References

AQA (2009). Uniform marks in GCE, GCSE and Functional Skills exams and points in the Diploma. http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF Accessed 16/02/10.

Baird, J.-A. (2007). Alternative conceptions of comparability. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, **25**, 3, 271–284.

Gray, E., & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication*, **7**, 32–37.

Newton, P.E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, **25**, 3, 285–292.

Oates, T. (2009). 'Standards are up this year' – what does this mean? The question of standards in public examinations. http://cambridgeassessment.files.wordpress.com/2010/01/the-question-of-standards-in-public-examinations-by-tim-oates1.pdf Accessed 17/5/10.

Ofqual (2009). GCSE, GCE and AEA code of practice, April 2009. http://www.ofqual.gov.uk/files/2009-04-14-code-of-practice.pdf Accessed 08/01/10.

Paton, G. (2010). GCSE and A level results being 'inflated'. *Daily Telegraph*. http://www.telegraph.co.uk/education/educationnews/7528383/GCSE-and-A level-results-being-inflated.html Accessed 17/5/10.

Stringer, N. (2008). *An appropriate role for professional judgement in maintaining standards in English General Qualifications*. Paper presented at the International Association for Educational Assessment annual conference, Cambridge, September 2008.

An American university case study approach to predictive validity: Exploring the issues

Stuart Shaw and Clare Bailey CIE

Introduction

Predictive validity research is fundamental to test validation (Davies *et al*., 1999). Predictive validity entails the comparison of test scores with some other measure for the same candidates taken some time after the test has been given (see Anastasi, 1988; Alderson *et al.*, 1995). In psychometric terms, predictive validity is the extent to which a scale predicts scores on some external (future) criterion measure. It is the prediction of criterion performance that is basic to validation. For tests that are used for university selection purposes it is vital to demonstrate predictive validity.

However, establishing predictive validity through relating secondary school performance to later academic performance is fraught with practical difficulties in mounting tracer studies and the problems associated with confounding intervening variables that obscure the effects of another variable (see Banerjee, 2003, for a critique of such approaches to establishing predictive validity). These difficulties notwithstanding, predictive validity is still regarded a vital aspect of the validation process. Moreover, predictive validity research is becoming increasingly necessary as test providers are being challenged to pay greater attention to issues of test comparability – both in terms of the relationships between their own assessment products and those offered by other competitor, examination boards.

A common need for predictive validity is inherent in the process of selecting students for university. Consequently, this article will focus on the research being conducted by University of Cambridge International Exams (hereafter simply 'Cambridge') to ensure that its international assessments prepare students well for continued studies in colleges and universities. The long-term purpose of the research is to highlight the predictive validity of Cambridge assessments and other students' characteristics to predict preparedness for and continued academic success at U.S. universities in terms of first year Grade Point Average (GPA).

This study takes a case study approach. The research reported here uses data collected from three cohorts of students enrolled at Florida State University. The data include information about each student's performance at high school, ethnicity, gender and first year GPA. Multilevel modelling has been applied to the data using the statistical software package MLwiN[1] to investigate the relationships between the variables, and in particular to determine which are the best indicators of academic success at university, whilst taking into account the effects of individual high schools. Issues relating to choice of predictive and university success measures, intervening variables, controlling for selection bias, data and measurement, and choice of research model will be discussed in the context of an American university.

U.S. secondary school indicators for success

Given the increase in the number of applications for admissions to colleges and universities for the limited number of seats in freshmen classes, students and universities in the U.S. must consider all available indicators for success in higher education. There are many ways a student can gain recognition to contribute towards their university application. The standard high school exam in the U.S. is the SAT (formerly known as the Scholastic Aptitude Test) although in some states an alternative, the

[1] www.cmm.bristol.ac.uk/index.shtml

[2] Concordance tables are published to find equivalences so that SAT scores can be used for the minority of students who take the ACT.

ACT (American College Testing), is more popular[2]. In this study we are considering students in Florida, where the majority take the SAT exam. Although standardised test scores have varying significance in the admission decisions of all students who qualify for admission at universities in the U.S, all potential U.S. university students must submit results of college entrance exams, either SAT or ACT, in order for an application to be considered complete in many universities. In addition to this, students can choose to take additional exams, such as those that are part of the Advanced Placement (AP), the International Baccalaureate (IB) or Cambridge's International A level programme (AICE)[3].

Advanced Placement has been a staple in U.S. education for over fifty years. Designed to promote excellence in secondary education, the programme desires to allow motivated students to work at their optimum capability. Nearly one million U.S. students now take at least one AP exam during their secondary careers. As Harvard, Yale and Princeton Universities were active participants in the study that led to the creation of AP, the acceptance of this credential is nearly universal among American universities.

In the late 1960s the International Baccalaureate was founded. While initially established as a single programme for internationally mobile students, the programme has flourished throughout the world, but nowhere greater than in the U.S. By 2005 over 1,000 secondary schools in North America offered the IB curriculum. The IB had to work diligently to have U.S. universities provide recognition similar to that provided to AP.

While Cambridge has been offering examinations for 150 years, it is relatively new in offering its curriculum in the U.S. The four year IGCSE/AS/A level curriculum and exams leading to an Advanced International Certificate of Education Diploma were introduced in Florida's Bay High School a little over fifteen years ago. Cambridge is experiencing the same curve of recognition as IB experienced in the 1970s and 1980s.

A tabulated comparison of secondary education in the UK and the US is shown as an appendix.

Explanations of terms used

For the benefit of readers who may not be familiar with the U.S. high school and university system we include here some explanations that may be helpful.

Cambridge Advanced International Certificate of Education Diploma: Cambridge awards a Cambridge AICE Diploma to students who have passed a prescribed number of subject examinations at the Advanced (A) level and/or the Advanced Subsidiary (AS) level. To qualify for a Cambridge AICE Diploma, students must pass at least one examination from each of three subject groups to include Mathematics and Sciences, Languages (both foreign and first), and Arts and Humanities. In the US, Cambridge International AS and A level examinations are sometimes referred to as 'Cambridge AICE' or 'AICE' examinations. Students passing AS and A level examinations may be awarded entry level or intermediary level university course credit by examination or advanced standing at US colleges and universities.

Advanced Placement: The AP programme is a curriculum in the US sponsored by the College Board[4] which offers standardised courses to high school students that are generally recognised to be equivalent to undergraduate courses in college. Participating colleges grant credit to students who obtained high enough scores on the exams to qualify. During their secondary studies a student may opt to take many AP courses, or as few as one. This curriculum is the most widely spread acceleration mechanism offered in the US and has been in place for over fifty years.

Credit hour: Each course that a student can enrol on is worth a certain number of credit hours. One credit hour is normally equivalent to 'one hour of classroom instruction and two hours of student work outside class over 15 weeks for a semester' so that a typical course is worth 3 hours, and this can vary from 1 to 5. Different institutions can vary how much credit is assigned to Cambridge AICE, AP or IB results.

Dual enrolment: Dual enrolment is normally concurrent enrolment where a high school student is taking a college course for both high school and college credit. This may be done by the student being released from his/her high school and taking the course on a college campus, or by the college approving the curriculum and allowing the student to remain on the high school campus and the college appointing the secondary school instructor as an adjunct faculty member at the college. Many students will earn a year of college credit in this manner, and some students will earn as much as two years of credit through dual enrolment. Many parents see dual enrolment as a money saving strategy to avoid high tuition costs at universities and state governments see this as a net saving since public school costs are lower than they would be at post secondary institutions.

High school GPA: High schools in the US determine how to calculate GPAs for purposes of generating a rank distribution. The system gives 4 points for a grade A, 3 points for a grade B and so on, and then takes the average, so that the final score is out of 4. (Given different weighting systems for advanced level courses, the GPA could exceed 4.) The lack of moderation in this process makes it more difficult to give standardised measures of high school performance, although there is evidence to suggest that HSGPA is nevertheless a good predictor (Betts and Morrell, 1999). One possibility is to sort students into categories based on their rank.

International Baccalaureate: The IB diploma programme is offered at over 3,000 schools in over 130 countries. The diploma programme is a two year programme and to receive an IB diploma a student must complete courses in social studies, mathematics, experimental sciences, their primary language and a second language. A sixth course must also be completed with a choice of an arts course, or a second course from the five disciplines mentioned above. In addition to the six courses, students must complete an extended essay, complete a course titled 'Theory of Knowledge' and complete a requirement of activity beyond the classroom. Three courses must be completed at the Higher Level while the other three can be taken at the Standard Level. College credit and placement may be earned, although the amount of credit and the score necessary to receive credit will vary by institution.

No Credit: Nearly all US high schools have what is commonly referred to as a 'college preparatory' curriculum. This curriculum is designed to

[3] http://www.cie.org.uk/qualifications/academic/uppersec/aice

[4] The College Board is a not-for-profit membership association in the US that was formed in 1900 as the College Entrance Examination Board (CEEB) www.collegeboard.com

[5] http://www.universityworldnews.com/article.php?story=20100625183517482

prepare a student for successful study at the college level. If no credit is included that could mean that no acceleration mechanism such as Cambridge AICE, IB, AP or dual enrolment has been included in the course of study or the student took an AP/IB/AICE curriculum, but did not score sufficiently to receive credit.

SAT and ACT scores: Almost all students take either the SAT exam or the ACT exam, and some take both. The SAT was revised in March 2005. The revisions were made to enhance the test's alignment with current high school curricula and emphasise the skills needed for success in college (see Lawrence, Rigol, Van Essen, and Jackson, 2003, for a detailed explanation of the changes).

The SAT is composed of three exams:

- Critical reading (SAT-CR)
- Mathematics (SAT-M)
- Writing (SAT-W)
- Total (SAT-Tot)

The score scale range for each section is 200 to 800 and the score scale range for the total is 600 to 2400. The official SAT website[6] states that, for 2006, a total score of 1800 means the candidate scored better than 80.8% of test takers. Admittance into many highly regarded American colleges requires scores above 1800, although entry will also depend upon a student's academic transcript (record of academic achievement) and extracurricular activities.[7]

Florida State University: a case study

This study takes a case study approach using data from Florida State University. Denscombe (2003) describes the key characteristics of case study research: spotlight on one instance; in-depth study; focus on relationships and process; natural setting; and multiple sources and methods. (For detailed explanations and discussions of case study research, see Denscombe, 2003; Bell, 2005; Cohen, Manion and Morrison, 2007; and Sharp, 2009.)

In general, case studies can be used to: (a) provide a thick description of complex interactions to enhance understanding of a range of social phenomena, (b) corroborate theoretical suppositions, and (c) generate and contribute to theory (Eisenhardt, 2002; Yin, 2006). Therefore, when giving consideration to case study methodology, it is necessary to understand it as "both a process of inquiry about the case and the product of that inquiry" (Stake, 2008, p. 121).

Florida State University (FSU) is a publicly supported institution located in the state capital of Tallahassee. FSU is a comprehensive, national graduate research university with 40,255 students of whom 8,557 are graduate students. FSU is home to the National High Magnetic Field Laboratory and their arts programme – dance, film, music and theatre – is widely regarded within the U.S. Recently FSU added a College of Engineering and a College of Medicine. The university also has a College of Law.

Exploring the issues

In what follows we outline some of the issues relating to the implementation of a predictive validity study in the context of an American university.

Choice of predictive success measure

A challenge to all models interested in prediction is the choice of predictive success measure.

The College Board encourages universities to use SAT and high school grades when making admissions decisions. However, high school grades are not necessarily a good means of comparing students' experiences and achievements prior to university. This is because high school grades reflect the standards and quality of a particular school or schooling system. These standards differ according to school area or region (e.g. urban or rural) and even individual schools. Moreover, inter-school effects are not always reflected in high school grades (Burton and Ramist, 2001).

The primary purpose of the SAT is to measure a student's potential for academic success in college. In this context, a number of studies have been undertaken which attest to the predictive validity of the SAT. (For a useful summary relating to the predictive utility of SAT, ACT and high school GPA [HSGPA] as indicators of university success see Cohn, Balch and Bradley, 2004.)

Cohen, Manion and Morrison (2007) used SAT scores, HSGPA and high school class rank to determine how well these predict college GPA. Data were collected from 521 students enrolled on Principles of Economics at the University of South Carolina in 2000 and 2001.They examined the frequency distribution of key variables and regression analysis (no multilevel model), with students grouped according to gender and race. It was found that having a SAT score of over 1100 (out of a possible 1600) and a class rank of over 70 gave a predicted college GPA of around 3.0.

A large-scale national validity study of the revised SAT (incorporating an additional section in writing and minor changes in content to the verbal and mathematics sections) was undertaken by Kobrin, Patterson, Shaw, Mattern, and Barbuti (2008). Their studies were based on data from 150,000 students from 110 four-year colleges and universities across the US entering 110 four-year colleges and universities in the fall of 2006 and completing their first year of college in May/June 2007. The writing section was shown to be the single most predictive section of the test for all students. The analyses also found the writing section to be the most predictive across all minority groups. The studies also revealed that:

- SAT is an excellent predictor of how students perform in their first year at university;
- SAT is a stronger predictor than high school grades for all minority groups (African American, Hispanic, American Indian and Asian);
- the recently added writing section is the most predictive of the three SAT sections.

Culpepper and Davenport (2009) studied a sample of 32,103 first-year students who were enrolled in one of 30 colleges or universities in 1995. They compared the attainment of students from different racial/ethnic backgrounds, and found that an African-American student with the same HSGPA, SAT or ACT score as a white student was likely to have a lower college GPA. The possible differential prediction of SAT scores for university performance by race highlights the need to control for race in models involving SAT scores.

However, not all studies have produced evidence that the SAT

[6] www.satscores.us

[7] Interpreting SAT Scores and ACT Scores. University Language Services. http://www.universitylanguage.com/guides/interpreting-sat-scores-and-act-scores/

identifies the students most likely to succeed at university. Lenning (1975) carried out three studies to determine whether ACT was as good a predictor of college grades as SAT for highly selective institutions. Although only three such institutions were studied, they found that ACT scores can be at least as predictive, and likely more predictive, of college grades at highly selective institutions than SAT scores.

Noble and Sawyer (1987) considered the ACT scores and HSGPA for students enrolled at 233 institutions across 2812 courses in October 1985. They computed regression statistics for each course. They found that including HSGPA gave a stronger prediction of college GPA.

Noble (1991) conducted a study of 30 colleges, mainly located in central and southern U.S, with a higher than representative proportion of public colleges. It was found that ACT is a reasonable predictor of college success, and that including HSGPA improves the predictive validity.

A study by Betts and Morrell (1999) also indicated that HSGPA (as well as SAT scores) are significant predictors of university GPA.

Choice of university success measure

Another challenge to models interested in prediction is the choice of university success measure. For example, a number of different university performance measures could be used. These may include:

- average GPA for first year (or other years if available)
- number of courses passed
- number of courses excelled in
- GPA in certain courses, for example, science/mathematics versus humanities
- university enrolment status (as of the second fall after high school graduation)
- university retention, that is, re-enrolment in a second year at the same institution (Robbins, Allen, Casillas, Peterson, and Le, 2006)
- certain measures of engagement, for example, more propensity to participate in research at university or study abroad, more likely to participate in a student activity of some kind, etc.

However, the ultimate choice of performance measure would depend on data available and whether the data provide a comparable measure across courses included in the study.

The concept of tertiary level academic success used here is determined by the persistence of a student within the university with a specific GPA. The definition of university GPA employed is based on the accumulation of all previous semesters' work. In this study we are considering the GPA for students attending just one university. However, future studies will entail collecting data from a number of universities which may create different challenges. For example, it would appear that U.S. universities demonstrate some degree of latitude in determining how to calculate GPAs.

Choice of research design and hypotheses

In order for the research to be well-founded, we must ensure that:

- the analysis of statistical indices (e.g. correlations, regression coefficients) is technically sound and in particular that it:
  - addresses a set of testable hypotheses, derived from a sound theoretical approach, and
  - uses appropriate empirical methodologies and data for the purpose
- any inferences drawn from the analysis are justified and that erroneous inferences in the public domain (as may be drawn by third parties) are either avoided, or otherwise addressed and corrected as appropriate.

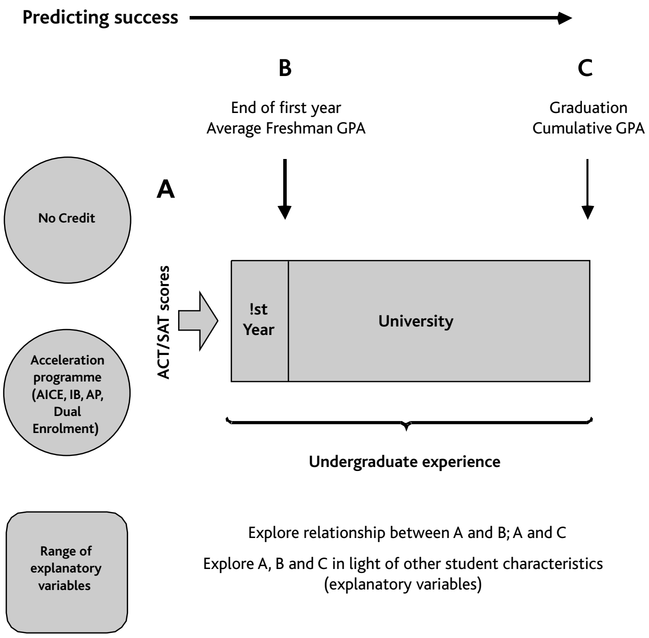
The principal hypothesis tested in this initial, exploratory study may be stated in the following way:

> *Students who follow the AICE, AP or IB programmes will achieve a significantly higher first year GPA than those with no credit, given the same SAT scores.*

The research designed to test this hypothesis may entail the formulation of several preliminary model specifications (each based on unit data where each student represents a single observation).

In order to estimate predictive validity it is necessary to determine the relationship between the success of students leaving high school following a particular programme of study and their success during, or at the end of, undergraduate study. Such a model is shown conceptually in Figure 1.

Figure 1: Predictive validity research design


Predicting success
B
End of first year
Average Freshman GPA
C
Graduation
Cumulative GPA
A
No Credit
ACT/SAT scores
!st
Year
University
Acceleration programme (AICE, IB, AP, Dual Enrolment)
Undergraduate experience
Range of explanatory variables
Explore relationship between A and B; A and C
Explore A, B and C in light of other student characteristics (explanatory variables)

A number of other models also have potential. For example, a test of predictive power using students who sit common examinations (i.e. a within-subjects design) – the hypothesis being that one assessment explains more variation in their university performance.

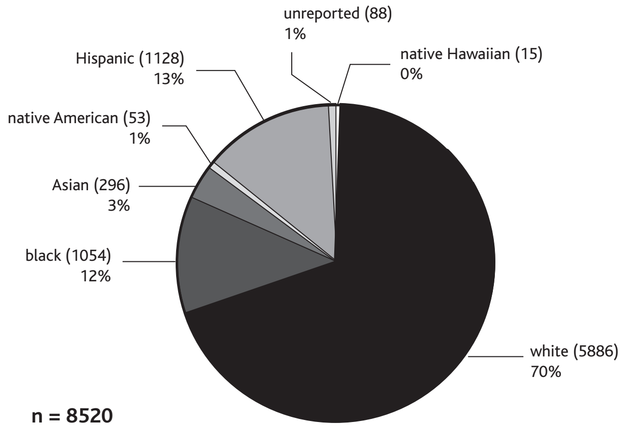
Choice of data and measurement model

The SAT score (total SAT score, SAT-Tot) has been used here as the choice of measure for high school performance. A point worthy of note is when students take the SAT. If students take the SAT late junior year or early senior year, then any additional acceleration programme may have an effect on their score.

To fit the multilevel models we used data based on records of over 8500 students who entered Florida State University during the academic years 2007/2008, 2008/2009 and 2009/2010.

Four datasets representing secondary educational programs were obtained from enrolment and admissions staff at the university. The largest data set (*n* = 6382) contained information on students with only

Figure 2: Pie chart to show the proportion of students of each race



the SAT (or ACT) score (hereafter referred to as having 'no credit'). The three other data sets contained information on students with Cambridge AICE credit (*n* = 144), with AP credit (*n* = 1188) and IB credit (*n* = 806). Figure 2 shows student data in terms of relative proportions by race.

Column headings for each of the four datasets include: FSU student number, year enrolled, race, gender, FSU GPA, high school GPA, SAT verbal, SAT math, SAT total, ACT (if applicable), high school attended, type of exam program followed (if applicable). The explanatory variables are set out in Table 1.

Table 1: Explanatory variables definition

Generic data requirements	
Variable	*Explanation*
FSU student number	Unique student identifier
Race	1 = white, 2 = black, 3 = Asian, 4 = native American, 5 = Hispanic, 6 = unreported, 7 = native Hawaiian/other Pacific islander
Gender	M = male, F = female
FSU GPA	Possible values from 0 to 4
High school GPA	Possible values from 0 to 4 (or in some cases more than 4)
Matriculation year	Year first enrolled at FSU
SAT verbal	SAT score for critical reading component
SAT math	SAT score for math component
SAT total	Total SAT score
ACT composite	ACT score
High school code	Local high school identifier
Type of credit	Exam program followed – Cambridge AICE, AP, IB or no credit
Credit hours	Number of hours credit gained on a college course

The four data sets were combined into an overall matrix. The structure of the data, which contain students from (i.e. 'nested within') a number of high schools, suggests the use of multilevel models. The multilevel software package MLwiN (Version 2.02 Rasbash *et al*., 2005) was therefore used.

Multilevel modelling is a way of finding a line of regression through different groups, nests or hierarchies of data (unlike standard multiple regression techniques which assume that the observations are independent, which is not the case here). Multilevel models recognise the existence of both hierarchical data and clustered data structures.

Multilevel modelling takes account of the context in which a variable exists. It is often used in sociological applications because individuals are affected by, or defined by, the groups they belong to. For example, patients receiving the same treatment for the same condition at different hospitals may experience different patient outcomes; students in different classes or in different schools may obtain different exam results (outcomes). A two-level model which controls for student outcomes within high schools would include residuals at both the student and school level. In effect, residual variance is separated out into an inter-school constituent (the variance of the school-level residuals) and an intra-school constituent (the variance of the student-level residuals). The school residuals ('school effects') represent unobserved high school characteristics that affect student outcomes, more particularly student performance. The unobserved variables lead to correlation between outcomes for students from the same school.

Recognising how groups of individuals can be nested can help build a more realistic picture, giving insight into where and how effects are happening, and this is what multilevel modelling aims to do (see Goldstein, 2011; or Bryman and Hardy, 2009, for a more detailed description of multilevel modelling).

Not using a multilevel model as a result of failing to recognise hierarchical structures makes it more likely that a significant difference is reported when in fact the difference is non-significant (i.e. a false positive or type 1 error): standard errors of regression coefficients will be underestimated, leading to an overstatement of statistical significance. Standard errors for the coefficients of higher-level predictor variables will be the most affected if the effect of grouping is ignored.

As the outcome variable (FSU GPA scores – first year examination marks) is continuous, the model fitted was:

$$y_{ij} = \beta_{0ij}x_0 + \beta_1 x_{ij}$$
$$\beta_{0ij} = \upsilon_{0j} + \varepsilon_{0ij}$$

where y_{ij} is the predicted outcome variable (FSU GPA score) for individual i in high school j, β_{0ij} is a constant, β_1 is the independent contribution of the predictor variable to the dependent variable, x_{ij} is a predictor variable, υ_{0j} is high school level residual error and ε_{0ij} is individual level residual error.

Multilevel models have been used in several predictive studies to take into account the hierarchical structure of educational assessment data. For example, Bell and Dexter (2000) used multilevel modelling to investigate the comparability of GCSE and IGCSE and suggested that a wide between-school variation can make results misleading. However, this is the first study to our knowledge that uses multilevel modelling to compare the predictive validity of different types of high school exam programmes in the US.

Initial findings

Figure 3 shows the total SAT scores and the FSU GPA for each student in the dataset according to the exam programme followed. It can be seen that there are a number of outliers at the FSU GPA level – students who perform well in their SAT score but who do not do so well in their first year of college. In every case where students exhibit a zero score for their

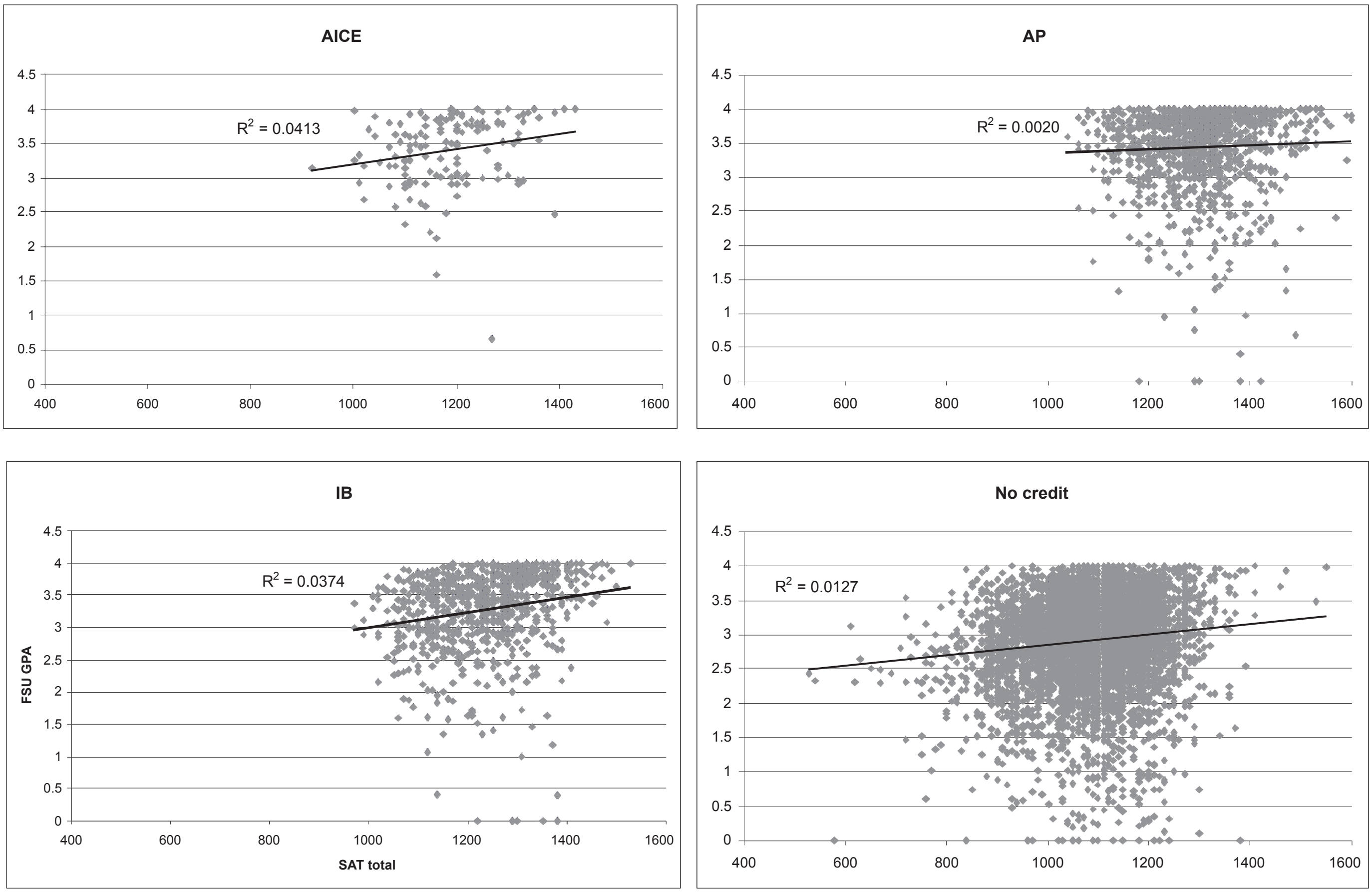


Figure 3: Scatter plots of the four datasets for each type of exam programme, showing SAT-Tot against FSU GPA and the line of regression and r^2 value

GPA it was noted that these were new students yet to receive a GPA. According to university admissions staff, any instances of low GPA scores are representative of underperforming students experiencing academic difficulties. It may be assumed, therefore, that these are special cases which a model could not reasonably predict. Consequently, any student with a GPA of less than 1.0 was excluded from the data set. It should also be noted that most of the student GPAs shown in Figure 2 fall within the range 2–4 (though this range is wider for 'no credit' students).

The SAT scores for students with no credit are considerably lower than those of the other three groups.

Using the refined dataset (excluding FSU GPA scores less than 1.0 and with the 488, or 5.7% of candidates missing SAT-Tot scores replaced with equivalent ACT) the model investigates the factors associated with the course of programme study (Table 2). Regression coefficients are statistically significant if they equal twice or more the value of the standard error (shown in brackets). Statistically significant effects are shown in bold type. It should be noted that school-level effects appeared to be much smaller than the individual-level effects: there is no statistical difference between schools.

Table 2: Effect of educational programme (given equivalent SAT scores) on FSU GPA

Base – No credit	*Regression Coefficient (Standard Error)*
AICE	0.351 (0.053)
AP	0.359 (0.023)
IB	0.222 (0.026)

Compared to students with no credit (and controlling for the effects of SAT scores, gender and race), having taken the AICE, AP or IB programmes were all associated with significantly higher first year GPAs.

- Students who took the AICE attained, on average, a GPA of 0.35 higher than those with no credit, given the same SAT score.
- Students who took the AP attained, on average, a GPA of 0.36 higher than those with no credit, given the same SAT score.
- Students who took the IB attained, on average, a GPA of 0.22 higher than those with no credit, given the same SAT score.

Discussion

The aim of this study has been to determine how well acceleration programmes in the U.S. prepare students for success at university. This general question can be extended: by using multilevel modelling, we can ask how well a given exam programme prepares a student who comes from a particular educational background. The study has explored the link between high school quality (in terms of programme followed) to first year university academic achievement using data supplied by Florida State University.

Consideration of the issues and exploratory analysis of the data collected so far has enabled us to test whether students who follow the AICE, AP or IB programmes achieve a significantly higher first year GPA than those with no credit, given the same SAT scores and controlling for the effects of race and gender. The results show that following an examination programme results in, on average, a better GPA than not following any extra credit.

Validity considerations

On the inclusiveness of validity, Bachman has argued that it is important to recognise that no one type of validity evidence by itself "is sufficient to demonstrate the validity of a particular interpretation or use of test scores" (1990, p.237). Validity is a multi-faceted concept requiring a range of types of evidence to support any claims for validity of scores on a test: "These are not alternatives but complementary aspects of an evidential basis for test interpretation" (Weir, 2005, p.13). However, for studies of this kind predictive validity work must take priority for tests designed for use in university selection if the tests are to be seen as fit-for-purpose.

According to Weir (2005), establishing predictive validity through correlating secondary school performance or standardised tests against later academic performance is impeded by practical and logistical difficulties. Such problems are particularly pronounced when implementing tracer studies and also when attempting to identify and control for a range of confounding intervening variables (See Banerjee, 2003, for a critique of approaches to establishing predictive validity.) Conceptually, therefore, any predominantly quantitative and *a posteriori* estimation of validity should be triangulated with qualitative data collected from, for example, individuals within one of the main stakeholder groups: the learners and their teachers. There is a requirement for any examination board to demonstrate and share how they are seeking to meet the demands of validity in their assessments and to make every systematic effort to ensure that their assessments achieve a positive influence or impact on general educational processes and on the individuals who are affected by the results. Predictive validity and impact studies are important contributions, therefore, to the validation process of any assessment.

Weiss defines impact – from the perspective of educational evaluation – as "the net effects of a programme (i.e. the gain in outcomes for program participants minus the gain for an equivalent group of non-participants)" (1998, p.331). Acknowledging the narrowness of this definition, Weiss broadens its scope by adding that "impact may also refer to program effects for the larger community … more generally it is a synonym for outcome". Investigating impact is regarded as being an essential aspect of determining the utility (or usefulness) of an educational assessment in terms of fulfilling its intended purpose, that is, its fitness for specific purposes (validity broadly interpreted) and contexts of use. Embedded within the concept of impact reside the notions of *processes* as well as *outcomes* (or products). Roy (1998) distinguishes between the two:

> *A study of the product is expected to indicate the pay-off value while a study of the process is expected to indicate the intrinsic values of the programme. Both are needed, however, to find the worth of the programme.* (1998, p.71)

As there are a number of variables that can weaken the reliability of the conclusions drawn from this study, it is intended that the findings from a series of US impact studies will be used to support any predictive validity estimates.

It is important to the interpretation of any predictive research, therefore, that impact data collection instruments and procedures (such as questionnaires and interview schedules) are used in order to understand the test impact better and to conduct effective surveys to monitor it (Hawkey, 2004). Currently data are being collected in order to ascertain stakeholder perceptions of Cambridge assessments in the US educational system. School lesson observations together with semi-structured interviews and focused discussion groups with both students and teachers have been conducted in an attempt to gather information on pedagogic practice, lesson content, learning/study approaches and perceived features of test validity and reliability. These data have been enlarged and enriched through the collection of views provided by Higher Education admissions and teaching staff on how examination results are used and how secondary educational study programmes provide an indication of tertiary level preparedness and success. It is hoped that by undertaking longitudinal research and eliciting participants' perspectives on their own behaviour, a number of recurrent patterns across data sets will emerge thereby revealing "multiple aspects of a single empirical reality" (Denzin, 1978). Such an approach will provide Cambridge with greater clarity regarding their own assessments in terms of "what goes on while a program is in progress" and "the end results of the program" (Weiss, 1998, pp.334–335). Impact research will enable a closer exploration of the relationship between the experience of students in the Cambridge curriculum and the level of preparation for college as well as the level of success at college.

Study limitations

The focus of the research has been a case study. Case studies include both a process of inquiry that is grounded in interpretations and a contribution to a product from that inquiry. Although a case study methodology is not without its criticism (being a bounded investigation which suggests that products are not readily generalizable), "compared to other methods, the strength of the case study method is its ability to examine, in-depth, a 'case' within its 'real-life' context" (Yin, 2006, p.111).

A case study approach uses a constructivist/interpretivist orientation toward data collection and analysis processes. A case study methodology recognises the need for:

- multiple perspectives (as evidence that contributes to case descriptions); and
- multiple methods (in order to isolate and scrutinise perspectives within case studies).

Its adoption, therefore, is justified as a mode of situated inquiry, favouring uniqueness over generalizability.

The size of the dataset was large – over 8 500 students. This means the reliability we can attach to the findings is increased. Even where the sub-sets were small – for example, of Cambridge AICE students there were 144 – they were still sufficiently large for the analyses to be carried out. There were some sub-sets that were small, for example native American and Hawaiian, which increases the risk of Type II errors. (This is the error of failing to observe a difference when in truth there is one – a false negative.)

A common challenge in studies of this type is controlling for selection bias. The choice of educational programme is not necessarily random. High schools have different characteristics and in mixed Cambridge /non-Cambridge high schools students may have a choice. Students also may choose a high school based on its use of programme. To control for such potential bias, it would be useful to have some control variable that is correlated with the choice of system but otherwise unrelated to the student's performance at university. Typically we would expect the choice of system and student performance to be quite related. It is not clear what determines the choice of acceleration mechanism. Is choice of educational programme influenced by type of high school, extrinsic and intrinsic motivational aspects, institutional ethos, affective characteristics, parental status, socio-economic constraints? Why do some students choose not to avail themselves of an acceleration programme? Clearly information of this kind would enhance our understanding of future predictive validity findings.

Future work

Further multivariate modelling work will include investigation of other variables which might explain student performance. Apart from a programme of learning these could include other students' characteristics such as socio-economic status, university enrolment status and university retention rates.

Other measures could include class type (whether Cambridge students do better with certain types of classes) or if certain behavioural measures, such as engagement with research or study abroad, might be enhanced. Apart from the freshman year cumulative GPA measure of achievement, other university performance outcomes could be explored, for example, four-year cumulative GPA scores; freshman year attrition rates; and four-year graduation rates. Additionally, it would be informative to compare SAT critical reading and SAT mathematics scores as there is some evidence that one is a better predictor of college success than the other.

All of the variables used for the above analyses come from university admissions records. Student transcripts from the administrative archives of the university provide information about university career (type and number of exam passed, frequency of study, credit hours, etc.) and data relating to some characteristics of the high schools attended (type of school, final grades). However, a questionnaire given to students when they enter university would enable the collection of additional information on the students' characteristics such as reasons for choice of educational programme and familial socio-economic status.

A valuable, longitudinal exercise would be to track an entire cohort of Cambridge students from one particular high school through to final year of study. Questionnaire surveys together with interviews throughout the duration of an AICE course could be undertaken in order to determine extent of workload, attitudes to course/assessment and teachers'/ students' perceptions of the course. This would be accompanied by follow-up interviews with students at university, the findings from which could be triangulated with GPA scores achieved at the end of the first year of undergraduate study and also at graduation.

Given the smaller numbers in the AICE, AP and IB groups, the case study nature of the research and the possible presence of unknown confounding variables between groups, it would be unwise to draw conclusions about the relative predictive strength of the three acceleration programmes. Further work will be required to collect more data from both Florida State University and other U.S. universities. Cambridge has already obtained considerably smaller datasets from the universities of Maryland, Virginia and Michigan and the process of data collection is expected to continue over time.

Acknowledgements

We would like to thank John Barnhill (Assistant Vice President for Enrolment, Florida State University) and Megan Benson (Director, Enrolment Management Operations Office of Admissions and Records, Florida State University) for providing us with the university data and for their assistance and advice during the project. Thanks also to Sherry Reach (Cambridge US Regional Manager) and Bill Kolb (Cambridge US Recognitions Consultant) for their invaluable contributions throughout the course of this work.

References

Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Anastasi, A. (1988). *Psychological Testing*. 6th edition. New York: Macmillan.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Banerjee, J. V. (2003). *Interpreting and using proficiency test scores*. Unpublished PhD thesis, University of Lancaster.

Bell, J. (2005). *Doing your research project: A guide for first-time researchers in education, health and social science*. 4th edition. Maidenhead: Open University Press.

Bell, J. F. & Dexter, T. (2000). *Using multilevel models to assess the comparability of examinations*. Paper presented at the 5th International Conference on Social Science Methodology, October 2000.

Betts, J. R. & Morrell, D. (1999). The determinants of undergraduate grade point average. *Journal of Human Resources*, **34**, 2, 268–293.

Bryman, A. & Hardy, M. A. (2009). *Handbook of data analysis*. London: Sage Publications.

Burton, N. W. & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980*. College Board Research Report No. 2001–02, College Entrance Examination Board, New York.

Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the behavioural sciences*. 2nd edition. Hillside, NJ: Lawrence Erlbaum Associates Publishers.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. 6th edition. Abingdon: Routledge.

Cohn, E., Cohn, S., Balch, D. C., & Bradley, J. (2004). Determinants of undergraduate GPAs: SAT scores, high school GPA and high school rank. *Economics of Education Review*, **23**, 277–286.

Culpepper, S. A. & Davenport, E. C. (2009). Assessing differential prediction of college grades by race/ethnicity with a multilevel model. *Journal of Educational Measurement*, **46**, 2, 220–242.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of Language Testing.* Studies in Language Testing 7. Cambridge: UCLES and Cambridge University Press.

Denscombe, M. (2003). *The good research guide for small-scale social research projects*. 2nd edition. Maidenhead: Open University Press.

Denzin, N. (1978). *Research Act: Theoretical Introduction to Sociological Methods*. New York: McGraw-Hill.

Eisenhardt, K. M. (2002). Building theories from case study research. In: A. M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion*. 5–35. Thousand Oaks, CA: Sage.

Goldstein, H. (2011). *Multilevel statistical models*. 4th edition. Chichester, UK: Wiley.

Hawkey, R. (2004). An IELTS Impact Study: implementation and some early findings. *Research Notes*, Issue 15, February 2004. University of Cambridge ESOL Examinations.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for Predicting First-Year College Grade Point Average*. Research Report, No. 2008–5. New York: College Board.

Lawrence, I. M., Rigol, G. W., Van Essen, T., & Jackson, C.A. (2003). *A Historical Perspective on the Content of the SAT*. Research Report No. 2003–3. New York: The College Board.

Lenning, O. T. (1975). *Predictive validity of the ACT tests at selective colleges*. Report No. 69 [050269000]. Iowa City, IA: American College Testing.

Noble, J. P. (1991). *Predicting college grades from ACT assessment scores and high school course work and grade information*. Report No. 91–3 [50291930]. Iowa City, IA: American College Testing.

Noble, J. P. & Sawyer, R. (1987). *Predicting grades in specific college freshman courses from ACT test scores and self-reported high school grades*. Report No. 87–20 [050287200]. Iowa City, IA: American College Testing.

Rasbash, J., Browne, W. J., Healy, M., Cameron, B., & Charlton, C. (2005). MLwiN Version 2.02. Centre for Multilevel Modelling, University of Bristol.

Robbins, S., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology*, **98**, 598–616.

Roy, S. (1998). A general framework for evaluating educational programmes. In: V. McKay & C. Treffgarne (Eds.). *Evaluating Impact*. 69–74. London: Department for International Development.

Sharp, J. (2009). *Success with your education research project*. Exeter: Learning Matters.

Stake, R. E. (2008). Qualitative Case Studies. In: N. K. Denzin & Y. S. Lincoln (Eds.). *Strategies of qualitative inquiry*. 3rd edition, 1–43. Thousand Oaks, CA: Sage.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Weiss, C. (1998). *Evaluation*. New Jersey: Prentice Hall.

Yin, R. K. (2006). Case study methods. In: J. L. Green, G. Camilli, P. B. Emore, A. Skukauskaite, & E. Grace (Eds.). *Handbook of complementary methods in education research*. 111–122. Washington, DC: American Educational Research Association/Lawrence Erlbaum.

APPENDIX: Comparison of secondary education in the UK and the US

	UK				*USA*			
Age	*Type of Institution*	*Year*	*Main Examination*	*Comments*	*Type of Institution*	*Grade*	*Main subjects/ examination*	*Comments*
14–15	SCHOOL	10		First year of GCSE course	HIGH SCHOOL	9	5 core subjects plus electives	• Students gain a Diploma in G12 • Credits for core and elective studies • Minimum number of credits needed; in Florida 24 • Many G11/12 pupils on Advanced Placement (AP) or Dual Enrolment (DE) as partof the credits • SAT taken in G11 and again in G12 if not good enough
15–16	“	11	GCSE (6–11 subjects)	Vocational courses also possible	“	10	5 core subjects plus electives	
16–17	SIXTH FORM or COLLEGE	12	AS (4–5 subjects)	Entry based on good grades in 4/5+ GCSEs	“	11	5 core subjects plus electives	
17–18	“	13	A2 (3 subjects)	The ‘best’ three AS subjects	“	12	3 core subjects plus electives	
18–19	UNIVERSITY	FIRST	First Year	Entry based on AS/A2 grades or points equivalent	COLLEGE	FRESHMAN	LIBERAL STUDIES	• Entry based on High School grades converted into GPA plus SAT score (plus in Florida community service) • They apply before receiving their Diploma • Offer based on minimum GPA + SAT scores in G12 • c.20% of students go to college
19–20	“	SECOND		“	“	SOPHOMORE	ASSOCIATE DEGREE	
20–21	“	THIRD	BACHELOR DEGREE	“	“	JUNIOR		
21–22	“	ONE	POST GRADUATE	Entry based on good first degree	“	SENIOR	BACHELOR DEGREE	