

Towards quality in multiple-choice assessment

COLIN CARMICHAEL

Colin.Carmichael@usq.edu.au

ROD ST HILL

University of Southern Queensland, Toowoomba

Australia

Abstract

To many researchers in the area of teaching and learning, the above title may be an oxymoron. However, the use of multiple-choice in assessment programs has been and continues to be a popular choice amongst academics in the tertiary sector. Anecdotal evidence suggests that, while academics using multiple-choice tests may check the overall associated mark distribution for a test, few check its psychometric properties (for example, test reliability) and fewer still check the psychometric properties of the items within the test. The building blocks of quality multiple-choice tests are quality multiple-choice items. The psychometric quality of an item can be determined in part through an assessment of its ability to discriminate adequately between students of high and low ability. In a pilot study of three Faculty of Business courses at the University of Southern Queensland, statistical tests associated with the application of Rasch models to multiple-choice test results were used to identify items that failed to discriminate between students of high and low ability. This article reports on the findings of this study, discusses the consequences of using items with low discrimination and examines simple precautions that academics can take in order to avoid using such items.

Key words: multiple-choice items, Rasch models, quality assessment

Introduction

Multiple-choice tests are used regularly in the context of higher education in Australia and elsewhere. Indeed it is common in large classes in the United States to rely solely on multiple-choice items in examinations. This is despite the fact that multiple-choice tests together with other forms of traditional assessment are regarded by some as being 'too narrow to provide sufficient information about student learning' (Garfield & Chance, 2003). While multiple choice tests can produce reliable results they are often of limited validity (Knight, 2000; Paxton, 2000). For example, Blackman & Darmawan, (2004) found in a study aimed to model the achievement of medical students that there was a negative correlation between student achievement on a 150-item multiple-choice test and the level of postgraduate qualification achieved by the students. Essentially, students with those abilities that we as academics regard as valuable were less likely to perform well in this particular multiple-choice test than students without this postgraduate training. Despite these criticisms, multiple-choice tests remain a popular form of assessment in many academic fields. In a survey of business school disciplines in the United States (Michlitsch & Sidle, 2002), it was found that 54% of faculty members used multiple-choice assessment and 67% regarded this method of assessment as being moderately to strongly effective. While good multiple-choice items can 'provide extensive content coverage with problems that require higher order thought' (Tanner, 2003) it is arguably their ease of marking that makes them an attractive proposition for cash-strapped academics managing very large classes.

In the context of the Faculty of Business at the University of Southern Queensland not only are the costs of assessing large classes an important motivation for using multiple-choice items, but also there is added economic pressure that arises from the need to moderate student results from multiple international operations primarily in Asia, Europe and the Middle East. Moderating short and long answer type assessment items from tens of different international locations is time-consuming and expensive, so more and more reliance is being placed on the results of multiple-choice test items. In some courses the final examination comprises entirely multiple-choice items. In

this context, the importance of reliable and valid items is paramount if quality is to be maintained.

Tests used to grade and/or assess student learning need to reflect both the content and stated learning objectives of interest to the teacher. For this reason, most academics design the tests used to assess their students' learning, although many rely heavily on test banks provided by text book publishers that have themselves been designed by academics. Some researchers, however, suggest that many academics do not have the requisite knowledge to assess the quality of these tests (Elton, 1998; Orrell, 2004) and publishers do not provide any indication as to the discrimination power of individual items. (Some do provide an indication of the level of difficulty of each item). Certainly the subject expert is best placed to assess the validity and content coverage of a test, but it is not clear that subject experts can assess reliability. At a test level a recommended measure of test reliability is the Coefficient Alpha (Kehoe, 1995), although Burton & Miller, (1999) suggest alternative approaches to this. Achieving reliability in tests is based upon the use of psychometrically sound test items. Tests must comprise individual items that range in difficulty, but, more importantly, are able to discriminate between students of high and low ability. Multiple-choice items that fail to discriminate usually fall into one of the following three categories:

- items that most students answer correctly;
- items that most students answer incorrectly; and
- items that are structured in such a way that enables students of lower ability to guess the correct option while students of higher ability choose an incorrect but attractive option.

Arguably all such items are of limited (if any) use in measuring students' achievement (Kehoe, 1995). In order to achieve quality in a test, academics must aim to include only quality items, that is, to exclude those items with low discrimination. However, in doing so academics also need to be mindful that they do not omit assessment of key learning objectives (McCoubrie, 2004).

How then can academics either developing their own tests or using items provided by a textbook publisher identify suitable test items and write new items that provide discrimination? In the pilot study reported in this paper, a sample of six multiple-choice examinations and tests were analysed from courses in the Faculty of Business at the University of Southern Queensland, Australia. The aims of this study were to:

- develop mechanisms for the identification of low discriminating items;
- gauge the extent that multiple-choice tests contain these items; and
- provide recommendations to academics using multiple-choice tests on how to identify such items.

Methodology

A selected sample of first-year business courses was obtained. Lecturers in these courses were invited to submit their test items for analysis. The courses included an introductory computing course, an introductory management course and an introductory law course. Data from both semester 1 (March to June) and semester 2 (July to November) of the 2004 academic year were analysed (see Table 1). In all instances only a fraction of electronic scripts (actual item responses for each student recorded electronically) were available. This was due to a number of reasons, including: failure of students to undertake the actual assessment task, the unavailability of actual student responses from overseas examination centres, and the inadvertent deletion of some files containing this information.

Table 1: Number of students enrolled and percentage of scripts available.

Course details	Number of enrolled students	Percentage of electronic scripts available
Computing 1	881	50
Computing 2	563	35
Management 1	1001	48
Management 2	569	30
Law 1	588	29
Law 2	684	24

(a) Statistical techniques for the identification of low discriminating items

Conventional methods for the identification of low discriminating items in a test rely on the comparison of students' performance on a given item and their overall test performance. These methods are founded on Spearman's Classical True Score Model, which assumes that students' actual performance in a test is linearly related to their ability (true scores). A common method is to calculate the Pearson correlation coefficient between students' performance on a given item and their total test scores.

Rasch models offer an alternative to conventional methods for test analysis. (See Hambleton & Swaminathan, (1985), for a good starting reference). They are based primarily on a student's performance on a given item rather than the whole test, and assume a non-linear relationship between this performance and the student's ability. More specifically, a Rasch model, as it applies to dichotomously-scored test items, attempts to describe the relationship between a student's ability and the likelihood of the student correctly answering a given item. This relationship can be shown graphically using an 'item characteristic curve' (see, for example, Figure 1). In this particular item characteristic curve, we see that students with low ability will have a much smaller chance of successfully answering the item than those with high ability. The difficulty of an item is by definition the point on the ability scale that corresponds to a probability of success equal to 0.5. (In this case, the difficulty is 0). The discrimination of the item is the gradient of the curve at this point and, in Rasch models, it is assumed that this is the same for each item in the test. Student abilities in Rasch models are estimated from their total test score, subsequently scaled and expressed as logits (the natural log of their odds ratio). An average student will have an ability of 0 logits, while a talented student will have an ability of 3 logits. In Rasch models, item difficulties are measured on the same scale as student abilities; conventional methods do not even try to define measurement scales.

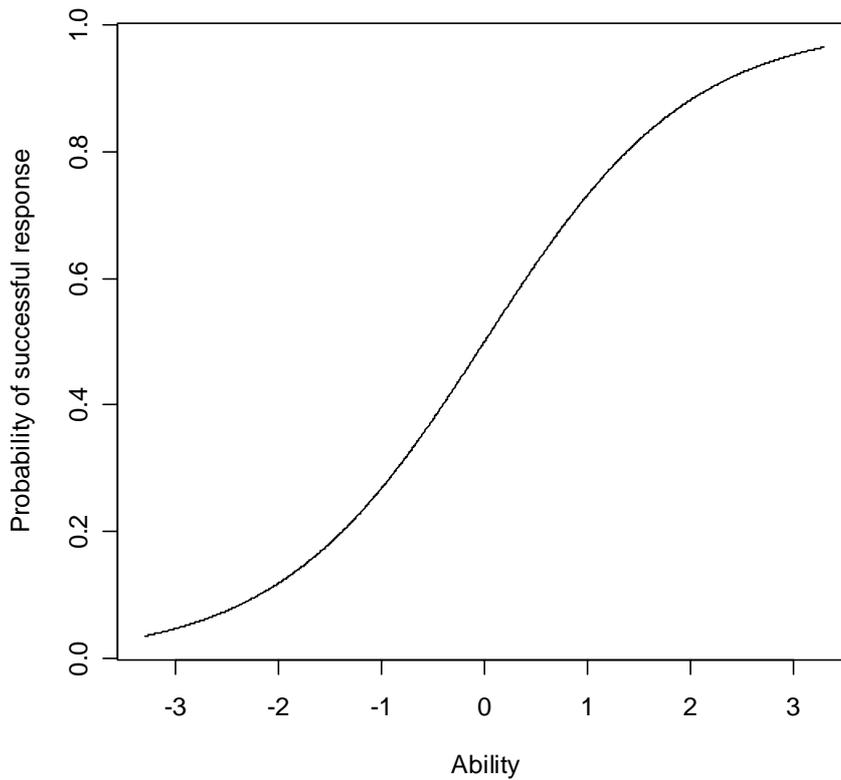


Figure 1: Typical item characteristic curve

Mathematically, the Rasch model is similar to the logistic regression of student responses on their estimated ability (as per their total test score). The model, as it applies to each item in the test, can be written:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times a ,$$

where p is the proportion of students who correctly answer the item and a their ability. In the Rasch model it is assumed that the discrimination parameter $\beta_1 = 1$. Statistical tests associated with the use of logistic regression can therefore be applied with this model.

In this analysis, the Rasch model was fitted to each set of examination results using the joint maximum likelihood method which has been the primary approach to item parameter estimation since its formulation by Birnbaum in 1968 (Baker, 1992). In this method, both student abilities and item difficulties are estimated. Logistic regression was then applied to each set of student abilities in order to re-estimate the

discrimination parameter β_1 . These parameters were subsequently tested to see whether they were significantly different from zero at the 5% level of significance. The test employed was based on an analysis of deviance method devised by McCullagh & Nelder, (1989) and will be referred to subsequently as the deviance test. The software used was written specifically for this study and was based on the statistical package R.

As recommended by Hambleton & Swaminathan, (1985) plots were also generated for each item showing the 'ideal model' and the actual data. To construct these plots, student responses were divided into 10 subgroups based on estimated ability. The proportion of each subgroup that correctly answered the item was then plotted against the average ability for that subgroup. (See, for example, Figure 2 which shows student responses to an item of low discrimination). These plots proved to be a good visual aid for academics; however their usefulness is compromised by small sample sizes and items with few options.

Rasch models are not easy to understand without considerable knowledge of probability and have strong underlying assumptions that are difficult to meet (Burton, 2005; Loyd, 1988). For this reason, it was felt that conventional measures for identifying low discriminating items should also be investigated. A common method is to measure the degree of correlation between students' performance on a given item and their overall performance in the test. This coefficient will be referred to subsequently as the discrimination coefficient. Based on the estimated standard error of the Pearson correlation coefficient, we were able to establish sample-dependent benchmarks (see Table 2). For example, discrimination coefficients below 0.1 in a sample of 400 are not significantly different from zero (at the 5% level of significance) while in a sample of 100 this benchmark would increase to 0.2.

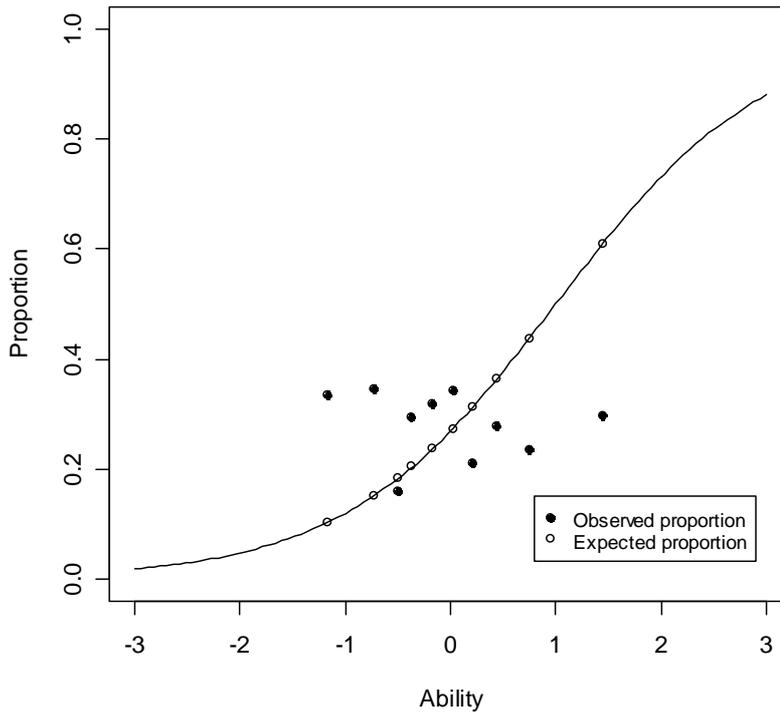


Figure 2: Observed and expected responses for item displaying low discrimination

Table 2: Suggested lower limits for discrimination coefficients

Sample size	Lower acceptable limit of discrimination coefficient
100	0.2
200	0.15
400	0.1
N	$\approx \frac{1.96}{\sqrt{N-3}}$

It was found that the two tests used in this study, namely the deviance test and the discrimination coefficient test (described above), displayed a high degree of agreement (see Table 4), with the latter identifying fewer items as having low discrimination. The deviance test is known to identify items incorrectly as non-discriminating in small samples. This may explain some of the discrepancies that exist between the two tests, however further research needs to be undertaken in this area.

(b) Analysis of items displaying low discrimination

Items that were identified as having low discrimination were subsequently analysed in an attempt to identify attributes that may cause such poor psychometric behaviours. In a study that examined the quality of multiple-choice items, Hansen & Dexter, (1997) cited 12 accepted attributes of good multiple-choice design. One in particular looked at the importance of not providing verbal clues that in some way eliminate one or more of the distractors (see Table 3) and another the importance of not using ‘all of the above’ as an option.

Table 3: Question faults that should be avoided in multiple-choice items

No.	Question fault
1.	Similarity of wording between the stem of the question and the correct option.
2.	More detail or use of textbook language in the correct option than in other options.
3.	The use of absolute terms such as ‘never’ in the distracters.
4.	Using pairs of options that are all inclusive
5.	Two or more options with the same meaning

Items that displayed significantly low discrimination in this study were analysed for violations of these particular principles of good item design. We based this part of the analysis on those items identified using the discrimination coefficient test as this was by the far the more conservative of the two tests. This then provided a basis for professional development of the staff involved in creating multiple-choice test items.

Results

(a) Identification of low discriminating items

Items that were found to have discrimination not significantly different from zero were identified using the two tests described earlier. Results are shown in Table 4.

With the exception of one semester examination and using the more conservative test, less than 5% of items within a given examination were found to have low discrimination. Previous research (Carmichael, Fahey, & Plank, 2005) indicated that such a low proportion of items is unlikely to affect final student grade distributions adversely. Nevertheless, in our quest for quality, such items need to be reviewed and if necessary modified.

Table 4: Items displaying significantly low discrimination

Examination	Details of test		Items of low discrimination identified by:	
	No. items	No. options	Deviance test	Discrimination coefficient test
Computing 1	60	4	10, 19, 35, 41, 43, 46	10, 35, 46
Computing 2	60	4	13, 24, 41, 45, 48	24, 45
Management 1	40	2 and 3	16	16
Management 2	50	2 and 3	2, 3, 7, 10, 13, 14, 16, 18, 23, 27, 29, 31, 45, 46, 48, 49, 50	10, 13, 14, 16, 23, 27, 29, 31, 48, 49, 50
Law 1	20	5	5, 8	5
Law 2	20	5	6, 8	6

Both of the tests used to identify low discriminating items are based on the assumption that the measures of student ability are reliable. In cases where guessing may be more pronounced, the total test score is a less reliable measure of student ability and both of these tests are less conclusive. For example, the large number of identified items in the Management 2 examination may in fact be a result of a significant guessing factor, as the examination included 11 True/False items and the sample was relatively small. Consequently it is likely that the examination in question had structural problems rather than problems with individual items. This hypothesis is supported by the low coefficient alpha value for the examination ($\alpha = -0.21$). The coefficient alpha value is a measure of the average correlation between all items of the test. In an educational context, it is suggested that the lower limit of coefficient alpha is 0.5 (Kehoe, 1995), in other words there should be some degree of correlation between items in the test.

(b) Analysis of low discriminating items

Whilst the above methods are useful for identifying poorly performing items after the event, some means of identifying these items at the test design stage would be of more use to the academic. Accordingly those items that had been identified using the more conservative discrimination test were assessed against the good characteristics of multiple-choice tests discussed earlier in the methodology section and shown in part in Table 3.

Of the 19 items identified by this test, six showed obvious violations of the principles of good design discussed earlier. These provided clues for lower ability students. One example is item M210 shown in Table 5, which contains more wording in its correct option. An analysis of answering patterns from the bottom one third of students (based on their total test score) and the top one third is shown in Figure 3. This shows that more students in the lower third ability group were selecting the correct option (presumably through guessing) than those in the upper one third ability group.

Of the remaining 13 items that displayed significantly low discrimination, five were True/False items. Arguably achieving a good discriminating True/False item is more good luck than good management because students have a high chance of selecting the correct option through guessing. One of the remaining items had been identified by the lecturer after the examination as having contained a printing error. A further two items had difficulties less than -2.5 (in other words were answered correctly by more than 99% of the students). There were no obvious reasons to explain the low discrimination in the remaining five items and a further testing of these items in a larger sample may be necessary.

Table 5: Item violating second principle of design

Item M210: If the manager were to use the traditional or mainstream theory of job satisfaction to improve the level of job satisfaction among his employees he would need to:

1. Satisfy the intrinsic aspects of their work, such as recognition and achievement.
2. Satisfy the extrinsic conditions surrounding their jobs, such as pay, physical work conditions, job security.
3. Satisfy the aspects of their work, such as equity of pay, work conditions, the mental challenge of the work, and supportiveness of fellow workers. **(Correct option).**

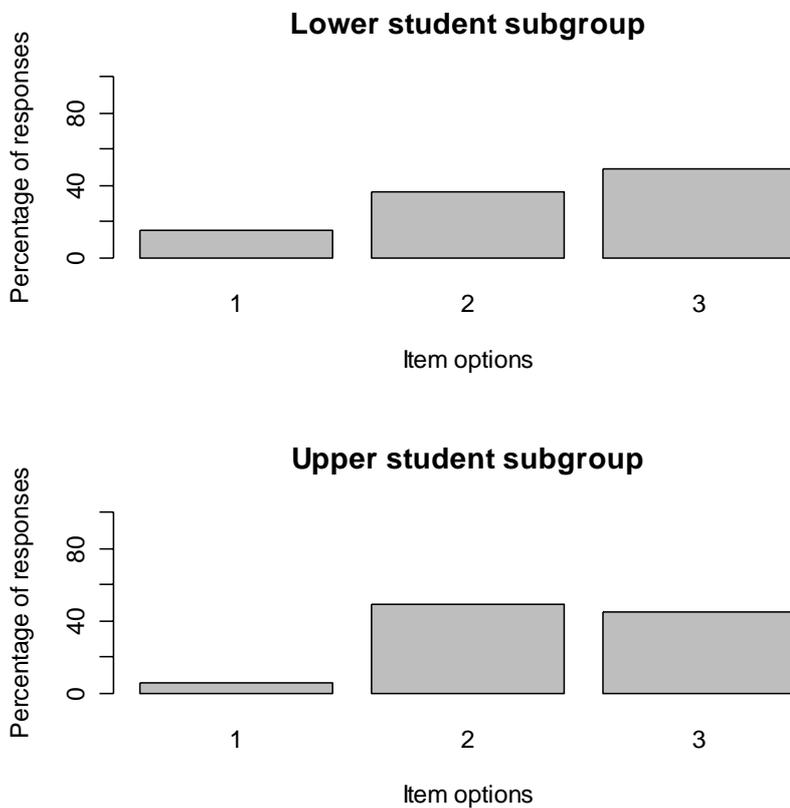


Figure 3: Answering patterns for item M210 by ability level.

Conclusions

The purpose of this study was to assess the quality of a sample of multiple-choice items through an analysis of their ability to discriminate between students of high and low ability. Each of the tests contained some items whose discrimination was not significantly different from zero. Arguably there is little point in using an item in which students of lower ability have the same or better chance of success than students of higher ability.

Quality multiple-choice items are difficult to design and using someone else's items will not necessarily rectify this problem. As an example, in a random sample of 40 multiple-choice items from 12 auditing textbooks, Hansen and Dexter (1997) found that 28% had one or more design violations. How then can quality truly be achieved in multiple-choice items? In this study it was found that simply following principles of good design could reduce the number of defective items, but even those items that appear on the surface to be well designed may in fact fail to discriminate adequately. It is recommended that examiners therefore consider a routine post-test analysis of their multiple-choice items. Such an analysis should include calculation of test reliability coefficients, item difficulty estimates and item discrimination estimates. Whether practitioners use techniques associated with the use of Rasch models or more conventional techniques is academic, provided that they actually undertake this analysis. Academics can then opt to adjust student grades and/or rectify the items prior to their subsequent use.

Although quality multiple-choice assessment has been the topic of discussion in this paper, it is quality assessment that is the significant issue in higher education generally. This is especially the case in Australia where the Federal Government has recently established the Carrick Institute for Learning and Teaching in Higher Education. The Institute has a number of responsibilities, one of which is to improve assessment practices within the higher-education sector. We believe that assessment practices in Australian higher education can be improved primarily through greater teacher awareness. Academics need to be aware of the limitations of any assessment instrument that they use. For example, multiple-choice tests, while easy to mark, are

difficult to create and, unless they are created carefully, can be of limited validity. The multiple-choice test, therefore, should be only one of several methods used to assess students' learning outcomes.

References

- Baker, F. B. (1992). *Item Response Theory - Parameter Estimation Techniques*. New York: Marcel Dekker, Inc.
- Blackman, I., & Darmawan, I. G. N. (2004). Graduate-entry medical student variables that predict academic and clinical achievement. *International Education Journal*, 4(4), 30-41.
- Burton, R. F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education*, 30(1), 65-72.
- Burton, R. F. & Miller, D. J. (1999). Statistical modeling of multiple-choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing. *Assessment & Evaluation in Higher Education*, 24(4), 399-411.
- Carmichael, C., Fahey, P. & Plank, A. (2005). Assessing our assessment. *Working paper, Dept. Mathematics & Computing, USQ*.
- Elton, L. (1998). Are UK degree standards going up, down or sideways? *Studies in Higher Education*, 23(1), 35-43.
- Garfield, J. B. & Chance, B. L. (2003). New approaches to gathering data on student learning for research in statistics education. *Statistical Education Research Journal*, 1(2), 38--41.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff.
- Hansen, J. D. & Dexter, L. (1997). Quality multiple-choice test questions: item-writing guidelines and an analysis of auditing test banks. *Journal of Education for Business*, 73(2), 94-102.
- Kehoe, J. (1995). Basic Item Analysis for Multiple-Choice Tests. *Practical Assessment, Research and Evaluation*, 4(10).
- Knight, P. T. (2000). The value of a programme-wide approach to assessment. *Assessment & Evaluation in Higher Education*, 25(3), 237-251.

- Loyd, B. H. (1988). Implications of Item Response Theory for the Measurement Practitioner. *Applied Measurement in Education*, 1(2), 135-143.
- McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teachers*, 26(8), 709-712.
- McCullagh, P. & Nelder, J. A. (1989). *Generalised Linear Models*. London: Chapman and Hall.
- Michlitsch, J. F. & Sidle, M. W. (2002). Assessing Student Learning Outcomes: A Comparative Study of Techniques Used in Business School Disciplines. *Journal of Education for Business*, 77(3), 125-130.
- Orrell, J. (2004). *Beyond the "Blindfolded High Jump!" Professional practice to improve the quality of assessment*. Paper presented at the Evaluations and Assessments Conference, Melbourne, Australia.
- Paxton, M. (2000). A linguistic perspective on multiple-choice questioning. *Assessment & Evaluation in Higher Education*, 25(2), 109-119.
- Tanner, D. E. (2003). Multiple-choice items: pariah, panacea, or neither of the above. *American Secondary Education*, 31(2).

Acknowledgements

The authors would like to gratefully acknowledge the Faculty of Business, USQ for research funding support and the dedicated lecturers involved in the study for the use of their data.