

Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment

Annika Wolff, Zdenek Zdrahal
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA
+44 (0) 1908 659462
{Annika.wolff,
zdenek.zdrahal}@open.ac.uk

Andriy Nikolov
fluid Operations AG¹
Altrottstraße 31
69190 Walldorf, Germany
+49 (0) 6227 3849-567
andriy.nikolov@fluidops.com

Michal Pantucek
Czech Technical University¹
Dept. of Cybernetics
Karlovo namesti 13, 12135 Praha 2
+42 (0) 22435 7666
pantumic@fel.cvut.cz

ABSTRACT

One of the key interests for learning analytics is how it can be used to improve retention. This paper focuses on work conducted at the Open University (OU) into predicting students who are at risk of failing their module. The Open University is one of the worlds largest distance learning institutions. Since tutors do not interact face to face with students, it can be difficult for tutors to identify and respond to students who are struggling in time to try to resolve the difficulty. Predictive models have been developed and tested using historic Virtual Learning Environment (VLE) activity data combined with other data sources, for three OU modules. This has revealed that it is possible to predict student failure by looking for changes in user's activity in the VLE, when compared against their own previous behaviour, or that of students who can be categorised as having similar learning behaviour. More focused analysis of these modules applying the GUHA (General Unary Hypothesis Automaton) method of data analysis has also yielded some early promising results for creating accurate hypothesis about students who fail.

Categories and Subject Descriptors

H.2.8 [Database Applications]: Data Mining; D.4.8 [Performance]: Modelling and prediction

General Terms

Algorithms, Experimentation.

Keywords

Predictive modelling, retention, student data, virtual learning environment, distance learning

1. INTRODUCTION

In learning analytics, student data is analysed in order to provide insight into what students are doing with learning materials. Some applications of learning analytics feed back directly to students about their own behaviour in order to help them to become more strategic learners.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

LAK '13, April 08 - 12 2013, Leuven, Belgium
Copyright 2013 ACM 978-1-4503-1785-6/13/04...\$15.00.

Others inform tutors or module design teams about student behaviour, to better facilitate student support. A key aim of the latter is to improve student retention. This paper will discuss work conducted by the Open University using data from the Virtual Learning Environment (VLE) combined with other data sources to predict students at risk of failing a module. This work suggests that changes in students activity on the VLE is a reliable indicator of failure. Additional, ongoing, work will also be discussed which in early stages seems to show that it is possible to use the GUHA method (using LISp Miner) with the VLE data to generate hypotheses in the form of rules about factors contributing to student failure. These rules can then be applied to new data sets and predict accurately which students will fail their course.

2. RETENTION AND ONLINE LEARNING

Studies have shown that online courses have larger attrition rates than traditional bricks and mortar establishments [4]. Investigations into the differences between 'click' and 'brick' establishments suggest several contributory factors, but with conflicting results among different studies as to which have the biggest impact on student drop out [3, 4, 14]. Possible factors include a difference in the student demographics: there are usually more students from lower economic backgrounds, those with less formal qualifications and a higher proportion of disabled students. Also, there may be difficulties with the technologies needed to study, greater time constraints and less academic 'preparedness' for study. Also, due to lack of face to face contact, students can feel isolated and unsupported by their tutors. On this latter point, a study by Frankola [5] reveals this to be a major contributing factor to student drop out.

Previous research has indicated that initiating telephone contact with students can improve retention figures [15, 16]. The difficulty faced by large distance learning institutions is that to resource this contact on a large scale is financially unviable. What is needed is some intelligence as to which students will most benefit from an intervention, to allow resources to be targeted more effectively [2].

2.1 Student tracking

A 'broad-brush' approach for identifying the students who might benefit from more focused support, particularly at the start of a module, is to use a tracking system [11]. This can flag up those students who fit the sorts of profile shown to be more likely to

¹ This work was carried out at the Knowledge Media Institute

fail. However, since these precise factors may vary by institution, it is advisable for each institution to undertake some analysis of their own data to determine which factors are most informative in their specific case [6]. A further issue is that, whilst these students are more at risk than others, the vast majority of them will go on to be successful, while others who don't fit that profile will fail. Therefore, these factors can be more usefully integrated with other data sources that reveal something of the students' behaviour and performance on the module. This can include, for example, data from Virtual Learning Environments, or other static data sources such as assessment submission patterns and outcomes.

Integrating these data sources increases the burden on the tutor in terms of analysing and interpret the tracking statistics. Visualisations can go a long way towards helping tutors to manage the data load (e.g. 11), but the more data that is involved, then the harder it is for tutors to know what to look for or to uncover the interesting patterns that might tell them who is struggling, or even why. Factors that might on the surface seem a good predictor of failure may become unreliable when the myriad of other factors involved in undertaking an online learning course come into play.

2.2 Predictive modelling

One possible solution is to build predictive models [10]. Computational methods can identify consistent patterns in learner behaviour that are hard for the human tutors to perceive, but which can be used to accurately predict what they will do next (in this case, either pass or fail the course). Course Signals [1] visualizes predicted performance to both tutors and students themselves. The model is built on several data sources, including engagement with resources on the learning management system, as well as assessment performance, past history and demographics. The authors demonstrate the effectiveness of the system for improving retention. Smith and Sweeley [17] used data such as frequency of logins to learning system and some interaction with course materials (as determined by clicking) to try to predict student failure, weighting the model to look most at recent behaviour. They found a significant correlation between logging in, looking at materials and course outcome. The developed model had 70% accuracy in predicting unsuccessful students.

These predictive methods build models of 'average' learners against which individual students are compared. But distance learners are a diverse group, which means it is not possible to build a typical student profile [18]. Pistilli and Arnold suggest that the most effective models are built on a course by courses basis [12]. This is because most courses are structured differently and place different demands on learners. This is especially true if looking at activity data, where the course design will, to a large extent, dictate what learners ought to be doing in terms of engaging with learning materials, or chatting on forums etc.

3. BUILDING THE PREDICTIVE MODELS FROM OU DATA

In common with other learning institutions, the Open University is interested in using analytic techniques as part of an approach to improving the student experience and increasing retention figures. As the OU is a distance learning institution, most of the knowledge about students is held in various data sources. Crucially, this includes knowledge about their learning activity in the form of their engagement with the Virtual Learning Environment (VLE). Other static data sources contain information on students' assessment results, as well as other demographic

data, financing, disability flags etc. The aim was to develop models for predicting student failure using data from the VLE in combination with the assessment data: each OU module has a number of Tutor marked Assessments (TMA's) as well as a final exam. These contribute to the overall pass mark.

3.1 Understanding the modules

The models were developed and tested on historic data sets from three modules, which were chosen for having large student numbers and for making good use of the VLE for delivering the course content. The modules were from the different subject areas of art, mathematics and business. The student numbers and VLE accesses are in Table 1.

Table 1. Profiles of the three selected modules for developing and testing models

Module ID	No. of students	No. of VLE clicks
A	4397	1570402
B	1292	2750432
C	2012	1218327

These modules were also considered to be fairly typical of OU courses, and more importantly to reflect how modules will increasingly be delivered in the future. However, even so it is necessary to note that all modules were significantly different in some important respects. Each module can set different criteria for passing the module. There can be different numbers of TMA's, some of which may be optional or substitutable with other TMA's. Therefore, predictive models were developed on a module by module basis to determine whether or not it is important to take these differences into account.

3.2 Understanding the data

The VLE clicks are recorded by category, such as whether they were clicks for a module page, or clicks on a forum or discussion topic. The level of detail is relatively coarse-grained. Analysis was undertaken to understand whether or not it was necessary to filter out some categories of data, if they were not directly related to learning but may interfere with the predictive results. This analysis revealed that there were two categories of data that might be filtered. One category was the home-page for the module, that the user always visited before proceeding to look at further materials. However, since this page is accessed uniformly, it does not affect the quality of the data. Conversely, the other pages (such as updating personal details) were accessed so infrequently they had a similar non-effect. The remaining data categories related to forums, course content and online tools.

It was also necessary to spend some time ensuring that not only was each category of data was understood but also whether or not it was used consistently from one module to the next. While the issue of data cleaning for all data within the OU was not resolved, it was possible to gain enough knowledge about the data from the three selected modules to start building models.

3.3 Understanding the students

The overall aim is to be able to understand the students behaviour by looking at their data alone and to therefore predict struggling students, even when the students don't say anything themselves. In order to do this, it is important to understand how students would be using the VLE and how this usage would affect their final performance. Student's VLE activity was analysed on the three modules, firstly by looking at all clicks, and then by breaking it down by the VLE data categories. The first general finding was that student activity, on average, increases in the

week that an assessment is due. This finding is so marked that it is possible, without prior knowledge, to pinpoint the date of each assessment on each of the modules. This information is not usable in itself to predict failure of a TMA, since the student can access the VLE only at the last minute and still pass the assessment. Instead, it suggests the need to identify ‘activity’ time periods between assessments for analysing the students VLE accesses. A second important finding was that there is no such thing as an average student. This is consistent with the view of Thompson [17]. There were students who clicked a lot and still failed, or those who clicked hardly (if) at all and yet passed. These students may have a printed version of course materials, or may have been retaking the module and therefore required less VLE access. This was true of all modules. It was therefore not possible to find any general measure of clicking for a given module that students could be compared against.

4. PREDICTING PERFORMANCE DROP AND FINAL OUTCOME

The next stage involved developing classifiers to predict risk, which is defined as either:

- performance drop – predicting a previously well-performing student will fall below the pass threshold in the next activity period.
- final outcome – predicting whether a student will pass or fail the course. This was tested at different time-periods in the course, namely the different assessment submission points.

Based on the previous analysis of students clicking behaviour, the data from students who did not engage with the VLE was filtered out. Training and testing were performed with the historical data from three modules using 10-fold cross-validation (whereby the data is divided into 10 random samples, 9 of which are used to train the model and the remaining 1 is used to test it). Testing was done using VLE data only, TMA data only and a combination of both. The VLE data was not processed to exclude any categories of data (based on the findings discussed in section 3.2). For both types of models, decision trees were found to outperform SVM’s (state vector machines). The decision tree algorithm used was C4.5, a version of Quinlan’s ID3 [13].

4.1 Performance drop

The features of the performance drop model included the students’ assessment scores and the number of VLE clicks in a time window k (a period between TMA submission). The feature to predict was the nominal class label (“drop”/“no-drop”). As the goal is primarily to recognise the “at risk” students, the results are measured using the standard precision (p), recall (r), and F1 measures for the class “drop”¹. With the window size 3, the performance drop classifier was able to achieve high precision for

¹ The precision is defined as $p = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$ and recall as $r = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$, where “true positives” are the instances correctly recognised as belonging to the class “drop”, “false positives” are the cases where the drop was expected, but the student actually performed well, and “false negatives” are the cases where the student’s performance was not expected to drop, but actually dropped. Then, the combined F1 measure is defined as the combination of the two: $F1 = \frac{2pr}{p+r}$.

all three courses (between 0.77 and 0.98) and good overall accuracy (F-measure between 0.61 and 0.94), on the VLE+TMA data combination. These results can be seen in Figure 1. Interestingly, the number of VLE clicks occurring just before the TMA being predicted was found to be the most informative attribute: a student who used to work with the VLE before but then stopped is likely to fail at the next TMA. Thus, even the time window size $k = 1$ was sufficient to build the model, while increasing it could only lead to marginal increase in performance. TMA data on its own was not found to be very good at predicting performance drop. VLE data on its own was marginally better. The best prediction occurred when these two data sources were combined.

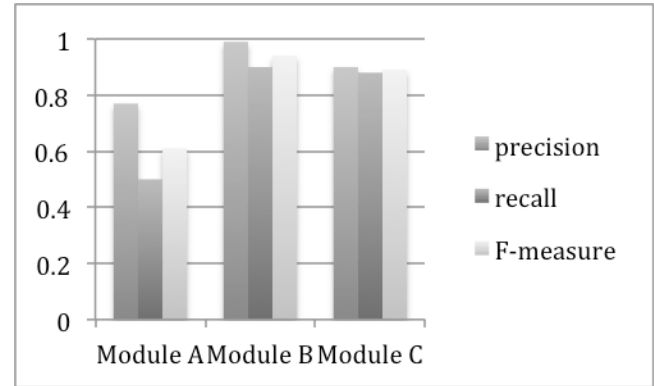


Figure 1. Predicting performance drop, using VLE and TMA data, with a window size $k = 3$.

4.2 Final outcome

The final outcome prediction model uses, as features, the scores for TMAs, the average TMA score, and the number of VLE clicks in periods between submission dates of each two subsequent TMAs. The findings from running the model on the available data was that precision is reduced as the module progresses, in other words it is easier to predict failing students in the early stages of a module. VLE clicks were again found to be better for prediction than the assessment scores. Again, VLE and TMA data combined are generally better for prediction, especially when compared to using only assessment data. See Figure 2 for an example using Module A data.

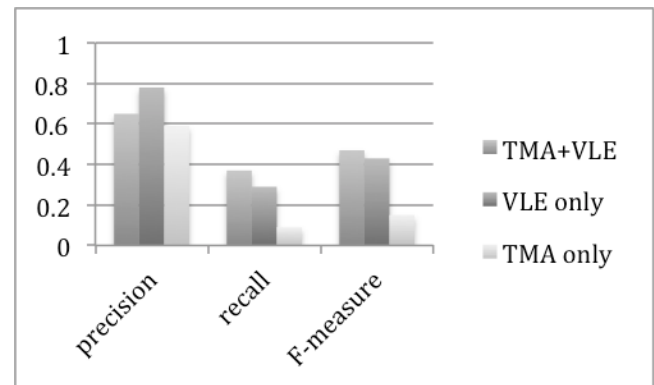


Figure 2. Predicting final outcome at TMA 3 for Module A - comparing TMA, VLE and combined.

4.3 Adding demographic data to module A

Demographic data was added to see if this would improve prediction of final outcome. The type of data to include was chosen in consultation with the student statistics team who have

used static data sources previously for prediction, though not combined with the VLE data. The demographic data was added to module A. It was found that this data did indeed improve prediction (see table 2). The selected demographic feature is not reported, as demographic data can be considered sensitive. However, it is likely that the chosen feature is specific to the Open University, since the demographics invariably vary between institutions. For example, the Open University solely offers distance learning and has a higher percentage of mature and disabled students than other institutions. It seems reasonable to suggest therefore that selecting which demographic data to use needs to be done on a case by case basis.

The importance of VLE, demographic or TMA data for prediction depends on the point in the module at which the prediction is being made. The data suggested that in the early stages the VLE data is best for prediction, then the demographic data and finally the assessment scores. However after the third assessment, the assessment scores become more informative. This seems reasonable, since the final result depends in some way on the assessment performance. Another possible explanation is that the students who drop out due to lack of motivation/difficulties, as evidenced by their VLE activity, tend to do so earlier in the module. Those who drop out later may do so for more unpredictable reasons, such as sudden personal problems.

Table 2. Module A with additional demographic data

TMA number	with demographics			without demographics		
	p	r	F1	p	r	F1
2	0.62	0.23	0.34	0.73	0.21	0.32
3	0.70	0.37	0.49	0.65	0.37	0.47
4	0.7	0.35	0.47	0.74	0.33	0.46

4.4 Applying models across modules

In a final test of using decision tree classifiers for prediction, the trained models were applied to different data sets. Unsurprisingly, the models generally proved best when applied to the same module. However, the quality on the ‘transferred’ models were still of reasonable quality and even in some cases better than on the original course it was trained on. These results suggest that there is a strong module-independent pattern of activity when looking at general student behaviour in terms of VLE activity. However, they do not give much precision for later stages of student drop out, nor do they provide module specific information for informing managers of why students might be struggling.

5. HYPOTHESIS GENERATION

The next step has involved trying to generate hypotheses that can provide more detailed explanations for student failure, since the explanations produced through the decision trees are fairly generalized. This work, in early stages but yielding promising results, has used LISp Miner to implement the GUHA (General Unary Hypothesis Automaton) method of data analysis to produce hypotheses about failing students in the form of a set of association rules [8, 9]. The GUHA method is a data mining technique for generating as many plausible hypotheses as possible from the data, in accordance with initial set up parameters for confidence (the probability that a generated hypothesis correctly classifies the cases), support (a minimum percentage of examples that fit the rule) and maximum number of antecedents.

In the first stage, previous findings from the decision tree were replicated when it was confirmed that it was not possible to get

useful results by looking at the VLE data in terms of total number of clicks. Instead, for each student, the attribute used for building the hypotheses was the *change* in the users VLE activity compared to a previous period activity, expressed as a percentage.

The TMA performance was also categorised into bands of <40% (the failure threshold), 40-60%, 60-80% and 80-100%. GUHA was trained to predict module failure (rather than performance drop).

Parameters were set to improve performance, e.g. setting the maximum number of antecedents to 5. The confidence was set to 70% (i.e. only generate rules that have a confidence of 70% or greater) so that it would not generate hypotheses that had equal chance of being invalid as they did of being valid. A suitable support value is dependent on the individual data set. In our case, because of the relatively low number of cases to classify in the large dataset (only a small proportion of students will fail) the support needed to be set very low in order for GUHA to find hypotheses (in some cases as low as 0.001).

GUHA produced a set of rules which, when applied to the same module data from the subsequent year (the model developed association rules from 2010 data which was then applied to 2011 data) produced extremely accurate results. As an example, shown in Table 3, GUHA has produced the following finding: a fail in TMA 4 can predict failure with 88% confidence in 2010 (537 cases) and 83% confidence in 2011 (516 cases). GUHA produced a more specific version of this rule that included the change in VLE activity between TMA’s 6 and 7. This improved performance to 94% confidence in both 2010 (472 cases) and 2011 (394 cases).

Table 3. Example of GUHA rules

	2010		2011	
	conf	supp	conf	supp
Tma4(<0;40))	0.88	0.1696	0.83	0.15
Tma4(<0;40)) & Vle6_vle7(<-133;-30)...<0;1))	0.94	0.1578	0.94	0.13

6. FUTURE WORK

Future work will continue to develop, refine and test both the decision tree and GUHA methods on expanding data sets and validate findings across multiple presentations of the modules. Additional demographic data will be added in pursuit of refinements. One possible area of interest is the use of disability data to highlight accessibility issues around module materials. Another key issue is to develop a formal representation of module design, to allow parameters to be set for each module that describes important aspects of the module that can help to improve the modelling. For example, to know the number and timing of TMA’s, whether or not they are compulsory, if they can be substituted, or if they are made deliberately harder or easier.

The overall goal of future work is to implement predictions using real-time data. The current barrier is the disparate nature of data sources, although this is being addressed through an upcoming data warehouse.

7. SUMMARY AND CONCLUSIONS

The results of this work indicate that using even fairly coarse-grain data about students’ activity on a VLE and combining this with some other data sources, it is possible to predict ailing students. Decision trees have been demonstrated to be suitable for

prediction, particularly at the start of a module, when there is commonly a high attrition rate. The main finding has been that the best predictor is based on changes in the student's own VLE activity, compared to their previous activity. In other words, if a student has started out clicking and then stops, this is more of a warning than if their clicking behaviour has been low to start with. Similarly, if a low-clicking student reduces clicking by only a small amount, then this may be significant in terms of the percentage drop compared to their previous clicking behaviour, rather than in terms of their overall level of activity.

GUHA has been investigated as a method both for improving the explanations of student failure and for being more able to accurately predict later drop out. The preliminary results demonstrate that GUHA is successful in this aim. For both decision trees and GUHA, the results are accurate when applied across different presentations of the same module. Whereas the decision tree, being more general, can have some success when applied across modules, the varying nature of module design means that the more detailed explanations of GUHA are unlikely to hold from one module to the next.

Taken overall, these findings suggest that online learners have different learning approaches and this is reflected in their use of the VLE. Online learners do not have to sit in front of a computer and click to learn, instead they may read a page once and make notes, or print it out, or save it onto their computer for offline use. It is not possible to draw conclusions about learner engagement solely based the amount they click, nor even by the types of activities they are clicking on. In terms of informing future research, both within the Open University and for other institutions, the suggestion is to develop profiles of online learning styles with which to classify the different online learning styles and from which it is possible to identify changes in behaviour that can indicate a developing problem. Similarly, there is a need to take into account the interplay between how a module is structured and how the VLE is intended to be used within that structure. This is especially true when making predictions using VLE activity, since it could easily be some feature of a particular module that influences VLE behaviour, such as an assessment has been made deliberately easy (which could mean less required activity) or else a lot of module materials need to be read or referenced for another (thereby increasing activity for that TMA).

The work described in this paper opens up several interesting possibilities for future work and already provides results that could be integrated with live data to produce information to help tutors to provide earlier interventions to struggling students.

8. ACKNOWLEDGMENTS

Our thanks to JISC for their funding and support of this work.

9. REFERENCES

- [1] Arnold, K.E. and Pistilli, M.D. (2012) Course Signals at Purdue: Using Learning Analytics to increase student success. *LAK'12*, 29 April – 2 May, Vancouver, Canada
- [2] Campbell, J.P. and Oblinger, D.G. (2007) Academic Analytics. White paper. *Educause*. <http://net.educause.edu/ir/library/pdf/pub6101.pdf>
- [3] Dekker, G., Pechenizkiy, M. and Vleeshouwers, J. (2009) Predicting Students Drop Out: a Case Study, In *Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09)*, pp. 41-50
- [4] Diaz, D. P. (2000). *Comparison of student characteristics, and evaluation of student success, in an online health education course*. Unpublished doctoral dissertation, Nova Southeastern University, Fort Lauderdale, Florida. Retrieved from <http://academic.cuesta.edu/ltseries/sitepgs/pdf/dissertn.pdf>
- [5] Frankola, K. (2002). *Why online learners drop out*. Crain Communications, Inc. Retrieved from http://www.kfrankola.com/Documents/Why%20online%20learners%20drop%20out_Workforce.pdf
- [6] Fusch, D. (2011) *Identifying At-Risk Students: What Data Are You Looking At?* Retrieved from <http://www.academicimpressions.com/news/identifying-risk-students-what-data-are-you-looking>
- [7] Galusha, J. M. (1997). *Barriers to learning in distance education*. Retrieved from <http://www.infrastruction.com/barriers.htm>.
- [8] Hájek, P., Havel, I., and Chytil, M. (1966). The GUHA method of automatic hypotheses determination. *Computing*, 1, 293-308
- [9] Hájek P., Holeňa and Rauch J. (2010) The GUHA method and its meaning for data mining. *J. Computer and System Sciences*, 76, 34–48.
- [10] Hammang, J.M., Campbell, J.P., Smith, V.C., Sweeley, D., Ushveridze, A. and Woosley, S. (2010) *Predictive Analytics: Building a Crystal Ball for Student Success*. Webinar. http://www.uregina.ca/orp/PapersPresentations/SCUP_Webinar_Sept29_2010.pdf
- [11] Mazza, R. and Dimitrova, V. (2004) Visualising Student Tracking Data to Support Instructors in Web-Based Distance Education, *13th International World Wide Web Conference - Alter-nate Educational Track*, 154-161.
- [12] Pistilli, M.D. and Arnold, K.E. (2010) *Purdue Signals: Mining real-time academic data to enhance student success*. About Campus, July-August, 22-24 DOI: 10.1002/abc.20025
- [13] Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- [14] Rivera, J. C. and Rice, M. L. (2002). A comparison of student outcomes & satisfaction between traditional & web based course offerings. *Online Journal of Distance Learning Administration*. The State University of West Georgia, 5(3).
- [15] Simpson, O. (2004) The impact on retention of interventions to support distance learning students, *Open Learning: The Journal of Open and Distance Learning*, 19, 1, 79-95
- [16] Simpson, O. (2008) Motivating learners in open and distance learning: do we need a new theory of learner support? *Open Learning: the journal of open and distance learning*, 23, 3, 159-170
- [17] Smith, V.C. and Sweeley, D. (2010) *Predictive Analytics: Increasing Online Student Engagement and Success*. Retrieved from http://www.uregina.ca/orp/PapersPresentations/SCUP_Webinar_Sept29_2010.pdf
- [18] Thompson, M.M. (1998). Distance learners in higher education. In C.C. Gibson (Ed.), *Distance learners in higher education* (pp. 9-24). Madison, WI: Atwood Publishing.