exploring, alongside other possibilities.' and that it would welcome '… the evaluation of other tests …'. Both the TSA and the BMAT are clear candidates for such consideration but, as already explained in Section 5.3, there is considerable doubt as to the relevance of tests such as the SAT to the UK Educational system (Fisher, 2005).

8.2 Some Comments and Conclusions

In conclusion, a number of points may be highlighted for consideration of the way forward. These are drawn from the above paragraphs and also from Fisher (2002)

- It has to be said that both the TSA and the BMAT come from a long background of Thinking Skills development and are seen to be reasonably reliable assessments. They thus provide a good basis for providing Admissions Officers with supplementary information that could be used alongside that currently used as predictors of university achievement. The predictive validity of these assessments has yet to be seen fully but, from the limited evidence presented herein, the test scores appear to be no worse, and possibly even somewhat better, than what is used at present (e.g. A-Level grades , Interview Scores, etc.).

- The admissions process is a difficult task and Admissions Officers may welcome information provided by a test that assesses 'how candidates think' as it may help them carry out this task.

- The TSA and BMAT do assess something distinctive - cognitive skills which are not directly assessed elsewhere.

- By assessing both Critical Thinking and Problem Solving skills in the TSA, many of the skills necessary for Higher Education are assessed directly but the nature of these skills should be reviewed to ensure that they are up-to-date.

- Questions included in the TSA should look like the kinds of questions students encounter in university, thus giving the tests face validity.

The (Thinking) skills assessed by the TSA/BMAT are teachable (see Fisher, 2005) and teaching materials or guidance should be provided. This point is one that has not been discussed herein but if Thinking Skills are taught in schools then not only will future applicants be better prepared for Higher Education and to be successful in their application to University but they are also likely to do better in their subject-based A-Levels as well. Evidence for the effect of teaching Thinking Skills on subsequent examination performance is given by Fisher (2005). This being the case there remains an open question as to whether effort should be made to develop materials for teachers to support the teaching of Thinking Skills in schools

8.3 Suggestions for Further Work

Throughout this report, areas of investigation that need attention have been mentioned as they occurred and a summary of these is provided in Appendix B. Much of this work can be conducted in the context of establishing co-operation with other universities and testing with a wide range of students to broaden the base for evaluation. This would allow the collection of a substantial data set for the evaluation of the nature, quality and predictive ability of both the TSA and the BMAT tests.

Appendix A

TSA Test Specification

The TSA consists of Critical Thinking and Problem Solving questions. These questions are intended to assess different aspects of Thinking Skills although it is recognised that there will be something that is measured in common between them.

A TSA test consists of 50 multiple-choice questions. Of these questions, 25 assess Critical Thinking and 25 assess Problem Solving.

The questions are classified, as below, into seven types of Critical Thinking questions and three types of Problem Solving questions.

Critical Thinking Skill Category	**Skill Assessed**	**No. of Questions**
1	Summarising the main conclusion of an argument	5
2	Drawing a conclusion when premises are given	4
3	Identifying assumptions	4
4	Assessing the impact of additional evidence	4
5	Detecting reasoning errors	4
6	Matching one argument with a second which has the same logical structure	2
7	Applying principles	2
Total - Critical Thinking		**25**

Problem Solving Skill Category	**Skill Assessed**	**No. of Questions**
FP	Finding Procedures	9
RS	Relevant Selection	8
IS	Identifying Similarity	8
Total - Problem Solving		**25**
	TSA Test	**50**

Appendix B

Areas for Further Work

Throughout this report, references have been made to areas of investigation that would be of benefit to the TSA and BMAT assessments. Most of these have been included in the main report but this appendix provides a note of the areas that would most benefit from further work.

Collation of full data set for the evaluation of:

Comparative prediction of combinations of sub-scores (as well as whole test scores)
A comparison of predictive ability by subject of application
Relationship between classical predictors and TSA/BMAT to establish worth of all variables (i.e. A-Level grades, Interview Results, School Reports, etc.)

This would require data to be collected over a period of time and, if possible, for this to be done in conjunction with at least one other University with a wider range of applicants.

Work on development of question types and question and test data to:

Ensure questions assess the appropriate Thinking Skills for Higher Education today and see whether this has any implications for the question types used.
Interpret factor structures in relation to question types and skills assessed
Eliminate as far as possible any cultural assumptions that may be implicit in questions
Reduce the effect of language use/ability on questions (especially CT questions)
Detect, identify and eliminate question and test bias
Improve test questions through question analysis and evaluation and so improve test reliability
Determine functioning of test parts – CT/PS

Much of this work could be done independently from the main validation samples. Some data would need to be collected - to obtain a sample of students with a range of ethnic backgrounds, for example - but the existing data could still yield more results than have been obtained already.

Other Points

One matter that comes up in the discussion of the use of Thinking Skills is that of teaching. There is little doubt that Thinking Skills can be taught and that such teaching improves students' score not only on tests of Thinking Skills (such as TSA and BMAT Part 1) but also on traditional measures of achievement such as A-level (see Fisher, 2002, 2005). It is a point worthy of consideration that UCLES might lead with the production of teaching/learning materials in this field.

A review of 'where rejected applicants go' and what success they have in whatever paths they follow would be invaluable but it is recognised that this would be a very complex, and so costly, project to mount.

All of this work could not be mounted at one time. Some parts are quite urgent if the TSA and BMAT are to stay ahead of the field (e.g. questions development) but much could be included over a number of years, perhaps as few as three, so that the true nature of the tests used could be determined.

References

Note on Availability

In preparing this report, a wide variety of documents has been drawn on including many old project documents, many of which are no longer available. The references in the text are provided as usual and each one falls into one of four categories. In the list of references below the category into which each one falls is indicated by a bold, capital letter **A**, **B**, **C** or **D**.

Category A

Reference to an existing publication or document that may be obtained in the usual way (e.g. from libraries, bookshops, etc.).

Category B

Reference to a document held in electronic form, in either Adobe pdf or MSWord format, which may be downloaded from the Cambridge Assessment website under the Assessment Directorate's section listing publications, conference papers and other articles.
http://www.cambridgeassessment.org.uk/research/confproceedingsetc/

Category C

Reference to a document held in electronic form, in either Adobe pdf or MSWord format, and which may be downloaded from the Cambridge Assessment website under the Assessment Directorate's section listing older (archive) documents of interest.
http://www.cambridgeassessment.org.uk/research/historicaldocuments/

Category D

Reference to an old UCLES project document that is no longer available.

References

Admissions to Higher Education Steering Group (2004). Fair Admissions to Higher Education: Recommendations for Good Practice. (Schwartz Report). **A**

Anastasi, A. (1968). Psychological Testing. New York: Macmillan. **A**

Bramley, T. (2001). MVAT 2000 – Statistical Report. Report prepared for the University of Cambridge Local Examinations Syndicate. **B**

Cambridge Reporter (2004). Undergraduates: Statistics of Applications and Acceptances for October 2003. University of Cambridge Reporter Vol. CXXXIV; Special Report No.12, Friday 12 March 2004. **A**

Chapman, J (2005). The Development of Thinking Skills Assessment. Report prepared for the University of Cambridge Local Examinations Syndicate. **B**

CIE (2004). Syllabus for A-Level and AS Thinking Skills. Cambridge: CIE. **A**

DfES (2004). 14-19 Curriculum and Qualifications Reform: Final Report of the Working Group on 14-19 Reform. (Tomlinson Report) London: DfES. **A**

Ebel, R. L. (1972). Essentials of Educational Measurement. New Jersey, Englewood Cliffs: Prentice-Hall. **A**

Fisher, A (1989). The Higher Studies Test. Report prepared for the University of Cambridge Local Examinations Syndicate. **B**

Fisher, A. (1990a). Proposal to Develop a Higher Studies Test: A Discussion Document. Cambridge: UCLES. **B**

Fisher, A. (1990b). Research into a Higher Studies Test: A summary. Cambridge: UCLES. **B**

Fisher, A. (1992). Development of the Syndicate's Higher Education Aptitude Tests. Report prepared for the University of Cambridge Local Examinations Syndicate. **B**

Fisher, A. (2002). Thinking Skills for the 21st Century. Paper prepared for CIE seminar: Thinking Skills for the 21st Century. Cambridge 18 November 2002. Cambridge: CIE. **A**

Fisher, A. (2005). 'Thinking Skills' and Admission to Higher Education. Report prepared for the University of Cambridge Local Examinations Syndicate. Cambridge: UCLES. **B**

Forster, M. (2004) Final Analysis of the 2003 TSA Tests. Report for the University of Cambridge Local Examinations Syndicate. **B**

Green, A. (1992). MENO: A Validation Study of Formal Reasoning Items. Report prepared for the University of Cambridge Local Examinations Syndicate. **D**

Guilford, J.P. (1973). Fundamental Statistics in Psychology and Education. New York: Appleton. **A**

Lucius, R. H. (2003). Selection Assessments: What you should know. Presentation to SHRM Atlanta Conference, 21 October 2003. **A**

Massey, A. J. (1994a). MENO Thinking Skills Service 1993/94 Pilot: Evaluation Report. Report from the Evaluation Service: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **B**

Massey, A. J. (1994b). Evaluation of MENO 1992/93 Pilot Data: Experience in Two Universities. Report from the Evaluation Service: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **B**

Massey, A. J. (1997). MENO and GCE Achievement in Singapore. Report from the Evaluation Service: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **D**

Massey, A. J. (2004). Medical and Veterinary Admissions Test Validation Study. Report from Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **B**

Massey, A. J., Shannon M. and Dexter, T. (2002). MVAT Scores and Outcomes of Applications for Medical and veterinary Courses in 2001. Report from the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **B**

Nonneman, W., De Metsenaere, M., Janssen, P., Cortens, I., Tanssens, R., Wels, G. and Claes, L. (1995). The 'MENO' Test at Flemish Universities. Report prepared for the University of Cambridge Local Examinations Syndicate. **D**

Nuttall, D. L. and Willmott, A. S. (1972). British Examinations: Techniques of Analysis. Slough: NFER. **A**

OCR (2003). Advanced Extension Award in Critical Thinking. Cambridge: OCR. **A**

OCR (2004). AS/A-Level GCE in Critical Thinking: Specification Summary. Cambridge: OCR. **A**

Robinson, P. (2002). Thinking Skills and University Admissions. Presentation made to CIE seminar: Thinking Skills for the 21st Century. Cambridge 18 November 2002.Cambridge: CIE. **A**

Rule, E. (1989). The Law Studies Test: Results of a Trial Administration. Report prepared for: University of Cambridge Local Examinations Syndicate. **D**

Salvatore, P. (2001). Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. Advances in Health Sciences Education **6** (2): pages 159-175. **A**

Shannon, M. (2004a). BMAT Scores and Outcomes of Applications to the University of Cambridge for Medical and Veterinary Courses in 2003. Report from the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **B**

Shannon, M. (2004b). BMAT Section 1 Factor Analysis. Results of analyses for UCLES. **B**

Shannon, M., Massey, A. J. and Dexter, T. (2003). MVAT Scores and Outcomes of Applications for Medical and veterinary Courses in 2002. Report from the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **B**

Thomson, A. and Fisher, A. (1993). A validation Study of Assessing Argument Items. Report prepared for the University of Cambridge Local Examinations Syndicate. **D**

UCLES (undated). Draft Test Material for a Higher Studies Test: A Consultation Document. Document produced by the University of Cambridge Local Examinations Syndicate. **D**

UCLES (1992a). Academic Aptitude Profile: Proposed Scheme. Consultation paper. **B**

UCLES (1992b). Academic Aptitude Profile: Rationale. Consultation paper. **B**

UCLES (1992c). Academic Aptitude Profile: Logical Reasoning: sample Questions. Consultation paper. **D**

UCLES (1992d). Academic Aptitude Profile: Guide to Writing and Editing Mathematical Reasoning Items. Consultation paper. **D**

UCLES (1993a). MN01-1: The MENO Thinking Skills Service 1993-1994: An Introduction. **C**

UCLES (1993b). The MENO Thinking Skills Service 1993-1994: Students' Guide. **C**

UCLES (1993c). The MENO Thinking Skills Service 1993-1994: Handbook of Administrative Arrangements. **D**

UCLES (1993d). The MENO Thinking Skills Service 1993-1994: Development and Rationale. **C**

UCLES (2004a). Thinking Skills Assessment. A student Handbook for TSA. Cambridge: UCLES. (See http://tsa.ucles.org.uk/index.html). **A**

UCLES (2004b). Analysis of (Subject B) Candidates' TSA Results 2003. Report from Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. **B**

US Department of Labor (1999). Testing and Assessment: *An Employer's guide to good practices.* A document by the U.S. Department of Labor, Employment and Training Administration (full document is available at http://www.hr-guide.com/data/G358.htm). **A**

Willmott, A.S. (2004). TSA Administration December 2003: A Preliminary Investigation of Question Bias. Report prepared for the University of Cambridge Local Examinations Syndicate. **B**