Difficulties in evaluating the predictive validity of selection tests

John F. Bell Research Division

When assessments are used for selection purposes there is a need to establish their predictive validity. Although there is literature on the predictive power of school examinations, much of it fails to appreciate the complexity of the issues involved leading Wood (1991) to comment that 'the question has proved a seductive topic for statistical dilettantes'. More recently, there has been a growth in the use of tests to assist in the admissions process of universities. There are two major reasons for this growth: the need to ensure fair access and the current inability of A-levels to distinguish between high attaining candidates (Bell, 2005a).

The most selective higher education institutions have been finding that the existing school examination system is no longer providing evidence of differences in individual merit for the highest attaining candidates. An important question that is asked of selection tests is 'do they predict future performance?' Textbooks on educational measurement usually recommend assessing this 'predictive validity' by calculating the correlation coefficient between scores on the selection test and scores on an outcome variable such as degree classification, or the score on a test at the end of the first year of the degree course.

One of the most important problems associated with evaluating the predictive validity of a selection test is that the outcome variable is only known for the selected applicants. Ideally, to evaluate predictive validity a random sample of applicants would be used. There are obvious difficulties in practice (a selective university is never likely to replace an existing selection procedure with a lottery). It is almost always going to be the case that there will be rejected candidates who will not have an outcome score.

To illustrate the effect of selection, a simulated data set of one thousand applicants was created (fuller details of this data set and the analyses described here can be found in Bell 2006, *in preparation*). It was assumed that the outcome, for example an examination mark, was related to an underlying trait and that the two selection methods are also related to the trait, that is, both tests correlate positively with the activity measure and with each other. One test will be referred to as the selection test (which is being evaluated) and the other as the original method (e.g. examination grades or interviews scores).

Table 1 : Correlations between selection methods and outcome

	Selection Test	*Original Method*	*Outcome*
Selection test	1.00		
Original method	0.28	1.00	
Outcome	0.56	0.54	1.00

The correlations in Table 1 have been set at what can be considered to be a realistic level. There are many factors that can determine outcomes in the real world that are not measured by any one test (indeed some influences can be the results of events that occur after the applicant has been admitted). The low correlation between the two selection methods indicates that they measure different traits and that both are important predictors.

There are a number of different types of selection procedure. The first type is a simple lottery, referred to as RANDOM. When lotteries have been used for selection they have either been used with other methods, either in the form of weighted lotteries, for example Dutch medical school admissions (ten Cate and Hendrix, 2001), or one stage in a medical admission (lotteries are used at one UK medical school to reduce the number of applicants to a manageable size).

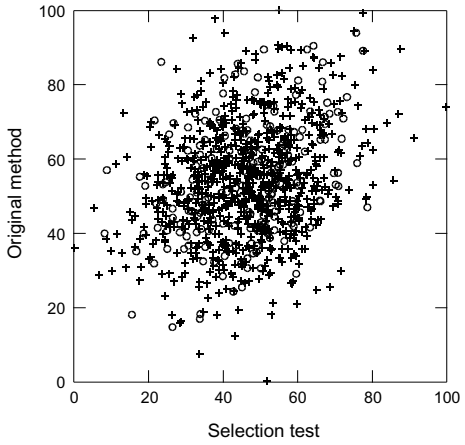
The next type uses only the original method. This involves taking the *n* highest scoring applicants on the original method where *n* is the number of available places (taking the best *n* applicants is assumed for all the remaining rules). This is the situation when a selection test is being piloted so it is referred to as a PILOT because it corresponds to a pilot year where the results of the selection test play no part in admissions decisions.

The next method will be referred to as EVAL and involves only using the selection test and ignoring the original method. This would represent the situation when a test that is the sole method of selection is being evaluated. Both PILOT and EVAL are examples of single hurdle rules.

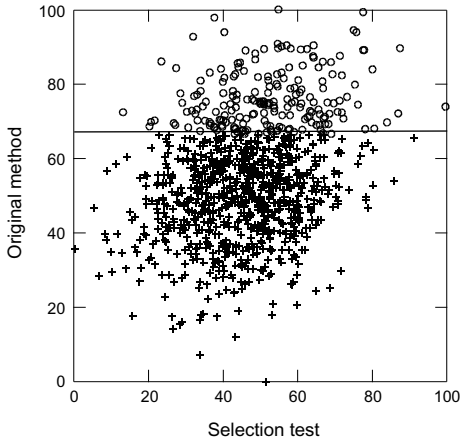
The remaining methods involve combining test scores. The first uses multiple hurdles and will be referred to as HURDLES. This involves selecting a fixed proportion of the entry with one test (e.g. the top 40% on the selection test) and then repeating this with another test (taking the 50% with the highest scores on the original method from the top 40% on the selection test). Multiple hurdles can be used when using all the selection methods on all applicants is prohibitively expensive so the first test is used to reduce the number of applicants for the second assessment. In this case, there are obviously multiple rules that could be applied depending on the percentages used for the first hurdle.

The next method of combining scores is compensatory and will be referred to as COMPEN. This involves taking a weighted sum of the scores. In this article, equal weights have been used but obviously others could be used. The effect of changing the weights is to change the slope of the line in panel (e) of Figure 1. In a compensatory method a very poor performance on one test can be compensated by a very good performance on another. This differs from the multiple hurdles method which guarantees a level of performance on all tests.

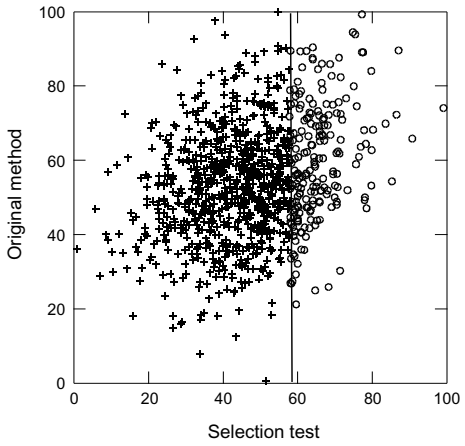
Finally, there are hybrid methods which use both hurdles and compensation and will be referred to as HYBRID (Figure 1(f)). These are probably the most realistic in practice (e.g. a University admissions decision might depend on obtaining at least a grade B for a particularly relevant subject – a hurdle – and exceeding a particular UCAS score – a compensatory method). In the example used in this article, a hurdle is set taking the top 40% using the selection test and then the top 20% using the compensatory rule described above.



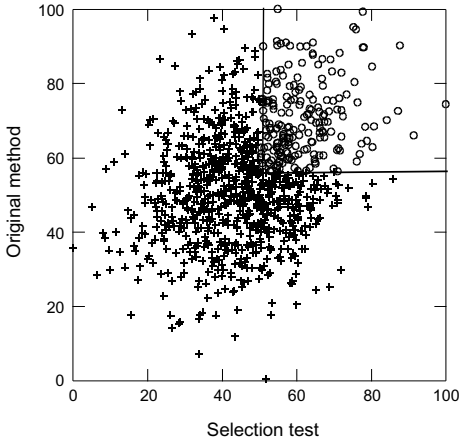
(a) Selected at random (RANDOM)



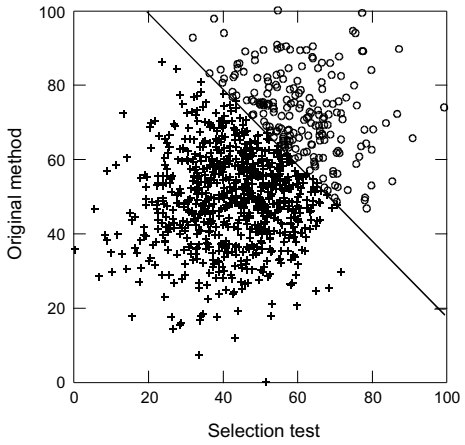
(b) Selected using the original method (PILOT)



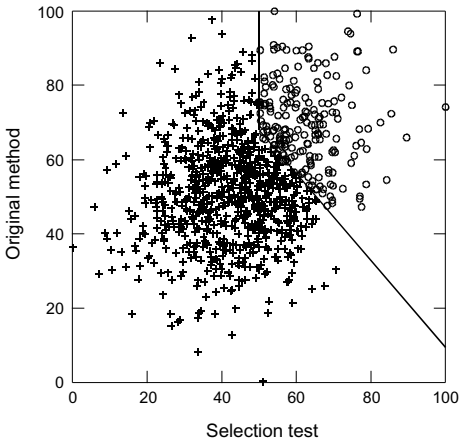
(c) Selected using the selection test (EVAL)



(d) Selected using multiple HURDLES



(e) Selected using COMPENsation method

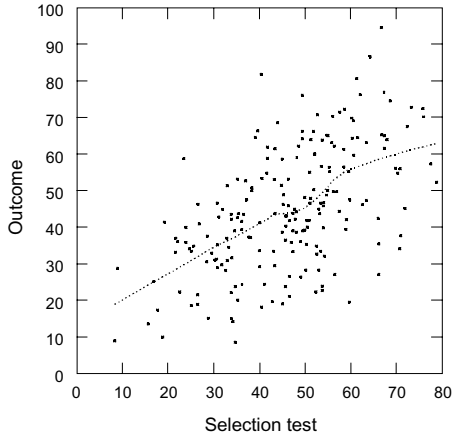


(f) Selected using HYBRID method

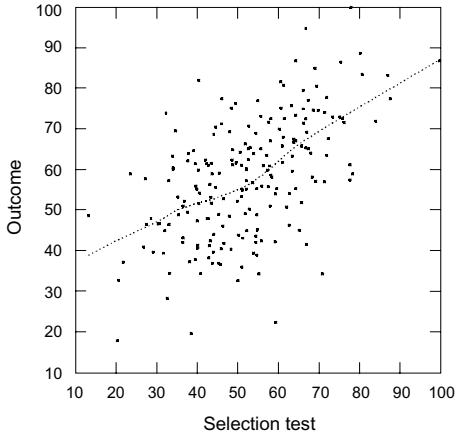
Figure 1 : Types of selection method

In addition, two other rules (RANORIG and RANCOMP) were defined for comparative purposes (these are not illustrated). In this case, it is assumed that it is only possible to obtain scores on the original method for 40% of the applicants. Rather than selecting the 40% with the selection test, this selection is used at random. These rules have been defined so that the outcomes can be compared with the multiple hurdle and the hybrid rules. The first rule is a random selection followed by the original method (the graph would be like Figure 1(b) but with fewer points and a line defined by a lower pass score because there are fewer candidates to select from) and the second is a random selection followed by the compensation method (like Figure 1(e) but with fewer points and the line closer to the origin). This is sometimes proposed as a solution when there are too many applicants to interview.

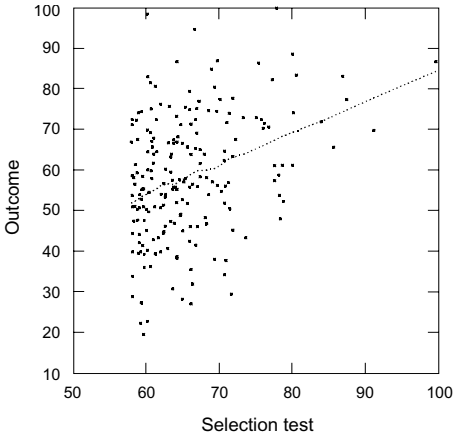
Note that the last five methods are examples from families of rules defined by the choice of weights and cut scores. This means that in the following discussion conclusions about the differences between these methods should be treated with care because they might not be using the optimal version of each rule.



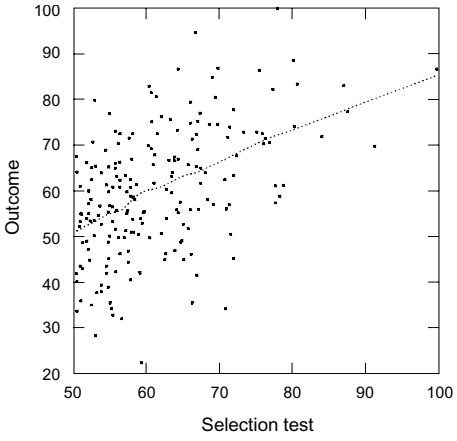
(a) RANDOM



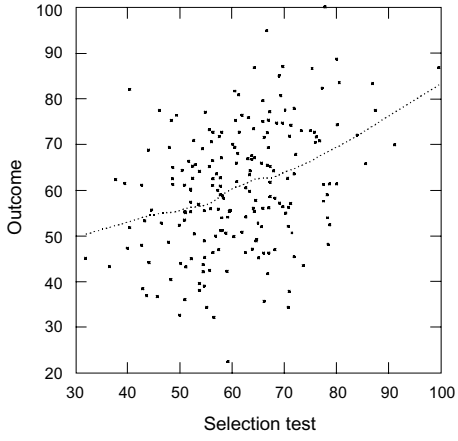
(b) PILOT `



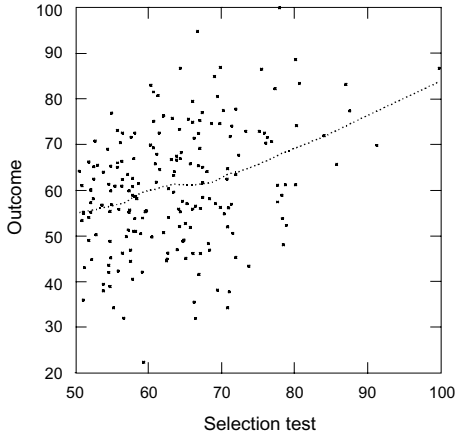
(c) EVAL



(d) HURDLES



(e) COMPEN



(f) HYBRID

Figure 2 :
Results of the selection

In the real world, only outcome data for the selected applicants is available. Scatter plots for the selected candidates are presented as Figure 2. Each part figure consists of a scatter plot of outcome against selection test for the selected applicants with a lowess smoothed line added. An inspection of the figures suggests that there is considerable variation in the strength of the relationship depending on the selection method used (this is most noticeable in the increasing spread of points even allowing for the changes in the axes).

Table 1 : Comparison of different selection methods

	Statistics			*Correlations*			*Grades*				
Method	*N*	*Mean*	*Sd*	*X1*	*X2*	*Xmean*	*A*	*B*	*C*	*D*	*E*
All	1000	45	17	0.56	0.54	0.69	20	40	60	80	100
RANDOM	200	45	17	0.55	0.64	0.72	23	37	59	80	100
PILOT	200	58	15	0.55	0.27	0.57	48	70	88	99	100
EVAL	200	58	15	0.35	0.52	0.55	45	73	87	96	100
COMPEN	200	60	14	0.49	0.31	0.51	50	79	92	99	100
HURDLES	200	60	14	0.49	0.31	0.51	50	79	92	99	100
HYBRID	199	60	13	0.35	0.28	0.47	52	80	94	99	100
RANORIG	199	53	15	0.55	0.39	0.61	34	59	80	95	100
RANCOM	200	54	14	0.32	0.33	0.48	36	62	80	86	100

(Note some rules involved ties so fewer than 200 were accepted)

In Table 1 some summary statistics about the different selection methods have been presented: the number selected, the mean and standard deviation of scores on the outcome variable for the selected applicants, correlations of the outcome with the selection test (X1), the original method (X2) and the mean of X1 and X2 (Xmean) respectively, and finally, the remaining five columns show the cumulative grade distribution for the selected applicants by dividing all the candidates into five equally sized groups based on the outcome scores. Thus, the mean score for the whole entry of 1,000 applicants is 45 with standard deviation of 17 and the correlations with three selection measures are 0.56, 0.54 and 0.69. By definition, 200 applicants in the whole entry obtained a grade A so a perfect selection method would give 100% A grade applicants. Inspecting the table reveals that the three methods that combine scores are the most successful at selecting good candidates (note it would be wrong to draw a general conclusion because neither cut-scores nor weights have been optimised). It is important to note that the predictive validity as measured by the uncorrected correlation coefficient declines as the selection methods become more effective.

Clearly, considering the correlation without considering the selection process can be very misleading. Suppose that the administration of an institution using the hybrid method squared the correlation and then concluded that the selection test only accounted for a relatively small 12% of the variation in the outcome and so abolished the selection test. If there were no change in the entry so the selection for the next year would generate results similar to the ones generated by the PILOT method, the percentage of grade A and B students would fall from 80% to 70%. This example suggests that the effectiveness of a selection procedure is better evaluated by considering the change in performance on the outcome variable rather than the correlation between scores on the outcome variable and the selection test.

One alternative to the simple correlation is to use a corrected correlation. However, the corrections vary with selection method and the availability of data. Sackett and Yang (2000) produce a very useful review of these methods. The correction not only depends on the selection method but also on the availability of the data on the original selection methods. In all cases, assumptions are made about the performance of the rejected applicants, the shapes of the relationships and the distribution of the errors.

More recently, research has been based on the fact that a selection method can be thought of as a missing data mechanism. With selection tests data are *Missing Not At Random*, abbreviated MNAR, and the missingness mechanism is termed non-ignorable. This has been applied to research into compensatory rules. In Sweden there is a complicated higher education admittance to higher education. This compensatory system involves applicants either being admitted on the basis of an admissions test or their school leaving certificate. Gustafsson and Reuterberg (1998) investigated modelling incomplete data (Muthén, Kaplan and Hollis, 1987) and found it to be a very efficient method for estimating the predictive validity of selection tests.

So far the assumption has been that if an applicant is accepted then they will take up the place at the institution. For most institutions this is not the case, since the most able applicants, although offered places, often choose to go to another institution. This is sometimes referred to as self-selection. It can have serious consequences when evaluating selection procedures. Consider two institutions P and Q. It is assumed that institution P is trying to select the best 20% and every one offered a place will take the place. Thus the second institution (Q) is only able to select from the remaining 80% of the sample. For the purposes of discussion, results for four decision rules have been generated:

SELF1: The top 20% and the next 20% are selected by the original method (as in Figure 3(a)).

SELF2: The top 20% and the next 40% are selected by the compensatory method (as in Figure 3(b)).

RANSELF1: A random selection is made from the 80% remaining after 20% is selected by the original method.

RANSELF2: A random selection is made from the 80% remaining after 20% is selected by the compensatory method.

In Figure 3 the crosses represent the applicants selected by institution P, the circles represent the applicants selected by institution Q and the pluses represent the rejected applicants. The two selections are very different. In the first the applicants attending Q have very varied scores on the selection test but do not vary much on the score from the original method. In the second the two scores are inversely related with an applicant with a high score on the original method having a low score on the selection test and vice versa. The effect on the outcomes is that predicted gains from having a high score on the original method will be cancelled out by the losses associated with a low score on the selection test. Obviously in the real world the effect of self-selection is not so clear cut because applicants apply to more institutions and do not necessarily apply to the best institution where they could have gained a place or have taken up the place if they applied and were successful.

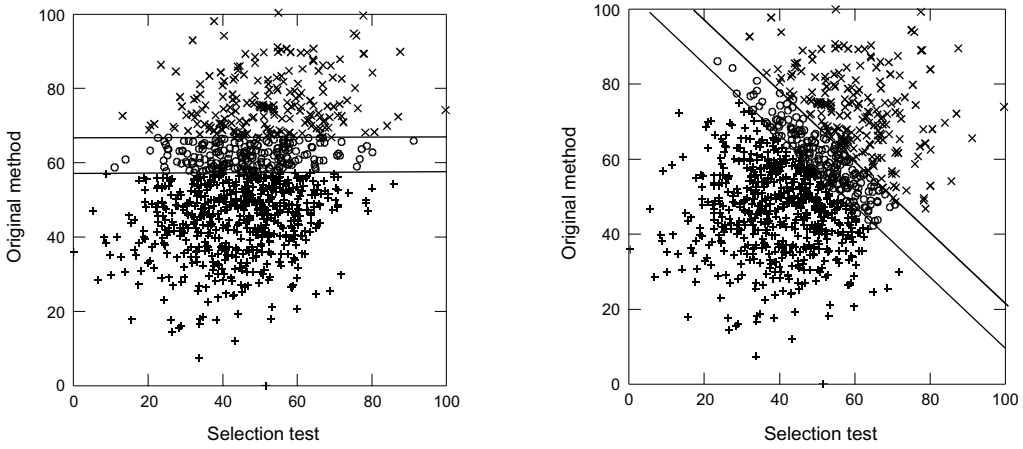


Figure 3 : The effect of self-selection on applicants accepted by two institutions

(a) Using original method for both selections (SELF1) **(b) Using compensation method for both selections (SELF2)**

The last two rules serve as baselines for the first two rules. RANS1 are the results for a random sampling after institution P had selected 20% by the original method and RANS2 is the same apart from the use of the compensation rule (i.e. taking a random selection of candidates from below the upper lines in Figure 3 ignoring the lower line). The effect of the self-selection is to reduce the relative proportion of good applicants that can be selected.

Using the summary statistics in Table 3, it is clear that if the effects of selection are ignored then this could lead to serious misinterpretation of the data. For the situation described by SELF1 then it might be concluded that the new test was greatly superior to the original method. Although the correlations for the two tests are similar in the whole population, the correlation of the original methods is much greater for the selection test. For SELF2 it is possible to erroneously conclude that the selection method was ineffective and they would do better switching to a lottery. Both the correlations for candidates attending in institution Q are close to zero. However, the institution would get a much poorer entry if they did so (i.e. 27% grade A candidates for SELF2 and just 11% for RANS2). Although this example is a simulation, it is not the case that it has just been cunningly contrived to illustrate an unlikely theoretical situation; such problems occur in real life. Linn and Dunbar (1982) found that the correlations between SAT scores and subsequent performance were low for a New York community college. This was the result of students who scored highly on the SAT almost always choosing to go to better colleges.

This simulated example is obviously a gross simplification given that institutions would not necessarily use the same selection procedures and more than one institution may be involved. However, recently there has been research into applying a range restriction to situations involving institutional and self-selection. Yang, Sackett and Nho (2004) proposed a procedure using non-ignorable double selection models and found that in simulations their model produced an unbiased estimate for the population correlation.

Evaluating rules in this situation is also more complicated. If institution Q improves its selection procedure then this would not guarantee an improved entry. This is because the quality of the available applicants can also decline if institution P also improves its selection method. Such a situation would occur if both institutions introduced a selection test at the same time.

The simulated data used in this paper has demonstrated that interpreting uncorrected correlation coefficients is difficult and, depending on the circumstances, can seriously underestimate the effectiveness of a selection test. The correlation coefficient can be corrected for the effects of selection but it is important to recognise that the correction method should match the selection procedure. Unfortunately, the corrections depend on assumptions about the rejected applicants. Although it is usually argued that the assumptions made about the rejected applicants are untestable, this is not quite true. Selection tests are usually used repeatedly, meaning that the effects of changes can be monitored. For example, consider the simplest case, where in the first year the selection test is piloted but not actually used to select students. Then the range correction formula for this situation can be applied. If in the second year the entry is identical in characteristics to the first year and the selection test alone is used then a prediction of the expected correlation can be made by inverting the appropriate correction formula. This can be compared to the observed correlation.

More fundamental, however, is the question whether using the correlation coefficient in the first place as a measure of the predictive

Table 3 : The results for the self-selection rules

	Statistics			*Correlations*		*Grades*				
Method	*N*	*Mean*	*Sd*	*X1*	*X2*	*A*	*B*	*C*	*D*	*E*
All	1000	45	17	0.56	0.54	20	40	60	80	100
SELF1	202	49	14	0.51	0.20	21	54	65	81	100
SELF2	200	52	12	0.05	0.01	27	69	83	96	100
RANS1	199	43	15	0.50	0.46	16	32	47	79	100
RANS2	200	42	14	0.50	0.28	11	32	55	77	100

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