ITEM SELECTION AND APPLICATION IN HIGHER EDUCATION

Andrew Boyle, Dougal Hutchison,
Dave O'Hare and Anne Patterson

Item Selection and Application in Higher Education

Andrew Boyle
Dr Dougal Hutchison
National Foundation for Educational Research
The Mere
Slough
Berkshire
SL1 2DQ.

Email: a.boyle@nfer.ac.uk

Dr Dave O'Hare
Anne Patterson
University of Derby
Kedleston Road
Derby
DE22 1GB

Abstract

Over the past ten years the use of computer assisted assessment in Higher Education (HE) has grown. The majority of this expansion has been based around the application of multiple-choice items (Stephens and Mascia, 1997). However, concern has been expressed about the use of multiple choice items to test higher order skills.

The Tripartite Interactive Assessment Development (TRIAD) system (Mackenzie, 1999) has been developed by the Centre for Interactive Assessment Development (CIAD) at the University of Derby. It is a delivery platform that allows the production of more complex items. We argue that the use of complex item formats such as those available in TRIADs could enhance validity and produce assessments with features not present in pencil and paper tests (cf. Huff and Sireci, 2001).

CIAD was keen to evaluate tests produced in TRIADs and so sought the aid of the National Foundation for Educational Research (NFER). As part of an initial investigation a test was compiled for a year one Systems Analysis module. This test was produced by the tutor (in consultation with CIAD) and contained a number of item types; both multiple-choice items and complex TRIADs items.

Data from the test were analysed using Classical Test Theory and Item Response Theory models. The results of the analysis led to a number of interesting observations. The multiple-choice items showed lower reliability. This was surprising since these items had been mainly obtained from published sources, with few written by the test constructor. The fact that the multiple-choice items showed lower reliability compared to more complex item types may flag two important points for the unwary test developer: the quality

of published items may be insufficient to allow their inclusion in high-quality tests, and furthermore, the production of reliable multiple-choice items is a difficult skill to learn. In addition it may not be appropriate to attempt to stretch multiple-choice items by using options such as 'all' or 'none of the above'. The evidence from this test seems to suggest that multiple-choice items may not be appropriate to test outcomes at undergraduate level.

Introduction

Collaboration

This paper is the result of collaboration between three parties, following initial contact at the fifth Computer Assisted Assessment (CAA) conference in 2001. The collaborators are described below.

The Centre for Interactive Assessment Development (CIAD) is located at the University of Derby. It developed TRIADs, an innovative assessment delivery engine, and also provides an assessment development service to tutors who are members of the TRIADs network. CIAD staff have presented several papers at previous CAA conferences (see McKenzie 1999, O'Hare 2001). Additionally, CIAD provides an advice service to University of Derby lecturers who develop computerised assessments.

The National Foundation for Educational Research (NFER) is a leading UK educational research organisation. NFER's Assessment and Measurement Department (AMD) produces assessments for educational and commercial clients, but has a particular reputation in the five-to-sixteen sector. AMD currently has departmental focus on computerised assessment and assessment for learning.

Anne Patterson is a lecturer in the School of Computing and Technology at the University of Derby. Whilst she is an experienced lecturer, she is not an especially experienced test developer. But she has made a practical commitment to improving the quality of computerised assessments delivered to her students.

CIAD and NFER aim to develop a joint research agenda. Also, their collaboration, whilst informal, is intended to be mutually beneficial; CIAD will be assisted in its research into complex item types, NFER will advance its current interests in computerised assessment and assessment for learning. One way of doing this may be to use TRIADs to develop curriculum-friendly assessments for use in primary schools.

CIAD and NFER hope to do research that will result in novel findings on the nature of complex items in computerised assessment. The current paper concentrates on a part of some exploratory work that has been undertaken. However this limited focus does allow an investigation of the conference themes listed in the abstract.

Test development methodology

NFER produces, and researches, high-quality assessments. These include a range of instruments, from the Foundation Stage Profile for Reception year children to National Curriculum tests and products for corporate and professional clients such as the Theory Test for Drivers and Riders.

In common with other assessment production institutions NFER has an established test development methodology. Sainsbury (2001) describes a seventeen-stage process stretching over two years to develop a Key Stage Two reading test. Clausen-May (2001, p.13 *et seq*.) explains how test specification ensures that instruments cover the curriculum adequately, and use the best question contexts. She also describes how item writing is a complex process in which skilled writers use their in-depth knowledge of item formats, and how items go through several quality-control stages (2001, p.48 *et seq*.).

When University lecturers develop tests for their undergraduates, the process may well be somewhat different. Assessment development is unlikely to be the principal responsibility of even experienced lecturers. And thus, test development is likely to involve writing questions in formats without the developed expertise that would be required in an institutional assessment provider. Also, lecturers may seek to minimise the workload of assessment development by using items from published sources.

Thus, it is not usual for University assessments to be produced with the same degree of formality and documentation as is the case in assessment institutions. Rather, a University may provide a central advice service (CIAD has this function within the University of Derby), and the test developer and an adviser may discuss the best ways to realise an assessment. However, the final decision on the content of the assessment resides with the curriculum expert (the lecturer) rather than the adviser.

The Current Test

The studied assessment was a test from a level one module in Systems Analysis. The module was offered to students on the Computing Science and Information Systems BSc and BTEC programmes (HND and HNC). As such the test takers were a mixed ability group.

The test consisted of 25 items. These items addressed a range of topics in the Systems Analysis syllabus. Additionally, the items were of several formats. The formats were: label diagram (LD), multiple choice (MC), multiple selection (MS), multiple text entry (MT), sequencing (SE), and true/false (TF). For the purposes of the current research, the items were categorised as simple or complex. Complex items were felt to exemplify innovative item types, and to be likely to produce a wide spectrum of scores. Simple items were more typical of traditional dichotomous (right/wrong) test items. Simple items were mainly multiple choice, but they were not limited exclusively to that type. The number of items of each format, and their simple/complex designation, is shown in **Table 1**.

Simple or complex item	Type	Total
Simple	MC	16
	MT	1
	TF	2
Simple Total		**19**
Complex	LD	3
	MS	2
	SE	1
Complex Total		**6**
Grand Total		**25**

Table 1: Simple and complex items

Assessment Stakes

Many writers on assessment have sought to distinguish low-, medium- and high-stakes testing systems. The stakes of a testing system derive from the impact that decisions based on the test have on test takers' lives (Bachman and Palmer, 1996, p.96). From a test developer's point of view, the stakes involved in a testing system also affect: the choice of acceptable reliability levels (*ibid.* at p.135), the test development procedures adopted (p.266) and even the way that human resources are deployed on the project (p.157).

Shepherd (2001) has constructed a useful summary of the properties of low-, medium- and high-stakes assessments.

	Stakes		
	Low	Medium	High
Decisions	None	Can be reversed	Difficult to reverse
ID individual	None	Important	Very important
Proctoring	None	Yes	Constant
Options	Study more	Pass, fail, work harder	Pass or fail
Item & test development	Minor	Takes time	Significant
Items created by	Subject expert	Subject expert	Subject expert
Statistics checked	Subject expert	Time to time	Psychometrician

Table 2: Assessment stakes

The current assessment was part of a module which students needed to pass in their first year (25 per cent out of one of eight modules taken in the year). Thus, decisions made on the basis of this assessment were not irrevocable. However, care was taken in test development, and test conduct procedures were rigorous ('proctoring' was taken seriously). Subsequent to running the 2001 test, considerable energy is being invested to improve future administrations of the test. This includes the current research. Thus, adapting Shepherd's categories, the current test could be described as 'low-to-medium stakes'.

The Value of Innovative Item Types in Computer-assisted Assessment

Huff and Sireci (2001, p.17) posit that innovative item types might enhance validity in computer-assisted assessment. They point out that pencil and paper items may 'measure knowledge, skills and abilities in an artificial way'. They state that multiple-choice items can be considered to be inadequate for assessing higher-level skills such as reasoning, synthesis and evaluation. They remark that innovative item types may aim towards more 'authentic' assessment of knowledge, skills and abilities. Also, computerised item types might potentially measure a construct domain more broadly and assess higher-level cognitive skills more efficiently than traditional paper and pencil tests. In doing so, computerised assessment could integrate positively into the learning process:

> 'the innovative item formats used in [computerised assessment] may have subtle, positive consequences for test developers, examinees and other stakeholders (i.e. improved consequential validity).' (*ibid.*)

However, Olson-Buchanan and Dragsow (1999, p.4), among many others, have pointed out that many computerised assessments are in effect translations of traditional pencil and paper tests to the new medium. Thus, they may miss some of the potential benefits of the innovative item formats.

Data

CIAD provided NFER with data from the test. The data described demographic characteristics of the students. Such data have been used to conduct differential item functioning *(dif)* analysis in the wider NFER/Derby collaboration. However, this analysis is not reported in the current paper. Thus, it is sufficient merely to record that 352 students sat the test, but that one left before the end.

Research methods

This research had two facets: a qualitative review of test items and a statistical analysis of test takers' responses. This statistical analysis was based on two measurement paradigms: Classical Test Theory (CTT) and Item Response Theory (IRT).

Classical Test Theory

CTT output three main types of measures. These were:

- a measure of the test's internal reliability;
- item discrimination;
- item facility.

Reliability is 'the consistency or stability of the measures from a test. The more reliable a test is, the less random error it contains.' (ALTE, 1999). In this research the 'Cronbach's alpha' reliability statistic was calculated. It was also possible to derive a pseudo-measure for 'item reliability'; that is, to calculate the test reliability if the item under study were removed. Thus, if the test's reliability went up if a given item were removed, the removed item could be

said to be relatively 'unreliable'. Conversely, if the test's reliability went down if the studied item were removed, then that item could be thought of as 'reliable'.

Item discrimination is 'the power of an item to discriminate between weaker and stronger candidates' (*ibid.*). In this research item discrimination was calculated as a correlation between item and test performance. Specifically, when test performance was calculated, the score on the item under consideration was excluded; so that the correlation was between the item and the rest of the test, rather than between the item and the whole test. Item facility is the proportion of persons who answered the item correctly.

Item Response Theory

Item Response Theory is underpinned by rather different assumptions than Classical Test Theory. Some basic distinctions between the two analytical approaches are listed below.

Classical Test Theory, as the name suggests, is the more ancient of the two paradigms. Its particular advantage is that it is based on relatively weak assumptions and so real test data are more likely to fit models based on CTT (Hambleton and Jones 1993, p.40). Models based on IRT, in contrast, make relatively strong assumptions about the structure of data; thus real test data are more likely to misfit an IRT model (*ibid.*). In particular, the forms of IRT that will be used in this research assume that measurement is mainly unidimensional. That is, the majority of variance in the data can be explained by a single underlying ability or trait.

In this research, the one-parameter, or Rasch, Item Response Theory model was used to analyse dichotomous items, and a generalisation, the Partial Credit Model, for polytomous items. Dichotomous and polytomous variables can be defined as follows:

- A *dichotomous* variable is one in which there are only two possible valid responses, e.g. male/female, alive/dead, right/wrong.
- A *polytomous* variable is a categorical variable in which the categories are ordered, e.g. strongly dislike, dislike, like, strongly like; scores of 0, 1, 2, 3, or 4 on a test item.

Using the earlier terminology, simple items tended to be dichotomous, and complex items could be considered as polytomous.

The Rasch model of IRT relates test taker ability to the probability of success on an item of given difficulty. **Figure 1** below shows the graph of probability of success against test taker ability for a range of difficulties. The higher a test taker's ability, the higher the probability of passing. A number of items could be displayed on the same graph. If that were done, as item difficulty increased, the curve would move to the right; i.e. at any given ability level, the probability of passing is lower if difficulty is higher.

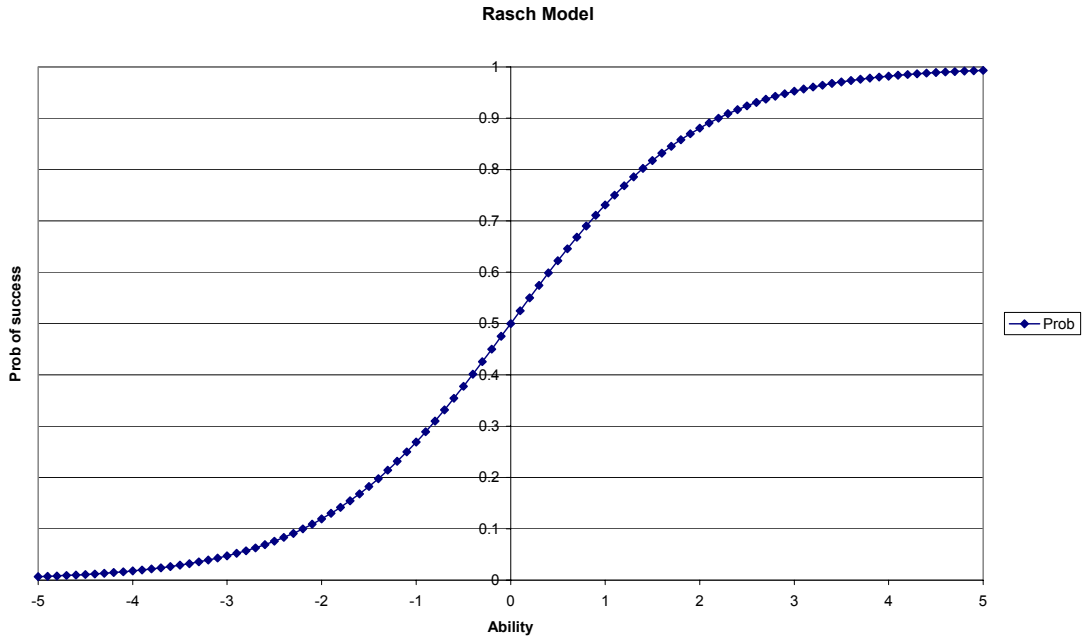

Figure 1: Probability of a correct response in the one-parameter model of IRT

The Item Characteristic Curve (ICC) in **Figure 1** describes an item that fits the Rasch model perfectly. In reality data from items will fit the model in varying degrees. It is important to note that in the current test items were not written with the intention of fitting the Rasch model; they were written to cover a curriculum adequately and to make best use of an assessment delivery platform. Thus it would not be surprising if some items misfit the model to some extent.

The Partial Credit model outputs a **step parameter**, which expresses the probability that a test taker achieved a score *k,* relative to the probability of scoring at one point lower on the scale (*k-1*). Further it was noted above (at page 274) that IRT analyses made stronger assumptions than their CTT analogues, and therefore items were more likely to misfit IRT models. In fact, in this paper, the analysis of misfit will be the main index to be reported and discussed. This misfit analysis will be followed up by graphical representations of misfitting items; to show exactly how misfit has occurred.

Findings

Classical Test Theory analysis

Alpha reliability was calculated as 0.6281. In public high-stakes examinations containing well-constructed objective items, reliability indices of 0.8 and higher are normally expected. Indeed, the reliability is lower than normally expected in tests that are supported by CIAD. The current reliability index may indicate that too high a proportion of test takers' scores reflected non-systematic variability, rather than the candidates' actual abilities in the subject of the test. But the low reliability may also be a function of the shortness of the test.

Alternatively, the ability underlying scoring may not have been unidimensional.

Some more positive findings were derived by sorting the items according to their 'item reliability'. In the table below, the five items where alpha reliability was lowest if the given item was deleted are shown (i.e. these items contributed most to the overall reliability of the test).

Number	Item type	Simple or complex item	Reliability if item deleted
Q02	SE	C	0.570
Q05	MS	C	0.576
Q07	LD	C	0.583
Q17	MS	C	0.586
Q21	LD	C	0.586

Table 3: The five items which, if deleted, caused the greatest decrease in reliability

The table shows that these five relatively reliable items were all complex items.

The second statistic is discrimination. Once again, an interesting finding can be derived by sorting the items according to their discrimination. The table below shows that the five items with the highest discrimination were all complex items.

Number	Item type	Simple or complex item	Discrimination
Q05	MS	C	0.465
Q02	SE	C	0.429
Q07	LD	C	0.374
Q17	MS	C	0.361
Q21	LD	C	0.355

Table 4: The five items with the highest discrimination values

The five relatively discriminating items in this table are the same items that had relatively high reliability, although ordered slightly differently.

Finally, the facilities can be considered. There is no ideal range for facility values. But, generally, if facilities tend towards 50 per cent, rather than items being very easy or very difficult, a greater reliability and spread of scores can be achieved. Most of the items in the current test had a facility within a broad central band. Only two items appeared excessively easy.

Item Response Theory

The Partial Credit model produced step parameters for each item. It also produced information about how the test, and items in it, fit the model. The overall degree of fit for the items in terms of the Partial Credit model had a chi-squared statistic of 243.2 with 125 degrees of freedom. This may indicate a weak fit to the model overall. When the fit of individual items was considered, it was found that ten items of the 25 did not fit well to the model.

Eight of these ten poorly fitting items were 'simple', and, further, six of the misfitting simple items were multiple choice.

The misfitting items were investigated in more depth graphically. In the graphs person-ability is plotted along the *x* axis, and the score that persons of given ability could be expected to achieve on the item, given an efficient operation of the Rasch model, is plotted on the *y* axis. The Item Characteristic Curve (ICC) is an idealised plot of the relationship between person-ability on the test overall, and the characteristics of the individual item, whilst the six small dots show the actual scores of groups of persons at particular points on the ability continuum.

Thus there appeared to be three main sources of lack of fit. Items 3, 19 and 20 appeared to have different slopes from the fitted curves. Item 3 is shown in **Figure 2** for exemplification.

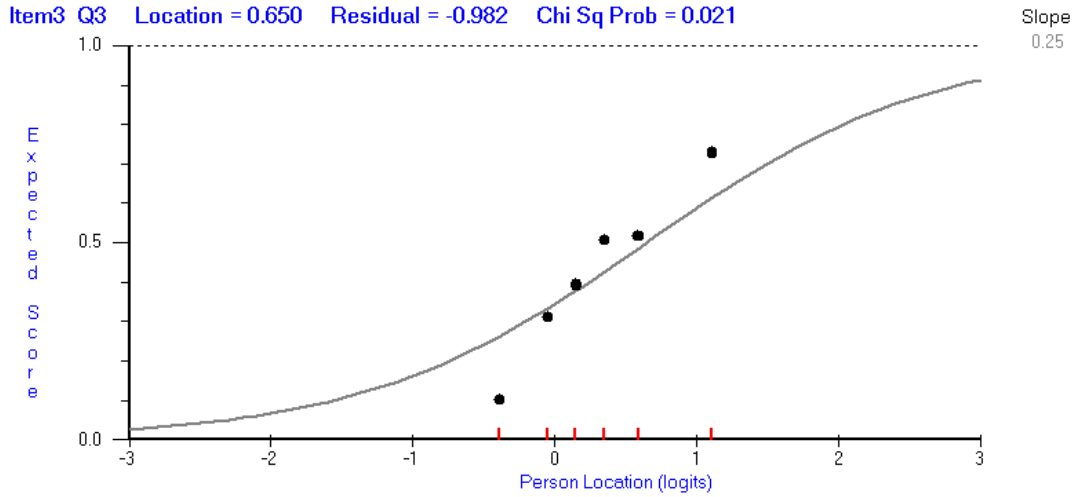

Figure 2: Fit of item 3 to the One-parameter Partial Credit model

The same may hold for questions 7 and 11, though they were also perhaps slightly on the easy side for many of the respondents.

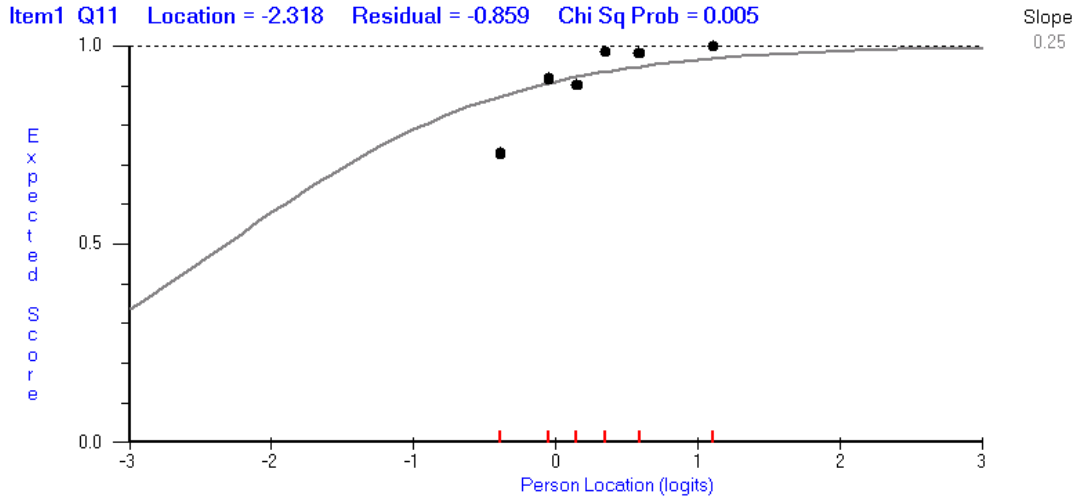

Figure 3: Fit of item 11 to the One-parameter Partial Credit model

Item 18 may have been slightly easy for some respondents, although its curve may denote a degree of non-systematic variation. Item 1 may have lacked discrimination at the lower end of the ability range. The other items: 8, 9 and 12 exhibited non-systematic behaviour and did not increase monotonically with increasing respondent ability. Item 12 below exemplifies this phenomenon.

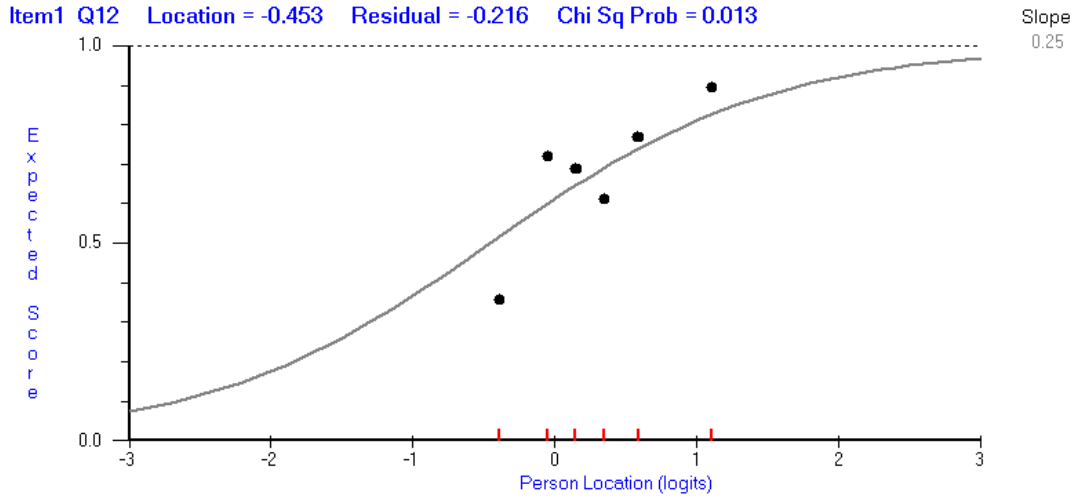

Figure 4: Fit of item 12 to the One-parameter Partial Credit model

Possible reasons for lack of fit could not be investigated further since candidates' responses were not available.

Qualitative Review of Items

Multiple-choice (MC) items were the largest single item type in the current test. Good multiple-choice items are difficult to develop (Bachman and Palmer 1996, p.193). There are a number of resources that describe good practice in the writing of multiple-choice items (some of these can be accessed from the CIAD web site: http://www.derby.ac.uk/ciad/dev/qdesign.html). Also, multiple-choice items were found to have performed especially poorly in statistical analyses of the current test. Therefore it was worthwhile to do a qualitative review of the multiple-choice items.

A specific observation was made about several of the multiple-choice items in this test. Several of the multiple-choice items had 'all' or 'none of the above' as an option. Such items may be put into a degree level test, because the test developer wishes to increase the cognitive demand on students. However, the selection of the 'all/none of the above' option is a cognitive activity of a different order to the selection of one of the factual options. If the actual responses that candidates chose had been available, it might have been possible to replicate the common finding that relatively few candidates select the 'all/none of the above' option in multiple-choice items.

It may also be that 'all/none of the above' items do not integrate positively into the learning process, as Huff and Sireci (2001) have hoped that computerised

assessment should (see page 273 above). There are a number of web sites in the United States that advise students on how to attack multiple-choice items, and how to deal with 'all/none of the above' in particular. The following is an example from the College of Liberal Arts and Sciences at the University of Illinois:

> 'Don't worry about the following choices: all of the above, none of the above, both B & C. Use the process of elimination and simply look at what you've crossed off in the previous choices. Sometimes these final options are correct (especially 'all of the above'), but sometimes they are [a] 'filler.''

Whilst the College of Liberal Arts and Sciences is doubtless providing good test-taking strategies for its students, it is doubtful that the application of these strategies in wider domains would result in more effective learning.

Thus, the presence of 'all/none of the above' is a relevant factor when conducting a qualitative review of multiple-choice items. There are a number of other things to bear in mind when reviewing multiple-choice items. These include the wording and structuring of options and stems, as this can affect the way items function.

In the current test seven items had an 'all/none of the above' type option, whilst nine did not. Also, there was variation in the length of stems in multiple-choice items, and in the presence of features such as negation and conditionality. At a test review meeting the three collaborators felt that perhaps the multiple-choice genre had been 'stretched' in order to make the items challenging enough for degree level test takers. It was hypothesised that by doing this, the test developer had tried to use the multiple-choice item type for a purpose for which it was not intended. Therefore, the suitability of multiple-choice items for undergraduate level tests was doubted.

Conclusion and Further Work

A low-to-medium-stakes test was constructed (with central support) by a University lecturer. The test contained items from published sources, and newly written items. Also, the test was a mix of complex interactional items from the TRIADs test delivery platform, and simple dichotomous items. The test was analysed and several indices were output from the Classical Test Theory and Item Response Theory paradigms. The statistical results were mixed. In particular, simple items had lower reliability and discrimination than complex items. In IRT analyses, once again it was simple items that misfit the IRT model most frequently. Specifically, many multiple-choice items misfit the model. Therefore, a qualitative review of the multiple-choice items was undertaken. It was hypothesised that features such as 'all/none of the above' could be associated with weakly performing items, and that in the current case the multiple-choice genre had been 'stretched' beyond its reasonable limits.

The three parties involved in this research will continue to work together. Particularly, it is hoped to use some of the test items that performed relatively

well in a renewed administration of the test in November 2002. However, some of the poorly performing simple items will be reviewed and rewritten - perhaps maintaining the focus of the items, but delivering them as complex interactional items using the many features of the TRIAD system.

This paper has discussed only a small part of the initial research that has been jointly carried out by CIAD and NFER. It is hoped that future work can look at the psychometrics of complex items, and investigate the empirical properties of good feedback to students. Further, it is hoped to study differential item functioning for test takers of different demographic groups, and for those who use different learning styles. Through this latter research avenue it is hoped to work towards providing a cognitively principled approach to computerised assessment.

Acknowledgements

NFER

Chris Whetton	Director of AMD and Project Director
Bethan Burge	Assistant Research Officer
Christine Taylor	Dissemination co-ordinator (library)

CIAD

Professor Don Mackenzie Director of CIAD and main developer of TRIADs

References

Association of Language Testers in Europe (ALTE). (1999) *Multilingual Glossary of Language Testing Terms* [CD-ROM] Cambridge: Cambridge University Press.

Bachman, L.F. and Palmer, A.S. (1996) *Language Testing in Practice* Oxford: Oxford University Press.

Clausen-May, T. (2001) *An Approach to Test Development* Slough: NFER.

College of Liberal Arts and Sciences of the University of Illinois (n.d.) *Strategies for Multiple Choice Questions* http://www.las.uiuc.edu/students/careeraids/practical_study/mult_choice.shtml (29 April 2002).

Hambleton, R.K. and Jones, R.W. (1993) *Comparison of Classical Test Theory and Item Response Theory and their application to test development.* Educational Measurement: Issues and Practice **12** (3) 38-47.

Huff, K.L. and Sireci, S.G. (2001) *Validity issues in computer based testing.* Educational Measurement: Issues and Practice **20** (3) 16-25.

Mackenzie, D.M. (1999) *Recent developments in the Tripartite Interactive Assessment Delivery System (TRIADs)* **in** Eabry, C. (ed.) Third Annual

Computer Assisted Assessment (CAA) Conference Proceedings http://www.lboro.ac.uk/
service/fli/flicaa/conf99/pdf/contents.pdf (26 April 2002).

O'Hare, D. (2001) *Student views of formative and summative CAA.* **in** Danson, M. and Eabry, C. (eds.) Fifth International CAA Conference Proceedings
http://www.lboro.ac.uk/service/ltd/flicaa/conf2001/pdfs/contents.pdf
(29 April 2002).

Olson-Buchanan, J.B. and Drasgow, F. (1999) *Beyond bells and whistles: an introduction to CAA* **in** Drasgow, F. and Olson-Buchanan, J.B. (eds.) Innovations in Computer Assisted Assessment. Mahwah, NJ: Lawrence Erlbaum Associates.

Sainsbury, M. (2001) *Test development research: a short guide.* TOPIC **25** Item 9.

Shepherd, E. (2001) *High, Medium and Low Stakes Assessments* http://www.science.ulst.ac.uk/caa/presentation/low_high/ (29 April 2002).

Stephens D. and Mascia, J. (1997) *Results of a Survey into the Use of Computer-Assisted Assessment in Institutions of Higher Education in the UK January 1997* http:www.lboro.ac.uk/service/fli/flicaa/downloads/survey.pdf
(29 April 2002).