validity is the best basis for evaluating a selection test. The objective of the selection test is to select the students who will perform best on the outcome measures. This leads to the conclusion that it might be better to evaluate the predictive validity of a selection procedure in terms of the improvement in the quality of those selected. This could be based on a change in mean score or proportion of satisfactory students. The case of a binary outcome is discussed in more detail in Bell (2005b, c).

This article shows that it is possible that by using simplistic analyses the benefits of using selection tests may have been underestimated. For example, in the late 1960s there was an experiment using a SAT-style test in the United Kingdom (Choppin *et al.*, 1972; Choppin and Orr, 1976). The results of the experiments were considered to be something of a disappointment despite the fact that the test had been carefully designed. There was a considerable degree of selection, for example, only 26% of those who sat the test were admitted to universities. The authors of the reports used simple correlations and regression to analyse the data. It is interesting to note the patterns of results for individual institutions for mathematics. The institution with the highest mathematics scores (presumably an institution not affected by self-selection) and so a very high degree of selection, had a correlation of 0.36 for both the mathematics and verbal scores. However, the correlations were much lower and in some cases slightly negative for an institution which would have been selective and been affected by self-selection. From the simulation it is clear these results are consistent with an effective selection test, although it is also true this need not be the case. The problem is that the analyses are based on simple correlations. This is not a criticism of the authors of both reports. Both theory and the technology have advanced a long way from the 1970s. However, it is reasonable to conclude that there is a possibility that the conclusions about the ineffectiveness of this test were erroneous.

In conclusion, when a researcher makes a sweeping claim about the ineffectiveness of an admissions test but bases their argument on an uncorrected correlation or a simple regression analysis and does not consider the effects of selection, then there is a distinct possibility that such a claim is mistaken. Higher education admissions are important and it is vital that care is taken with them. Thus it is vital that research into admissions tests address in full the complexities of the data that arise from their use.

Acknowledgements

The translations from Dutch were made by a colleague, Mark Claessen.

References

Bell, J.F. (2005a). Gold standards and silver bullets: Assessing high attainment. *Research Matters: A Cambridge Assessment Publication*, **1**, 16–19.

Bell, J.F. (2005b). Evaluating the predictive validity of a selection test. Part 1 – Replacing an existing procedure. *Submitted for publication*.

Bell, J.F. (2005c). Evaluating the predictive validity of a selection test. Part 2 – Supplementing an existing procedure. *Submitted for publication*.

Bell, J.F. (2006). The effect of the selection method on the evaluation of the predictive validity of a selection test. *In preparation*.

Choppin, B.H.L., Orr, L., Kurle, S.D.M., Fara, P. & James, G. (1973). *Prediction of academic success*. Slough: NFER Publishing.

Choppin, B. & Orr, L. (1976). *Aptitude testing at eighteen-plus*. Slough: NFER Publishing.

Gustafsson, J.-E. & Reuterberg, S.-E. (2000). Metodproblem vid studier av Högskole-provets prognosförmåga – och deras lösning. [Methodological problems in studies of the prognostic validity of the Swedish Scholastic Aptitude Test (SweSAT) – and their solution] *Pedagogisk Forskning i Sverige*, **5**, 4, 273–284. (In Swedish with extensive English summary)

Linn, R.L. & Dunbar, S.B. (1982). Predictive validity of admissions measures: correction for selection on several variables. *Journal of College Student Personnel*, **23**, 222–226.

Muthén, B., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, **42**, 431–462.

Sackett, P.R. & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, **85**, 112–118.

ten Cate, T.J.T. & Hendrix, H.L. (2001). De eerste ervaringen me slectie [Initial experience with selection procedures for admission to medical school]. *Nederlands tijdschrift voor geneeskunde*, 14 Juli:145, 28, 1364–1368.

Wood, R. (1991). *Assessment and Testing: A survey of research*. Cambridge: Cambridge University Press.

Yang, H., Sackett, P.R. & Nho, Y. (2004). Developing a procedure to correct for range restriction that involves both institutional selection and applicants' rejection of job offers. *Organisational Research Methods*, **7**, 4, 442–455.

PREDICTIVE VALIDITY

Using Thinking Skills Assessment in University admissions

Joanne Emery and John F. Bell Research Division

In the first issue of *Research Matters*, the difficulties involved in assessing high attaining candidates were discussed (Bell, 2005a). A particular problem is that elite institutions are faced with selecting among candidates with the same grades on existing qualifications. Most applicants to the University of Cambridge are predicted, or have already, at least three grade As at A-Level. Cambridge University admissions staff therefore requested that Cambridge Assessment (then known as UCLES) develop a 'Thinking Skills Assessment' (TSA) to assist in making admissions' decisions. When first proposed, the TSA was seen as a test that would form part of the admissions interview process so that it could be taken by applicants during their interview visits to Cambridge. This has the advantage in the Cambridge context of allowing the use of the test on a college-by-college and a subject-by-subject basis. At the time of writing, most Cambridge colleges use the TSA during the admissions process and the range of subjects for which it is used varies from college to college. The test provides *supplementary information* for use in helping to make admissions decisions. Obviously, to be meaningful, any such selection tool must be able to predict future performance. This issue of predictive validity is the focus of this article.

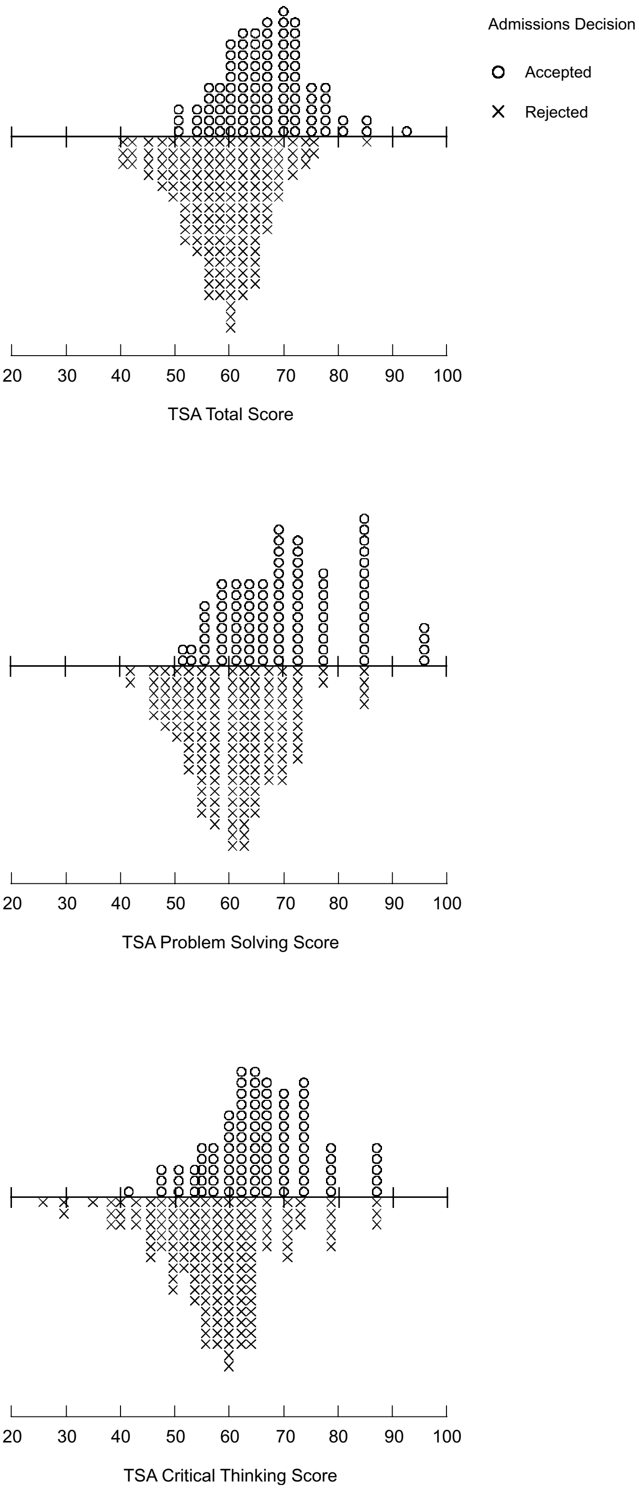
The Cambridge Thinking Skills Assessment (TSA) provides an assessment of two kinds of thinking: Problem Solving and Critical Thinking. Problem Solving describes reasoning using numerical, graphical and spatial skills. It requires developing strategies for tasks through thought and planning. Critical Thinking is often defined as 'reasonable, reflective thinking that is focussed on deciding what to believe or do' (Ennis, 1996). Central to Critical Thinking are the concepts of argument and evaluation. It requires the ability to interpret, summarise, analyse and evaluate arguments and ideas. With the TSA, the aim is to provide an assessment of *Thinking Skills*: intellectual skills that are independent of subject content and are generalisable to a wide range of subject areas. For example, the skill of Critical Thinking can be useful in subject areas ranging from the Humanities (interpreting documents and evaluating their arguments) and the Arts (following the reasoning of great thinkers) right through to the Sciences (appreciating advances in scientific development).

Cambridge Assessment has a long history of developing tests under the general heading of 'Thinking Skills'. An item bank of former Thinking Skills questions (items) was built up for this purpose. This gave an excellent starting point for the development of the TSA. The test consists of 50 multiple-choice questions, each with 5 possible answers, and has a time limit of 90 minutes. Questions assessing Problem Solving and Critical Thinking skills are mixed throughout the test and there are no penalties for incorrect responses. In December 2001, 289 Computer Science applicants took the TSA. This expanded to 472 in December 2002 with more colleges and more subjects taking part.

Up to this point the objective of the TSA work was the development and evaluation of the test itself but in January 2003 Cambridge Assessment added a second objective: that of experimental online delivery of the test. This software was developed specifically for Cambridge Assessment as a prototyping system. Both objectives were successfully achieved: there was a greatly enhanced take-up of the test, with 23 colleges taking part involving 4 main subjects (Computer Science, Engineering, Natural Sciences and Economics), and the administration procedures were based around the online system we had developed. A total of 1,551 tests were administered in that year: 1,114 paper tests and 437 online tests. An especially valuable feature was the administrative website used for making entries (registration) and returning results. Online tests were marked automatically and paper tests were marked using scanning technology with intelligent character recognition. A website (http:// tsa.ucles.org.uk) is available giving details of the TSA with example and practice materials.

This article reports on the 2003 TSA scores and the subsequent 1st year (Part 1A) examination results of Computer Science students (taken in Summer 2005). Of the 1551 candidates who sat the TSA in 2003, 238 applied to study Computer Science. Of these, 94 received an offer of a place and 144 were rejected. TSA scores are reported as a total calibrated score and as separate calibrated Problem Solving and Critical Thinking sub-scores. The calibration process allows the results of candidates taking different versions of the test to be reported on a common interval scale.

Figure 1 : Dot density plots showing the TSA 2003 score distributions of candidates who were offered a place (conditional or unconditional) and candidates who were rejected for the Computer Science course



Plots showing the TSA score distributions of Computer Science candidates who were offered a place (conditional or unconditional) and candidates who were rejected are shown in Figure 1. These plots are particularly helpful in evaluating whether the TSA is likely to be useful as a preliminary hurdle to reduce the number of interviews given (at the time of writing it is rare for an applicant not to be interviewed). It can be seen that few of the accepted candidates had low scores. If the test were to be used for pre-selection, two questions need to be considered: why any relatively low-scoring candidates were accepted and whether they could be spotted without an interview.

Correlations, whilst problematic, are the most familiar measure of predictive validity. Table 1 displays the correlation coefficients between the 2003 TSA scores and 1st year Computer Science examination performance in 2005. Pearson coefficients are given for all variables except 'rank' where a Spearman's rho is used due to the ordinal nature of the data. It should be noted that the coefficients displayed are uncorrected for the effects of selection. Correlations tend to produce underestimates where selection tests are concerned due to restricted score ranges. Although there are corrective formulae, none of them apply to this particular situation where the selection test is used in conjunction with other qualitative information. There are, however, some guidelines that can be applied.

General guidelines for interpreting validity coefficients

Validity coefficient value	*Interpretation*
above .35	very beneficial
.21–.35	likely to be useful
.11–.20	depends on circumstances
below .11	unlikely to be useful

(US Department of Labor, Employment Training and Administration, 1999)

Table 1 : Correlations between TSA 2003 scores and Part 1A examination outcome in Computer Science

	N	*TSA Total Score*	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Computer Science Part 1A Rank in Year	64	–.453**	–.439**	–.292**
Computer Science Part 1A Total Mark	64	.445**	.419**	.315**
Computer Science Paper 1 Mark	67	.488**	.477**	.327**
Computer Science Paper 2 Mark	64	.566**	.505**	.425**

** The correlation is significant at the 0.01 level (one-tailed)

The total TSA score and both the Problem Solving and the Critical Thinking components show highly significant positive correlations with 1st year examination performance in Computer Science. The relationships are slightly stronger for the Problem Solving component than the Critical Thinking component but show the greatest magnitude for the combined total score. Paper 1 of the examination covers topics on the Foundations of Computer Science, Operating Systems, Algorithms and Java Programming. Paper 2 is composed of questions on Digital Electronics and Discrete Mathematics.

Table 2 displays the means and standard deviations of the TSA scores of candidates achieving various Part 1A examination classes in 2005. Total examination marks are graded (in descending order of merit) as class 1, class 2:1, class 2:2, class 3, ordinary and fail. Students obtaining 1st class results tend, on average, to have gained higher total scores on the TSA than those who went on to obtain a 2:1, who, in turn, tend to have achieved higher scores than those obtaining a 2:2. This is also the case for the Problem Solving and Critical Thinking sub-scores. It is notable that the average TSA scores of students gaining a 3rd class outcome are higher (for total score and Critical Thinking) than those of candidates gaining a 2:2. It is quite likely that candidates who obtain such poor results do so for reasons that are not necessarily related to their academic abilities, given that they have very high academic achievement prior to their arrival at Cambridge. An admissions test can only identify those students who are capable of doing well: not necessarily those who will do well.

Table 2 : TSA descriptive statistics by examination class achieved in Computer Science

Part 1A Class	*N*	*TSA Total Score*	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Class 1	16	71.5 (7.99)	75.2 (12.09)	70.2 (8.70)
Class 2:1	18	68.5 (6.96)	74.6 (11.46)	65.8 (9.43)
Class 2:2	23	63.4 (7.36)	66.2 (8.60)	61.4 (8.32)
Class 3	7	63.9 (6.14)	62.7 (7.18)	65.7 (6.63)

The weakness of correlation analysis here is that it cannot include data for candidates who have been rejected. The TSA is used in a complex process which is compensatory in nature but not necessarily quantified. This means that there is no simple way of adjusting the coefficients for selection effects. However, there is an alternative method of evaluating predictive validity. When a selection procedure is based on the principle of maximising academic performance then this is the same as assuming that, for a given TSA score, the probability of obtaining a degree of a particular class is lower for applicants who were rejected compared with those who are accepted. There is no way of directly testing this. However, tau analysis has been developed to investigate this assumption (Bell, 2005b, 2005c).

The tau method uses logistic regression equations to calculate the probability that any given TSA score will result in the student who achieved it gaining a 1st class result. The students who were actually selected for course entry using the existing methods (predicted grades, interview performance, UCAS form information) are compared to the students who *would* have been selected if TSA scores alone had been used, in terms of how many 1st class outcomes they achieved (or *would* have achieved). The probable number of 1sts which would have been achieved with the TSA-only method is calculated by taking the top *n* highest-scoring TSA candidates (the same number as were actually selected) and simply summing together their calculated probabilities of success.

The above analysis requires assumptions to be made about the probabilities of success for the rejected applicants. The magnitude of these probabilities is related to the degree of confidence in the existing selection system. It is assumed that, for any given mark on the TSA, any candidate who was rejected by the existing system had a lower probability of success than one who has been accepted. The tau method quantifies this difference. This is achieved by multiplying the probabilities for the rejected candidates by a confidence factor (1 minus 'k'). The possible values of k can range from 0 (no confidence) to 1 (absolute confidence). Absolute confidence implies that the procedure has definitely selected the best candidates and no confidence suggests the selection procedure was essentially random (this is not plausible if the logistic regression for selected applicants is positive). In practice, the confidence level is not known. However, it is possible to investigate the predictive validity of the test by considering a range of confidence values.

Table 3 : Tau analyses comparing the probable success rates achieved using new (TSA-only) versus existing selection methods

COMPUTER SCIENCE PART 1A 2005 *confidence in existing system: k=0.75*

	New Predictor Measure		
	TSA Total Score	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Number of applicants	210	210	210
Number accepted	67	67	67
Number selected with new method	67	70	67
Actual number of firsts achieved	16	16	16
Predicted firsts for new method	16.2	16.3	16.4
Observed success rate	0.24	0.24	0.24
Predicted success rate of new method	0.24	0.23	0.24
new/existing	1.0	1.0	1.0
new/random	2.5	2.3	2.4
existing/random	2.4	2.3	2.3

confidence in existing system: k=0.5

	New Predictor Measure		
	TSA Total Score	*TSA Problem Solving Score*	*TSA Critical Thinking Score*
Number of applicants	210	210	210
Number accepted	67	67	67
Number selected with new method	68	67	67
Actual number of firsts achieved	16	16	16
Predicted firsts for new method	17.2	16.4	17.7
Observed success rate	0.24	0.24	0.24
Predicted success rate of new method	0.25	0.24	0.26
new/existing	1.1	1.0	1.1
new/random	2.1	1.9	2.1
existing/random	2.0	1.8	1.9

(Note: any differences between the numbers accepted and numbers selected with the new method are due to tied ranks in TSA scores)

The tau tables presented above show the case where k=0.75 (thus assuming high confidence in the existing system) and the case where k=0.5. The 'accepted' candidates are those who were actually selected by the colleges and for whom the number of 1st class results is known. The 'selected' group are those who would have been chosen on the basis of the TSA alone (the total score or its subscales). The 'random' group is akin to using a lottery method: its success rate considers the probable number of 1st class results for the entire applicant pool. The success rates and ratios presented above compare the proportion or likely proportion of students gaining a 1st class outcome using the old, new and random selection methods.

The results suggest that, even when confidence in the existing system is high, using the total TSA score alone would have resulted in at least the same success rate as was obtained using the existing selection methods. If confidence is lower (k=0.5) then the total TSA score and the Critical Thinking component both outperform the existing selection system. However, it is not necessarily the case that the same candidates would have been accepted. The comparison of existing versus new methods shows the effect of changing from using all the information, including the TSA, to using the TSA alone. The success rates for both methods, however, are vastly superior to a random selection of candidates from the applicant pool.

In this article we have demonstrated that a Thinking Skills Assessment is useful in the University admissions process as an additional source of evidence. Correlations with subsequent examination performance are impressive, given the problems of restricted score ranges in such highly selected candidates. Students attaining higher examination classes tended to have achieved higher TSA scores and the tau analyses suggest that selecting on the basis of the TSA alone would have produced at least the same number of Class 1 results. In conclusion, there are substantive differences in Thinking Skills between candidates with three grade As at A-Level and these differences predict their future performance. Thus a selection process involving the assessment of Thinking Skills is necessary.

References

Bell, J. F. (2005a). Gold standards and silver bullets: Assessing high attainment. *Research Matters: A Cambridge Assessment Publication*, **1**, 16–19.

Bell, J. F. (2005b). Evaluating the predictive validity of a selection test. Part 1 – Replacing an existing procedure. *Submitted for publication*.

Bell, J. F. (2005c). Evaluating the predictive validity of a selection test. Part 2 – Supplementing an existing procedure. *Submitted for publication*.

Ennis, R.H. (1996). *Critical Thinking*. New York: Prentice-Hall.

US Department of Labor, Employment Training and Administration (1999). Employer's guide to good practices. (accessed at: http://www.chrysaliscorporation.com/pdf/Testing_and_Assessment_Guide.pdf)