Preprint

The final version of this article was published as Simpson, O. (2006) 'Predicting Student Success' Open Learning 21(2) pp125-138

PREDICTING STUDENT SUCCESS
IN OPEN AND DISTANCE LEARNING
2006

Ormond Simpson

The Open University, UK

Abstract

This paper reviews some of the ways in which student success can be predicted in conventional and distance education. Predicting such success is particularly important for new students where the pre-course start information available is sometimes slight and withdrawal often occurs very early in a course. It suggests that in such cases statistical methods involving logistic regression analysis are the most useful rather than questionnaires or tutors' opinions. Identifying students with low probability of success allows support to be targeted on them. However there are ethical dilemmas to do with targeting support and openness with students about the results of any analysis.

Introduction

In any higher education system it will be important to attempt to predict the chances of any new student's success. There are a number of reasons for this – in an entry system determined by entry qualifications the more accurate the forecast of students' potential the fewer students selected who will go on to eventually drop out – the 'false positives'. Such false positives represent a waste of resource for both the institution and the students themselves. But equally important for social justice and the long term value to a country the more accurate the forecast the fewer potential students excluded who could have succeeded had they been admitted. In a recent UK study for example conducted by the Sutton Trust in lower-achieving state schools about 5% of students tested given the US SAT scored high enough for entry into a top US university yet almost all failed to get a UK 'A-

level' qualification good enough for entry to any leading UK university (Times Higher Education Supplement, 5 November 2004, p2)

In an open entry system the considerations may be a little different. Since there are no entry qualifications students are in a sense studying at their own risk and pay the price (in terms of their own tuition fees, time expenditure and perhaps diminished self-esteem) if they fail. However the institution itself is not without interest in being able to predict its new students' chances of success: I have suggested elsewhere (Simpson, 2004) that an institution may pay a high price for student dropout in reductions in government grant (a certainty in the UK and increasingly a possibility in the US and elsewhere), as well as loss of student fee income and increased expenditure on recruitment to replace withdrawn students. If it was possible to identify students with a low chance of success then it might be possible to focus support on such students and produce an increase in retention rates as a result.

There is also an argument in any open entry educational system that a policy of pure caveat emptor 'let the buyer beware' is unethical. In an unrestricted market it is perfectly legal to sell someone a purchase knowing that whilst it is fit for general purposes it may not be fit for their particular purpose. But the case of the salesman who persuades someone to buy an expensive high specification pc knowing that they only need a cheap machine for email and word processing may be uncomfortably close to the open entry institution which knowingly allows students to register for courses on which they may have only a low chance of success.

Administrative and psychological methods of predicting student success.

Consequently there have been a number of attempts to predict students' chances of success. Some of the methods used have been very simple - for example Wright et al (2002) gave 393 new medical students a simple administrative task – to provide a passport photograph. They found that of the 93% of students who complied 8% went on to fail their first year exams. Of the 7% who did not provide the photograph 48% went on to fail. Other more sophisticated methods have used socio-psychological tests of various kinds such as Shin (2003) who used a questionnaire to measure a student's 'Transactional Presence' – the psychological distance between a student, the institution, tutors and other students. Yet other methods focus on the student's motivation and there have been successful attempts to show links between students' success and their motivation as measured by various tests for different models of

motivation such as 'Self-Efficacy theory', 'Achievement Goal theory' 'Interest Development model' and so on.

Such methods are clearly important for the insights they offer into students' intellectual and emotional frameworks but may not necessarily offer clear guidance as to how such frameworks may be adapted to improve their success. In addition there is the possibility of some lack of clarity as to what is really being measured and some tautology in the result. For example Correia et al (2004) showed that for full time students there was a clear link between their 'Belief in a Just World' (a psychology concept measured by a standardised test) and their educational success. But 'Belief in a Just World' may itself be not a fundamental measure and may just reflect some other more basic aspect of the student's emotional framework. And clearly a student who is getting good grades may well have their 'Belief in a Just World' fostered as a result of getting those grades.

Finally the accuracy of questionnaires in predicting success may not be that great. DeTure (2004) found that scores on cognitive style and self-efficacy were poor predictors of student success in online distance education and Bernard et al (2004) found that whilst it was possible to develop questionnaires to forecast student success, the student's previous course grade was usually a better predictor than any questionnaire.

Faculty opinions in predicting student success.

An early example of using faculty opinions to predict student success in a distance education context was reported by Thorpe (1988) who asked UK Open University tutors (adjunct faculty) to identify students they thought were 'at risk' of failure on initial entry courses. Tutors were then funded to offer extra support to such students, generally in the form of one or two hours extra tuition. The identification of such students was left to tutors largely on the basis of their experience although it subsequently became clear that the main criterion they were using were students' previous educational qualifications. An analysis of the results based on the previous educational qualification (peq) of the students showed that predictions based on peq were reasonably accurate and that interventions based on those predictions had a marked effect on student success. There were clear differences between students who had received extra support and those who had not – see Figure 1.

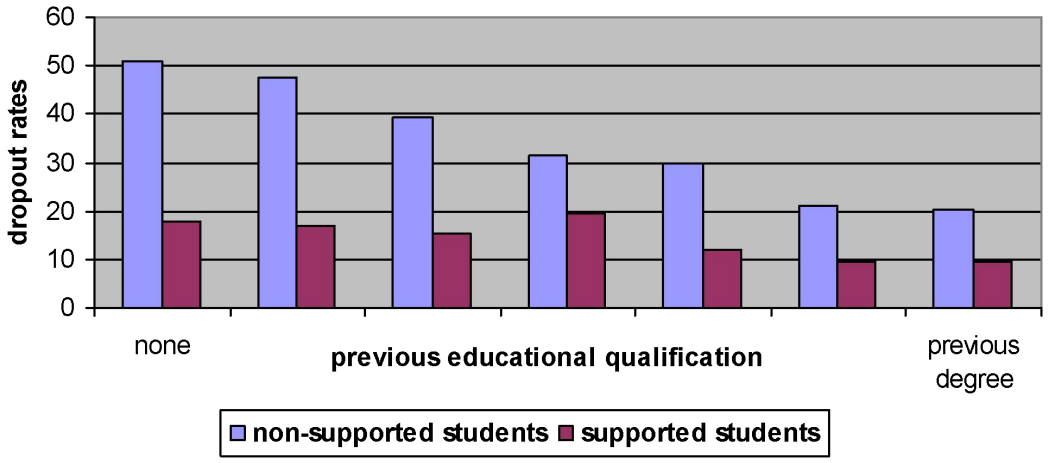

Figure 1 Dropout rates versus previous educational qualifications for first year UKOU students

The results showed a clear increase in student success amongst students targeted for extra support and the project was continued for a number of years. Nevertheless institutional inertia and lack of attention to student retention meant that the lessons learnt from this exercise were largely forgotten until recently, when UK Government funding regimes meant that there was higher awareness of student retention issues. However it seemed that using tutors to identify 'at risk' students might not be the most effective method for identifying vulnerable students on the grounds that:

i. dropout was a 'multi-factorial phenomenon' (Woodley, 1987) that involved many more factors than a student's peq,

ii. and that in any case such identification could often only take place after the start of the course when remedial action might well be too late.

Thus it became important to investigate more sophisticated methods.

Statistical methods of predicting student success.

Statistical methods of predicting student success have been in use for a number of years in conventional higher education. The Noel-Levitz Corporation in the US (www.noel-levitz.com) offers a service which will identify new students' chance of withdrawal based on a logistic regression analysis of previous students' records. Such an analysis uses previous students' known entry characteristics (such as sex, previous educational qualifications, age and so on) to relate to their subsequent success. Clearly some of these factors are likely to have a

much greater effect on a student's chances of success than others so the analysis produces an algorithm weighted for different factors. This algorithm can be applied to new students entering the institution to predict their chances of success in turn, assuming that the dependency of success on the various factors does not chance substantially from year to year.

In the UK a logistic regression model was used by Napier University in Edinburgh for its full time students (Johnston 2002). Data was collected by means of a questionnaire shortly after the new students' arrival. This was linked to their eventual success and a logistic regression analysis undertaken and a predictive model formulated. This was then used to design a second questionnaire which was given to new students in following years. The analysis of this questionnaire was conducted by students themselves and the results used in an interview with the student's tutor. Examples of some of the questions are shown below in Figure 2 together with an example of the scoring system used (there are 14 questions in all).

Score	**Questions**
	1. How old were you at the beginning of October? 18 years or less 19 to 23 years 24 or more years
	2. If you have 'Highers' (a Scottish educational qualification), how many do you have? (If you have both Highers and 'A' Levels then calculate 1 'A' Level = 2 Highers and select the nearest category below) 1-2 3 4-5 6 or more
	3. If you have 'A' Levels (and no Highers), how many A Levels do you have? 1 2 3 4 or more
	4. What type of accommodation do you stay in? At home Napier-owned accommodation Private accommodation sharing with other students only Other
	5. If you have a job during term-time, for how many hours are you normally employed each week?

	None	1-10 hrs	11-15 hrs	16+hrs

Score Sheet for Questionnaire

1. Age group

	Points
18 years or less	0
19 to 23 years	2
24 or more years	11

2. No of Highers (or Highers and A Levels)

	Points
1-2	2
3	7
4-5	8
6 or more	14

3. Number of 'A' Levels (no Highers)

	Points
1	2
2	7
3	8
4 or more	14

4. Accommodation

	Points
At home	0
Napier	7
Private	0
Other	4

5. Hrs of employment

	Points
None	8
1-10 hrs	11
11-15 hrs	6
16+hrs	0

Figure 2 The Napier 'At risk' Diagnostic Test

The aim of the interview was to help the student identify areas where changes might be necessary in order to increase their chances of retention. For example the logistic regression analysis showed a link between dropping out and taking paid work whilst being a student - students working more than 15 hours a week had a higher chance of dropping out. Thus it could be suggested to such students that at the least they should reduce their working hours.

Interestingly one of the functions of the analysis is that it can also show up unexpected correlations. In the example above it appeared from the analysis that students working between 1-10 hours/week had the highest chance of success; students not working at all or working 11-15 hours/week actually had a lower chances of success. This finding remains to be explained. It may be that there is another underlying characteristic which affects working hours as well as chances of success.

Predicting success in distance education

It was clear that such a process could be applied to distance education students and that the large numbers of students involved in the UKOU (35,000 new students each year)

might make the process statistically reliable there. An early approach was made by Lindsay (1977) who used logistic regression methods. His investigation explored a longitudinal description of student progress using factors such as age, sex and previous educational level and apparently found good agreement particularly using previous course performance as students moved through their studies. Unfortunately his article is very brief and does not include any statistical data. However he did compare his predictive model with predictions made by an admissions counsellor interviewing students, students themselves and their tutor after the tutor had known the students for three months. He found that the most accurate predictions were made by the model, followed by those made by tutors and the students themselves and that the counselling interview predictions were much less accurate.

Burt (1996) also used a mathematical model to predict the performance of student groups based on their previous success in the UKOU. He discovered that groups that had been successful in the past were more likely to continue and be successful. Ashby et al (2004) in a very detailed regression analysis of more than 400,000 new students using 18 items of data some gathered from new students after they had started (such as financial assistance application) found excellent agreement between the model's prediction and actual results.

Again all these models use data which is generally not known until well after course start or after the student has actually completed some study with the institution. If it desired to use such models to deal with student dropout then it has to be remembered that this is often very heavily front-loaded in distance education – for example nearly 40% of new UKOU students withdraw before the first assignment (Simpson, 2003). So for retention activities to have much effect they have to be targeted on students at the point of registration or very shortly thereafter. But relatively little may be known about new students at registration in a distance education operation – the cost of collecting data is a large factor in recruitment and registration expenditure so the amount of data acquired is restricted.

In addition some factors which must be important to student success are not taken into account such as student motivation and resilience and other personal characteristics (Simpson, 2002). It might have been possible to collect further characteristics by means of a questionnaire as used at Napier University but the cost and time of collection which would probably involve several stages of chasing up responses (some students register only a matter of one or two weeks before course start) made that seem prohibitive. In addition the

results from Wright et al (op cit) suggested that the most vulnerable students would probably be those least likely to return the questionnaire.

In the event the data available for analysis at new student registration in the UKOU is only

- Sex
- Age
- Previous education qualification
- Course choice
- Socio-economic status (inferred from occupational status)

Other factors are collected such as special needs and financial award status but often at a later stage, sometimes after course start. Since the aim of this project was to identify vulnerable new students as early as possible these statuses were excluded from the analysis but of course the fewer the data included in the analysis the less accurate the predictions.

The analysis was conducted by an Open University tutor who was also studying for an MSc and it was published as an MSc thesis (Woodman, 1999). It turned out that the most important factors linked to success for new students were (in order):

1. Their chosen course level. Students entering on level 1 (first year of degree equivalent) tended to have a higher success than students entering on level 2 (second year of degree equivalent) courses.
2. The credit rating of a course. Students entering on 15 credit point courses (equivalent to one eighth full time study) were more successful than student entering on 30 or 60 point courses
3. a student's previous education qualifications
4. their course choice (arts students were more likely to be successful than maths and science students for example),
5. their socioeconomic status (the higher the more successful)
6. sex (women more successful than men)
7. age (middleaged students more likely to be successful than younger or older students).

These factors were then used to predict the chances of success of the succeeding year's cohort of new students in 2002. In practice this was achieved by extracting the data of new students' characteristics into a specially designed Access™ database which contained an algorithm which then calculated a 'predicted probability of success' (pps) for each new student in the cohort.

Some of these factors made only relatively small difference in student progress. For example the difference in pass rates between men and women is only about 10 percentage points. But when all the factors are taken together the differences predicted can be much larger. In the first analysis undertaken using this method the predicted probability of success varied from 84% (for a well-qualified woman studying arts course with other positive characteristics) to 9% (for an unqualified man studying technology courses and other negative factors). The output data from the analysis was in the form of a spreadsheet that showed a predicted probability of success (expressed as %) for any particular student. An extract from that output is shown as Table 1 (the full spreadsheet contained predictions for nearly 3500 new students taking about 5300 courses between them). The initial letter of the course code denotes the faculty so M- are maths courses, S- are science courses , T- are technology courses and A- are arts courses.

Student	Sex	Course code	Predicted probability of success % (pps)
1	M	S281	9.4
2	M	M358	13.1
3	M	T171	13.6

- which ran through to

Student	Sex	Course code	Predicted probability of success %
5321	F	A103	82.1
5339	F	A103	84.4
5340	F	A103	84.4

Table 1 Extract from the spreadsheet of predicted probability of success.

The distribution for the predicted probability of success (pps) for the 5300 students-courses involved in this prediction is shown in Fig 3.

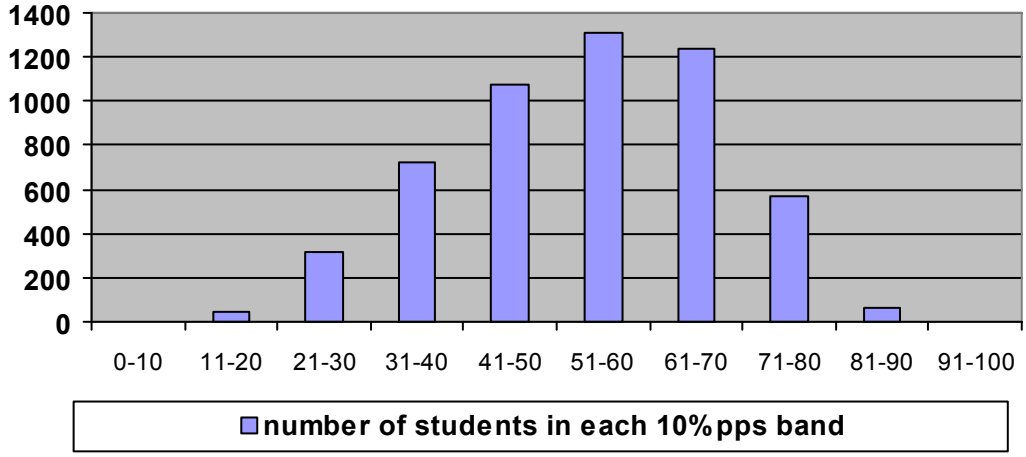

Fig 3 the number of students in each 10% 'predicted probability of success' bands

As might be expected the number of students in lower and higher bands is relatively small and most students cluster in the 30-80% bands.

Accuracy of predicted probability of success model.

The predictions were compared with the actual pass rates of the new student cohort at the end of 2002 in 5% pps bands – see Table 2.

predicted probability of success band %	Actual average pass rate of students in that band
5-10	0
10-15	0
15-20	20
20-25	23
25-30	29
30-35	31
35-40	34
40-45	38
45-50	50
50-55	53
55-60	57
60-65	62

65-70	67
70-75	72
75-80	76
80-85	80

Table 2 Comparison of predicted pass rates and actual pass rates in 5% bands (n=5287)

There is good agreement between predicted and actual values although for the lowest and highest bands the number of students involved is very small – there is only one student in the 5-10% pps band out of 3,500 for example.

Of course for any one student the critical factor is not their probability of passing but their actual result which is pass or fail. The overall pass rate is around 44% and if the pps distribution is divided at that point the prediction of pass/fail is correct in 65% of cases – i.e. for any one student with a pps above 56% a prediction of a pass will be correct in 65% of cases.

The courses analysed above are largely conventional distance education courses. But when it comes to online learning the same factors appear to apply with the addition of familiarity with computers. For example Dupin-Bryant (2004) in a study of pre-entry variables in online learning found that prior education and IT skills were the most important determinants of student retention.

Using the Predictive model

The stimulus for using the predictive model in the UKOU was the recognition that in order to increase its retention rates the university needed to undertake proactive contact with its new students rather than waiting for new students to contact it. It was realised that such contact had to be individual and targeted so that it was almost inevitably made through individual phone calls. However the cost of proactively contacting all the 35,000 new students each year was thought to be prohibitive so a way was sought of targeting new students so that interventions would produce the greatest effect.

It was assumed that proactive contact with students with a high predicted probability of success was unlikely to increase that pps by any substantial amount and that better effects would be produced by concentrating contact on students with lower pps's. Thus proactive contact was concentrated on students in the below 56% pps category. This

also accorded with the University's mission to support students from educationally disadvantaged backgrounds.

In the event as reported by Simpson (2004) there was an increase in retention of around 4-5% amongst the contacted students compared with a control group of identical pps's. This is similar to a result reported by Mager (2003) of a 5% increase in student retention using similar methods at Ohio State University with full-time students.

However the assumption that contact with higher pps students would have a lesser retention effect was not tested and it remains to be shown at what levels contact may be most effective. The evidence to date suggests that contact at very low pps levels does not make as much difference as contact in the middle of the range but this is based on very small student numbers. It does however raise various ethical issues.

Ethical issues in using the predictive model.

There are a number of potentially sensitive issues to do with acting on the data supplied by the predictive model.

- Limited accuracy. Using that data to target resources on students is open to the criticism that the data is limited in accuracy. This is particularly true as a substantial proportion of new students do not give full personal data. It is occasionally the case that an apparently vulnerable new student was phoned only for the adviser to discover that – for example – the student has a high level qualification that they did not enter on the form for some reason and which would have given them a much higher pps had it been known.
- Arbitrary limits. The cut off point in pps value above which students are not contacted is usually set quite arbitrarily by the resources available. Thus a student with a pps value of 54% may be contacted but a student with a pps value of 55% may fall outside the contact zone.
- Targeting favours certain groups. Inevitably targeting means that certain groups within the population receive disproportional levels of support. Thus men who are predominantly represented in the lower half of the list are likely to receive more support than women.
- Holding data . There are also ethical issues around the holding of such data. At the moment the pps data is held on various separate and informal databases throughout different parts of the UKOU and although students are entitled to ask for that data under the British Data Protection Act it is unlikely that they know that it exists.

However the University is shortly to introduce a Customer Relationship Management (CRM) system and it is likely that a pps will form part of a student's record on that system. It is much more likely that students will then become aware of that data and ask to see it.

- Being upfront with students. Indeed there are arguments that say the data should be made available to students in some way. Should a university allow a student with apparently only a 20% or less chance of success to embark on a course without some kind of strong warning that that is the case? On the other hand just to contact such a student with that information could be irritating for someone who had just omitted pertinent data which vitiated the prediction or very demoralising for someone who through sheer motivation would have overcome their disadvantages and succeeded.

One possible route through this ethical maze would be to provide the data but mediated through an appropriate person such as the tutor in the Napier model. This is likely to be difficult in a mass education system such as the OU for cost reasons but it may be possible to encourage students to undertake a 'self-assessment' using a questionnaire similar to the Napier but designed to be self-assessed by the student. An example developed by the author is shown below. However this has not yet been developmentally tested.

DRAFT

HOW GOOD ARE YOUR CHANCES OF PASSING?

Everyone who starts with the OU has a chance of succeeding. Of course you'll need commitment, time and energy. And a sense of humour will help!

There are also factors in your background which we know may affect your performance in your first year. This questionnaire is designed to help you
- become aware of the factors which may affect your performance
- to identify factors which might apply to you particularly
- to point to actions which you might be able to take on some of the factors to improve your chances of success.

Start with a score of 60 points. Answer each question in turn and add or subtract a point score as you go along.

	Initial Score : 60 points
1. Are you male or female? Male : subtract 5 Female: No change	Revised Score: points
2. How old are you? Under 30 : Subtract 13 Age 30 or above : No change	Revised Score: points
3. What level is this course? Level 1: Add 23 Level 2 : Add 11 Other: No change	Revised Score: points

4. What Faculty is this course? A : Add 16 D or L: Add 8 E or K: Add 7 M : Add 6 S : Subtract 3 T : Add 1 Other: No change	Revised Score: points
5. What is the credit rating of this course? 15pts : Subtract 23 30pts : Subtract 9 60pts : No change	Revised Score: points
6. How many courses are you taking in total this year? 1 course : Add 5 2 or more courses : No change	Revised Score: points
7. What are your current highest educational qualifications? Degree or equivalent : Add 17 A-level or equivalent : Add 12 O level, GCSE or equivalent : No change None to CSE : Subtract 21 Other : No change	Revised Score: points
8. How would you classify your occupation? Working- professional occupation : Add 10 Working- other occupation : Add 5 Not working or other: No change	Revised Score: points
	Final Score : Points

How did you score?

100 or above: The outlook is very bright for you. You'll undoubtedly have your share of challenges but you should be able to get things off to a good start. But you should still make sure that you are on the right course for you and have the right knowledge background.

75 to 99: This will be a challenge you've taken on and it will be useful to see if you can reduce your point score in some way. You may not want to change sex (!) but you could change your course to a lower level or amount, increase your current educational qualifications by taking a short course of some kind – the 'Openings' courses are ideal – and so on.

Under 75: You'll still be able to succeed but if you can increase your score that would really improve your chances. For example do think about changing to a lower level course just for the first year – you can step up the pace later on. If you are taking more than one course then again do think of switching to just one or a fewer number of points. But remember that determination can overcome everything so don't be put off studying by your score.

Benchmarking

One side-effect of the predictive model is that it may contribute towards the evaluation of the student experience in other ways. For example there is substantial variance between UKOU courses in terms of their pass rates. Part of this variation may be due to different courses attracting different populations of students. Using the predictive model may allow 'benchmark' pass rates to be calculated for courses – the pass rates they should have for their particular student intake – and allow for more accurate comparisons. A similar process might be possible to settle the question of whether attendance at face to face tutorials assists student progress. However benchmarking remains to be evaluated as a method in the UKOU.

Conclusions

The predictive modelling of student success can be sufficiently accurate to be worth using for targeting support onto vulnerable students. There is some limited evidence of the effectiveness of this approach particularly if statistical methods rather than questionnaires are used. However the model also raises ethical issues about the use of the data notably whether and how such predictions can be shared with students. But the model may also be useful for setting benchmarks in the evaluation of courses and student support.

Ashby, A. Slee, A. and Moss, G.(2004) *'Progress report from the project team on the IET Modelling Student Data Project'* Internal report to the UKOU Retention-Next Team

Bernard, R.M. Brauer, A. Abrami, P C. Surkes, M. (2004) 'The development of a questionnaire for predicting online learning achievement' Distance Education **25** (1) pp31-47

Burt, G. (1996) 'Success, confidence and rationality in student progress' *Open Learning* **11** (3) pp 31-37

Correia, I and Dalbert, C. (2004) 'Belief in a Just World as a personal resource for university students' - paper presented at the International Conference on Motivation - Cognition, Motivation, and Effect, Lisbon.

DeTure, M. (2004) 'Cognitive style and self –efficacy: predicting student success in online distance education' American Journal of Distance Education 18 (1) pp21-38

Dupin-Bryant, P. (2004) 'Pre-entry variables Related to Retention in Online Distance Education', *American Journal of Distance Education* **18** (4)

Lindsay, B. (1977) 'The prediction of the academic performance of Open University students' *Teaching at a Distance* 10 pp 49-50

Johnston, V. (2002) Presentation at the conference *'Holistic student support'* University of Central Lancashire, Preston.

Mager, J. (2003) Report at the Noel-Levitz National Student Retention Conference San Diego

Shin, N. 'Transactional Presence as a critical predictor of success in distance learning' Distance Education **24** 1 pp70-86

Simpson, O. (2004) 'The impact on retention of interventions to support distance students': *Open Learning*, **19** (1)

Simpson, O. (2003) 'Student Retention in Online, Open and Distance Learning' RoutledgeFalmer, London

Simpson, O. (2002) 'Supporting Students in Online, Open and Distance Learning' RoutledgeFalmer, London

Thorpe, M (1988) *Evaluating Open and Distance learning*, Longman, Harlow, UK

Woodley, A (1987) Understanding adult student drop-out, in *Open Learning for Adults*, ed M Thorpe and D Grugeon, pp 110–124, Longman Open Learning, Harlow, UK

Woodman, R. (1999) *'Investigation of Factors that influence Student Retention and Success Rate on Open University courses in the East Anglia Region'*, dissertation submitted to Sheffield Hallam University for the degree of MSc in Applied Statistics.

Wright, N. and Tanner, M.S. (2002) Medical student compliance with simple administrative tasks and success in final exams – a retrospective cohort study, *British Medical Journal* 7353, 29 June 2002 pp1554-1555.