The Effects of Linguistic Factors on Student Performance on Economics Multiple Choice Questions

Viet-Ngu Hoang, QUT Business School, Phone: 061 7 331 84325, Fax: 61 7 313 81500, Email: vincent.hoang@qut.edu.au

Lyn May, QUT Education Faculty, Phone: 61 7 31383462, Fax: 61 7 31383988, Email: lynette.may@qut.edu.au

Tommy Tang, QUT Business School, Phone: 61 7 3138 2737, Fax: 61 7 313 81500, Email: tt.tang@qut.edu.au

ABSTRACT

This paper proposes a framework to analyse performance on multiple choice questions with the focus on linguistic factors. Item Response Theory (IRT) is deployed to estimate ability and question difficulty levels. A logistic regression model is used to detect Differential Item Functioning questions. Probit models testify relationships between performance and linguistic factors controlling the effects of question construction and students' background. Empirical results have important implications. The lexical density of stems affects performance. The use of non-Economics specialised vocabulary has differing impacts on the performance of students with different language backgrounds. The IRT-based ability and difficulty help explain performance variations.

INTRODUCTION

Multiple-choice (MC) questions are commonly used in introductory Economics units in tertiary education (Siegfried and Kennedy 1995, Watts and Becker 2008, Watts and Schaur 2011). According to a 2010 survey of academics in the field of Economics in the United States of America, MC questions were heavily used in introductory Economics courses, accounting for approximately 42% of total grades (Watts and Schaur 2011). The popularity of MC questions in exams may be due to the advantages of MC testing, such as low grading costs, the potential for timely feedback to students, freedom from scoring bias, the potential for wider sampling of course content, less measurement error (Walstad 1998), and a high correlation between MC test scores and constructed-response test scores in many types of questions (Chan and Kennedy 2002; Walstad 1998). The Test of Understanding of College Economics data revealed that, on average, MC questions accounted for 65.5% of variations in total grades (Siegfried and Kennedy 1995).

An important strand of literature in Economics Education has focused on the driving factors that determine student performance (Anderson, et al. 1994; Becker, et al. 1991; Orhan, et al. 2009; Swope and Schmitt 2006; Walstad and Robson 1997). Factors related to students (i.e. age, gender, academic ability, learning strategies and social-economic and cultural backgrounds), instructors, question construction, and other aspects of teaching and learning can influence student performance. In a context where students come from different language backgrounds, the linguistic abilities of students and the linguistic aspects of the questions are important features of these driving factors.

In 2010, more than 22% of tertiary students studying in Australian universities were international students and in several universities international students accounted for more than 40% of total enrolment (ABS 2011). Of all the broad fields of education, 'Management and Commerce' has the largest international enrolment (52%) In the majority of study programs in this field, introductory Economics is compulsory. Since the eighties, given the increasing presence of international students, literature has highlighted the extent to which problems encountered by international students have been attributed to difficulties with language (Samuelowicz 1987). Lumsden and Scott (1987), using a dataset of more than 3,000 students in seventeen universities and colleges in the United Kingdom, found that four percent of non-native English speaking (NNES) students scored substantially lower than native English speaking (NES) students in MC exams. Anderson et al. (1994) found that the final grade of students undertaking an introductory Economics course at the University of Toronto was positively correlated with their high school performance in an English subject.

The language background of teaching assistants (TAs) has also received attention in Economics Education research (Becker and Powers 2001; Belton, et al. 2002; Borjas 2000; Marvasti 2007; Watts and Lynch 1989). The presence of TAs for whom English was a second language was found to have an adverse impact on student grades in several studies (see, for example, Watts and Lynch (1989), Borjas (2000) and Marvasti (2007). However, Becker and Powers (2001) and Belton et al. (2002) reported that the presence of foreign graduate TA had positive effects on students' final grades. These seemingly contradictory findings point to the need for further research into the impact of the language background of TAs on teaching, learning and assessment.

Linguistic issues in MC testing are complex, and cannot unquestioningly be attributed to the English language proficiency of students or instructors. Lumsden and Scott (1987)

hypothesized that failure to understand a key word can reduce Economics students to guessing the meaning of the MCQs. Studies in other disciplines have considered the potential disadvantage to international students of timed, reading-intensive MC tests, and the threat to construct validity posed by this construct-irrelevant aspect of difficulty (Paxton 2000; Smith 2011). However, analysis of the impact of the linguistic complexity of MC items in Economics tests on student performance has been far from conclusive.

The present study investigated the role of linguistic factors on student performance with both methodological and empirical contributions. The remainder of this paper is organised in the following way. In the methodology section, a two-stage approach to analysing student performance in MC questions is proposed. In the first stage, item response theory (IRT) is used to estimate student ability, which then is used to detect the problem of differential item functioning (DIF) using logistic regression methods. In the second stage, a Probit-type model is estimated using the binomial responses of students to non-DIF items to analyse the determinants of student performance, which include both demographic and linguistic factors. Empirical results are then presented and their pedagogical implications discussed.

METHODOLOGY

One common approach to analysing the determinants of students' performance in Economics subjects is to regress the overall total scores of tests or final grades on various driving factors using linear or nonlinear regression specifications. The present article, however, examines the determinants of student performance in *each* MC item; hence it is important to remove the impact of item bias and 'guessing' factors. To do this, item response theory (IRT) can be used to remove 'guessing' factors from student performance and to detect differential item functioning (DIF) questions.

IRT models specify the relationship between unobservable student ability and item parameters (Linden and Hambleton 1997; Osterlind and Everson 2009). The most common IRT model for the analysis of MC questions is a three-parameter model that characterises test items by three parameters: discrimination (*a*), difficulty (*b*), and pseudo-guessing (*c*) (Birnbaum 1968; Lord 1980). The relationship between three item parameters, ability and performance for each item is depicted by an item characteristic curve (ICC).

Figure 1 displays a typical ICC curve for a single item. The ICC suggests that higher ability (on a horizontal axis) is required for a higher probability (on a vertical axis) of getting a correct response to the item. The pseudo-guessing parameter *c* (equal to the value of a lower asymptote) incorporates the phenomenon that on MC questions even the worst students will sometimes guess correctly. The difficulty parameter *b* (equal to the value on the ability axis when the slope of the ICC curve is maximised) measures the item's overall difficulty. The discrimination parameter *a* (equal to the value of the slope of the ICC at the inflection point) captures the extent to which the likelihood of getting a correct answer changes with respect to ability.

[Insert Figure 1 about here]

Walstad and Robson (1997) noted that IRT models produce *invariant* item parameters and student ability; that is, the three item parameters are not dependent on the sample to which items are administered and student ability will be the same from different sets of items except for sampling errors. Details of the IRT estimation are described in Walstad and Robson (1997) and will not be repeated here.

DIF occurs when the performance on an item for students of two groups (or more) differs after conditioning on academic ability (Dorans and Holland 1993); hence DIF items contain

biases and fail to provide reliable and valid test scores. If DIF items are not removed from the test, DIF biases can affect the entire analysis of performance determinants. Many parametric and nonparametric techniques can be used to detect DIF (Teresi 2006; Teresi and Fleishman 2007). In fact, the IRT can be used in the parametric framework to conduct DIF analysis.

In Economics Education research, Walstad and Robson (1997) used a three-parameter IRT model to estimate item parameters and student ability for two separate groups of students categorised by gender. The authors then constructed ICCs for males and females and used the area between ICCs for each item as a measure of the degree of DIF. This approach of detecting DIF is based on the fitting of ICC curves in separate groups, which assumes there are clear categories of the variables to be examined for DIF. For gender, this is a reasonable assumption. However, this approach cannot be applied directly to the analysis of group variables (e.g. age, education, the levels of language proficiency) (Crane, et al. 2004). To deal with group variables, logistic regression (LR) method has been recommended (French and Miller 1996; Swaminathan and Rogers 1990; Teresi and Fleishman 2007).

Following Swaminathan and Rogers (1990), the standard LR model for predicting the probability of a correct response to an item is:

(1) $$P(u=1) = e^{z} \big/ (1+e^{z})$$

where u is the response to the item and

(2) $$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)$$

where θ is the observed ability of an individual student, g represents group membership (i.e. g=1 if the student is a member of group 1, 0 otherwise), τ_i s are parameters to be estimated by maximum likelihood methods.

The difference in the log of the likelihood functions obtained in regressions with and without τ_2 is used to test for uniform DIF and the difference in the log of the likelihood functions obtained in regressions with and without τ_3 is used to test for nonuniform DIF. Uniform DIF refers to the probability of answering the item correctly being greater for one group than others uniformly over all levels of ability; that is, there is no interaction between ability level and group membership in (2). Nonuniform DIF relates to nonuniform differences in the probability of answering the item correctly over all ability levels.

The proposed analytical framework has two stages. In the first stage, a three-parameter IRT model will be employed to produce estimates of item parameters (particularly the difficulty parameter) and student ability. The estimated student ability will be used in the logistic regression to detect DIF. DIF items then will be removed from the test. Data on non-DIF items will be used in the second stage, which involves estimation of a Probit function to establish a statistical relationship between student performance and 'driving' factors. The difficulty parameter and academic ability derived the IRT model are also included in the driving factors. Due to the binomial nature of the dependent variable (i.e. correct or incorrect answers to non-DIF questions of individual students), the Probit model is recommended (Cameron and Trivedi 2009). Formally, the Probit regression of interest is specified as:

$$y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i \tag{3}$$

where y_i is a binomial response to non-DIF questions of student i, $\mathbf{X}_i$ denotes the (Kx1) vector of 'driving' factors, and ε_i is the error term. Note that student ability and item difficulty estimated in the first stage are also present in vector $\mathbf{X}_i$.

It is important to note that this framework differs from existing approaches in a number of ways. Firstly, in contrast to many studies (Becker and Powers 2001; Belton, et al. 2002; Borg

and Stranahan 2002; Borjas 2000) the proposed framework utilises the IRT model to remove DIF biases before examining performance determinants. Secondly, Walstad and Robson (1997) also removed DIF items but their study focused on the total test score, while our framework uses responses to individual MC questions to assess performance determinants. Third, student ability estimated from the three-parameter IRT model is used in conjunction with collegiate grade point average (GPA) as proxies of students' academic aptitude. This use of student ability is particularly useful when international students are present in the dataset and their information related to high school performance is not comparable.[1]

DATA

Data from 952 students who undertook an introductory economics unit at QUT Business School in the first semester in 2011 is used in this study. This unit covers both microeconomics and macroeconomics topics and utilises the conventional two-hour lecture and one-hour tutorial format. Since this unit is one of the eight compulsory core units in business degrees, students enrolled in this unit undertake a range of disciplines (e.g. marketing, accounting, economics, finance, management, etc.) as their major and have diverse language and mathematics backgrounds. Assessment in this unit has three components: a mid-semester MC test (30% of final grade), a research paper (30% of final grade) and a final exam (40% of final grade). The MC test is a one-hour test with 30 items. Students answered on computerized answer sheets which were machine marked. Students were allowed to use hand-

[1] Note that collegiate GPA was found to be the best proxy for students' academic aptitude but should be used in conjunction with other indicators such as high school GPA or scholastic aptitude test scores (Grove, et al. 2006).

held, battery operated calculators (not capable of communication) and a bilingual dictionary (no hand written notes).

Table 1 displays descriptive statistics of academic and demographic data relating to students who participated in this study. The average age of the students is 20.6 years, with ages ranging from 17 to 50 years. Male students account for 49% of the cohort. Nearly 19% are NNES students, 8% hold an international student visa, and 60% completed high school in Queensland.

[Insert Table 1 about here]

There are several linguistic aspects of the questions that can be included in analysis: the length of stems and each question as a whole, the lexical density of stems and questions, and the proportion of vocabulary categorised as high-frequency, academic, and 'off-list' in each MC question. Length is measured by the number of words in a text. The longer questions will involve more reading and processing time, and would thus be potentially more difficult for students for whom English is a second language. Greene (1997), however, reported that item length was not significantly related with student performance in closed-ended questions (in his case, true-false questions). Instead, he found that the easier the readability of a passage (measured by the Flesch Reading Ease score) is, the lower the probability of error on a true-false question about information contained in the passage. The linguistic complexity of a text can be measured in different ways; for example, commonly used readability measures, including the Flesch Reading Ease Readability Formula, use the average sentence length and the average number of syllables in each word in order to calculate the difficulty of a text. While these measures are useful, they do not take into account another essential aspect of readability: the extent to which a text is composed of content words that carry information (usually nouns, verbs, adjectives and adverbs) and words that have a grammatical function

(for example, articles, prepositions and conjunctions). Lexical density refers to the proportion of content words in a text, and affects readability since the higher the lexical density, the more information is packed into the text, and therefore the more difficult it is considered to be. Higher levels of lexical density (typically above .7) are characteristic of written academic texts (Read, 2000). Thus we decided to include lexical density of stems instead of length as a key indicator of linguistic complexity. A web-based vocabulary profiling software program was used to calculate lexical density. This program also identified the percentage of words in each stem and question from the 2,000 most frequently used words in English, the Academic Word List (AWL), and the Off-list words. The Off-list words were low frequency and included highly discipline-specific and culture-specific vocabulary (Cobb 2011). It is important to note that while words including 'economist' and 'consumer' are included in the AWL, more specialised economic terms such as 'stagflation' are categorised as Off-list. The Off-list category was useful for the identification of specialised vocabulary from other disciplines, including 'neo-natal', and culture-specific vocabulary, including 'chiko roll'. It is reasonable to assume that both local and international students are familiar with the frequently used words as well as the AWL; hence the presence of these words would not create extra difficulty to the students. On the other hand, the presence of the non-Economics Off-list words has the potential to cause difficulty to students unfamiliar with vocabulary from other disciplines, and students unfamiliar with the Australian cultural context.

To account for other features that might contribute to the difficulty of the MC questions, we have also included dummy variables that reflect whether questions are dependent on each other and if questions require numeracy, graph or diagram reading skills. We also invited five economists with various teaching experience (i.e. one Professor, two senior lecturers and two

lecturers) to assess the difficulty/clarity of the questions[2] and as a result two questions were categorised as 'unclear' or 'confusing'. Also, a linguist who had previously studied Economics was asked to read through the questions with the focus on the difficulty of distracters and as a result two questions were identified as 'deliberately overly complex'. Hence, another dummy is created to account for the effect of unclear and overly complex question on student performance. Table 2 reports the frequency of these dummy variables in the test.

[Insert Table 2 about here]

The demographic variables used in the analysis included students' age, gender, GPA, language background (NNES or NES) and whether they completed high school in Queensland. The effects of gender on academic performance have been well researched but consensus on the relationship between them has not been reached (Borg and Stranahan 2002; Greene 1997; Hirschfeld, et al. 1995; Lumsden and Scott 1987). Besides using the IRT-generated ability measures, we also included their GPA to capture general academic ability. The binary variable "State" captures the effects related to relocation and being away from family and social network as well as any effects caused by differences in high school curricula between Queensland and other states in Australia.

[2] The rankings of difficulty by the five economists were not consistent and their correlations with the difficulty level estimated by the IRT model were low. However, the ranking by the economist who was the coordinator and the lecturer of the unit was most highly correlated with the IRT-based difficulty estimates (i.e. correlation coefficient of 0.46). In this study, only the IRT-based measure of difficulty was considered.

The focus of the study is on linguistic aspects of MC questions and it is possible that the impact of linguistic features may depend on the language background of students. To investigate these differential effects, interaction terms of the four linguistic dummy variables (i.e. Off-list, Deliberately complex/potential confusing construction, Graph/Numeracy and Dependence) with Language were added to the model. An interaction term between Difficulty and Dependence was also included post-hoc (to be discussed later).

ESTIMATIONS AND RESULTS

The three-parameter IRT model was estimated in Stata using the openirt module (Zajonc 2009). Estimates of student ability from this three-parameter IRT model were then used to detect uniform and nonuniform DIF items in Stata's DIFd module (Crane, et al. 2005). No uniform and nonuniform DIF items were detected in the test in relation to the English background of the students; hence the complete set of responses of 952 students to 30 questions was analysed in the second stage.[3]

The second stage specified the complementary log-log equation due to the dependent variable has a high proportion (more than 70%) of value one (i.e. correct responses) (Cameron and Trivedi 2009). We estimated three different models: (i) Model 1: the full model in which all independent variables and interaction terms were present; (ii) Model 2: a version of Model 1 without GPA; and (iii) Model 3: a version of Model 1 without Ability. Models 2 and 3 were examined to see if the use of both GPA and Ability as in the full model (Model 1) would improve model performance. Chi-square tests of the nested models were conducted and test

[3] We also conducted the DIF analysis with respect to gender and results showed no item of the 30 questions were DIF.

results preferred Model 1. Table 4 displays the test results. In the following sections, we focus our discussion on the results of Model 1. Table 3 also reported the estimates of Model 1's specifications for two groups: NNES and NES students. Table 5 showed the estimates of the partial effects at the mean.

[Insert Tables 3, 4 and 5 about here]

As shown in Table 4 most of the explanatory variables were statistically significant except Age, Gender, State, and three of the four interaction terms between Language and the linguistic dummies. The signs of coefficients were mostly consistent with expectations. GPA and student ability have positive relationships with students' performance. Difficult, complex items, lexical density of stems, language background (NNES status) and the use of graphs, diagrams or tables have negative relationships with student performance. The difficulty and ability levels had the strongest partial effects at the mean.

Note that Ability and GPA are moderately correlated (correlation coefficient=0.60). But as shown in table 4, nested model comparisons show that the presence of both Ability and GPA improved model performance. There are several interesting observations. The impact of Ability on student performance is much stronger than GPA (i.e. the partial effects of Ability and GPA are 0.138 and 0.006). This is not unexpected as the IRT-estimated ability measure is specific to the test, whereas GPA tends to represent general academic performance through the entire semester in which the test contributed a minor portion[4].

[4] At QUT, students undertake four units in a standard semester. This test carries 30% of the final grade of one unit; hence its contribution in GPA is reasonably minor.

The second observation relates to the effect of Gender on student performance in the nested models. Table 6 presents the coefficients of Gender and their p-values produced by the three specifications of the Probit model. Models 1 and 2 which include the IRT-generated Ability reported a negative relationship between student performance and gender, suggesting that females performed better than males, although the relationship is statistically non-significant. However, when Ability was taken out of the model, the sign of the coefficient changed to positive and became highly significant. This suggests that the better performance of male students can be accounted for their superior discipline cognitive ability. This finding is consistent with an observation that an average ability score estimated by the IRT model is higher for male students than for female students.[5] In other words, after controlling for ability, gender difference disappears. Also note that this test did not have any questions that exhibit DIF with respect to gender.

[Insert Table 6 about here]

The third observation is that given Model 1 was tested as being superior to models 2 and 3, one could argue that it is important to include specific discipline cognitive ability (i.e. the IRT-based ability as in this empirical study) in modelling student performance and the resulted estimation can be biased due to model misspecification if this key variable is omitted.

It is also observed that those questions that require students to understand graphs or perform calculations were found to be harder, as shown by the negative coefficients of this variable. There are at least two important interpretations of this finding. First, these questions may test

[5] The difference in the average ability score between male and female students was 0.045. However, a t test does not reject a hypothesis of equal mean values.

graph reading or calculation skills rather than testing the understanding of Economics concepts. If this is the case, the validity of these items is questionable. Hence, test constructors should pay attention to answering these two questions: what is really being tested and what is the question designed to test. Second, if applying the economic concepts into graphs or calculating are important skills to be examined, then this finding support a view that MC questions can be designed to make exams not only more challenging but capable of assessing high level understanding and application.

As expected, NNES students performed worse than NES counterparts, and both groups performed worse in those questions that were more linguistically complex. Test developers, therefore, should be aware that the wording structure of information, particularly the lexical density of the stems of MC questions, has a significant impact on student performance. While it may be tempting to include contextual features that make the questions appear more authentic and interesting to students, based on the findings from this study, test developers should avoid the use of long, overcomplicated stems and the provision of information that is unnecessary for the testing of a particular concept or application, if NNES students are not to be disadvantaged. There are also implications for the provision of concurrent language support for NNES students studying Economics units.

One would also expect high lexical density would adversely affect NNES more than NES students. But the coefficient of lexical density of stems was negative for NES and was positive for NNES students. The results also showed the positive coefficient of the interaction term between Language and lexical density of stems. Is it possible that NES, working in their first language, tend to skip through stems packed with information and as a result overlook key information essential to answering the question. Undoubtedly, this hypothesis is important and should be investigated further.

The use of local, culture-specific terms and specialised academic vocabulary from other fields (non-Economic Off-list words) affected student performance between NES and NNES groups in different ways, as shown by positive and negative coefficient values for the two respective groups. While it is not certain why the presence of Off-list words enhances the performance of NES students, an important implication of this finding is that adding Off-list vocabulary may put NNES at an unnecessary disadvantage if this knowledge is not relevant to the concept being examined and thus also constitutes a threat to construct validity.

The test contains five pairs of dependent questions, i.e. a question requires information or understanding of content or context from a previous question. It is possible that if the previous question is a difficult one, students will have more problems answering the latter question. In an attempt to explore this hypothesis, we added an interaction term between Dependence and Difficulty. The negative coefficient of this interaction variable shows that if a student found a particular question difficult, s/he would also have less chance of answering related questions correctly.

CONCLUSION

This paper revisited the analysis of determinants of student performance on MC questions with the focus on the linguistic aspects of questions and the language background of students. We proposed a two-stage analytical framework. In the first stage, the IRT theory is deployed to estimate the academic ability and the difficulty level of questions. The estimated ability is subsequently used in the logistic regression model to detect DIF items. The second stage establishes relationships between performance and explanatory variables in the Probit-type model using data on non-DIF items. Instead of regressing total grade of the test, we proposed to use responses to each non-DIF questions as the dependent variable.

The empirical results of this study have several important implications for test constructors. First, the linguistic complexity of questions affects student performance and can be a source of discrimination between different demographic groups of students. Second, the inclusion of questions that require the ability to read graphs and/or perform calculations using information from tables can be used to test in-depth learning but can constitute a threat to validity if these skills are not directly related to disciplinary knowledge.

The empirical study also yielded two important methodological implications for further research in Economics Education in particular and education research in general. First, the difficulty measure estimated from the three-parameter IRT model performed well on explaining student performance on MC questions. We recommend that future research in Economics Education should consider the use of an IRT-based difficulty parameter besides (or instead of) difficulty levels ranked subjectively by instructors as has been done in other studies (Smith, et al. 1994). The second implication concerns the use of discipline specific measure of academic aptitude in the analysis of performance determinants. The empirical study provides evidence that the use of IRT-based ability explained well the variation of student performance. One can view IRT-based ability as discipline specific measure of academic aptitude and the omission of such discipline specific ability indicator may cause misspecification problems. As shown in the empirical study, the use of only GPA reported that male students performed better than female students but this relationship disappeared when IRT-based ability was included into the model.

TABLES

Table 1: Descriptive statistics of main variables

Students' academic-social-economic background	Average	Min	Max
Age	20.63	17	50
Gender (proportion of male students)	0.49		
Language background (proportion of nonnative English speaking students)	0.19		
International student	0.08		
State (proportion of students who did high school in Queensland)	0.61		
Collegiate grade point average (GPA)	4.72	1	7
Students' ability (estimated from the 3-P IRT model)	0.02	-5.00	5.00

Table 2: Aspects of additional potential difficulty

Variables	Frequency (out of 30 items)
The presence of Non-Economics off-list words (Dummy 1)	4
Deliberately over-complex/potentially confusing question construction (Dummy 2)	4
Presence of tables, diagrams and graphs (Dummy 3)	7
Dependence of items (Dummy 4)	10
Lexical density of stems (average density)	0.59

Table 3: Results of complementary log-log models (specifications 2 and 3)

Variables	Whole sample		Native English Speakers Group		Nonnative English Speakers Group	
	Coeff.	St. error	Coeff.	St. error	Coeff.	St. error
Constant	0.235	0.091	0.252	0.099	-0.092	0.192
Difficulty	-0.492*	0.011	-0.496*	0.012	-0.475*	0.024
The presence of Off-list words (Dummy 1)*	0.118*	0.030	0.118*	0.030	-0.029	0.063
Deliberately over-complex/potentially confusing question construction (Dummy 2)	-0.141*	0.034	-0.134*	0.035	-0.228*	0.072
Presence of tables, diagrams and graphs (Dummy 3)	-0.121*	0.031	-0.121*	0.031	-0.177*	0.064
Dependence (Dummy 4)	0.250*	0.033	0.244*	0.033	0.282*	0.068
(Difficulty)x(Dummy 4)	-0.059**	0.021	-0.068*	0.024	-0.020	0.049
Lexical Density of Stems	-0.279*	0.091	-0.296*	0.092	0.358**	0.189
Age	0.001	0.002	0.001	0.003	0.001	0.005
Gender (Male)	-0.013	0.017	-0.017	0.019	0.005	0.040
State (did high school in QLD)	0.001	0.020	0.001	0.023	0.003	0.043
Collegiate grade point average (GPA)	0.018***	0.009	0.017	0.010	0.021	0.021
Ability (the three-parameter IRT model)	0.390*	0.009	0.390*	0.011	0.388*	0.022
Language Background	-0.267***	0.117				
(Language)x(Lexical Density)	0.571*	0.205				
(Language)x(Dummy 1)	-0.146***	0.068				
(Language)x(Dummy 2)	-0.056	0.075				
(Language)x(Dummy 3)	-0.058	0.070				
(Language)x(Dummy 4)	0.007	0.074				
*, **, ***, ****: significant at 0.5%, 1%, 5% and 10% LOS.						

Table 4: Chi-square tests

Null hypothesis	Chi-square test statistics	Critical value at 0.5%
Model 1 is preferred to Model 2 (1 degree of freedom)	357	7.879
Model 1 is preferred to Model 3 (1 degree of freedom)	1960	7.879

Table 5: Partial effects

Variable	Partial Effects
Difficulty	-0.174
Non-Economics Academic Vocabularies (Dummy 1)	0.041
Tricky Question (Dummy 2)	-0.051
Combination of Numeracy, Graph & Literacy (Dummy 3)	-0.043
Dependence (Dummy 4)	0.087
Collegiate grade point average (GPA)	0.006
Ability (the three-parameter IRT model)	0.138

Table 6: Coefficient and p-value of the variable Gender

	Coefficient	p-value
Model 1	-0.013	0.461
Model 2 (Model 1 without GPA)	-0.018	0.277
Model 3 (Model 1 without Ability)	0.234	0.000

FIGURES

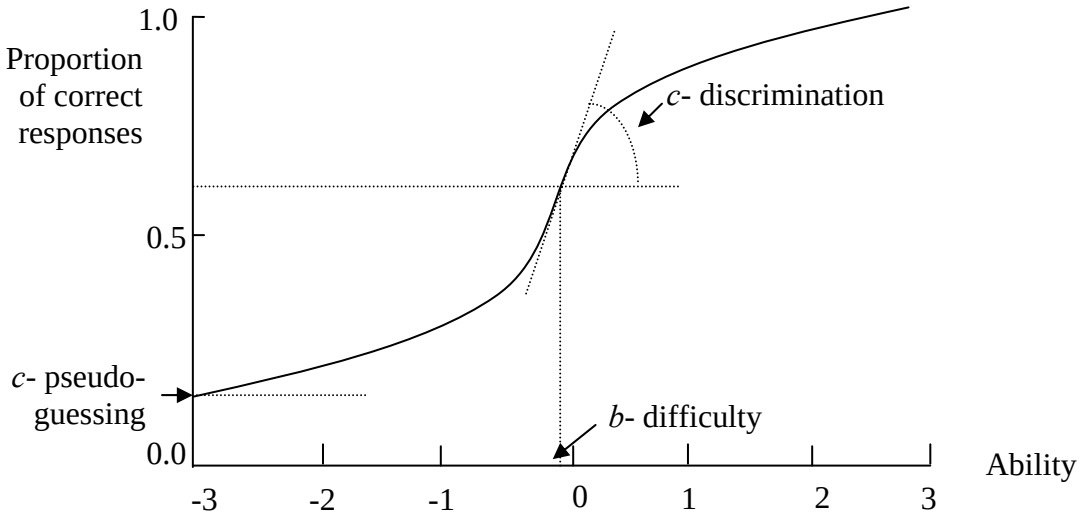

Figure 1: Item chateristic curve in three-parameter IRT models

ACKNOWLEDGEMENT

We express our gratitude to Dr. Louisa Coglan, Professor Uwe Dulleck, Professor Tim Robinson, and Dr. Dipanwita Sarkar (in an alphabetic order) for their involvement in this study. Special thanks also go to Ms. Michele Smith and Ms. Katalina Mok for data collection and Mr. Cuong Vu for research assistance work. We would like to express our thanks to participants at 2012 Australasian Teaching Economics Conference for useful discussions.

REFERENCES

ABS. 2011. "Australian Social Trends December 2011 International Students." ABS catalogue no. 4102. Australia Bureau of Statistics: Canberra.

Anderson G., Benjamin D., and Fuss M.A. 1994. "The Determinants of Success in University Introductory Economics Courses." The Journal of Economic Education, 25:2, pp. 99-119.

Becker W., Highsmith R., Kennedy P., and Walstad W. 1991. "An Agenda for Research on Economic Education in Colleges and Universities." The Journal of Economic Education, 22:3, pp. 241-50.

Becker W.E. and Powers J.R. 2001. "Student Performance, Attrition, and Class Size Given Missing Student Data." Economics of Education Review, 20:4, pp. 377-88.

Belton F., Masanori H., and Bruce A.W. 2002. "Foreign Gtas Can Be Effective Teachers of Economics." JOURNAL OF ECONOMIC EDUCATION, 33:4, pp. 299-325.

Birnbaum A. 1968. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability," in Statistical Theories of Mental Test Scores. F.M. Lord and M.R. Novick eds. Mass.: Addison-Wesley: Reading.

Borg M.O. and Stranahan H.A. 2002. "Personality Type and Student Performance in Upper-Level Economics Courses: The Importance of Race and Gender." The Journal of Economic Education, 33:1, pp. 3-14.

Borjas G.J. 2000. "Foreign-Born Teaching Assistants and the Academic Performance of Undergraduates." American Economic Review, 90:2, pp. 355-59.

Cameron A.C. and Trivedi P.K. 2009. Microeconometrics Using Stata. Texas: A Stata Press Publication.

Chan N. and Kennedy P.E. 2002. "Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and "Equivalent" Constructed-Response Exam Questions." Southern Economic Journal, 68:4, pp. 957-71.

Cobb T. 2011. "Web Vocabprofile." http://www.lextutor.ca/vp. Accessed on 15 November

Crane P., Gibbons L.E., Jolley L., and van Belle G. 2005. "Difd V. 1.0, ." University of Washington: Seattle, WA.

Crane P.K., van Belle G., and Larson E.B. 2004. "Test Bias in a Cognitive Test: Differential Item Functioning in the Casi." Statistics in Medicine, 23:2, pp. 241-56.

Dorans N.J. and Holland P.W. 1993. "Dif Detection and Description: Mantel-Haenszel and Standardization," in Differential Item Functioning. P.W. Holland and H. Wainer eds. Hillsdale, NJ: Erlbaum, pp. 35-66.

French A.W. and Miller T.R. 1996. "Logistic Regression and Its Use in Detecting Differential Item Functioning in Polytomous Items." Journal of Educational Measurement, 33:3, pp. 315-32.

Greene B. 1997. "Verbal Abilities, Gender, and the Introductory Economics Course: A New Look at an Old Assumption." The Journal of Economic Education, 28:1, pp. 13-30.

Grove W.A., Wasserman T., and Grodner A. 2006. "Choosing a Proxy for Academic Aptitude." The Journal of Economic Education, 37:2, pp. 131-47.

Hirschfeld M., Moore R.L., and Brown E. 1995. "Exploring the Gender-Gap on the Gre Subject Test in Economics." JOURNAL OF ECONOMIC EDUCATION, 26:1, pp. 3-15.

Linden W.J.v.d. and Hambleton R.K. 1997. Handbook of Modern Item Response Theory. New York: Springer.

Lord F. 1980. Applications of Irt to Practical Testing Problems. Hillsdale, N.J.: Erlbaum.

Lumsden K.G. and Scott A. 1987. "The Economics Student Reexamined: Male-Female Differences in Comprehension." The Journal of Economic Education, 18:4, pp. 365-75.

Marvasti A. 2007. "Foreign-Born Teaching Assistants and Student Achievement: An Ordered Probit Analysis." The American Economist, 51:2, pp. 61-71.

Orhan K., Fathollah B., and Thomas T. 2009. "Factors Affecting Students' Grades in Principles of Economics." American Journal of Business Education, 2:7, pp. 25.

Osterlind S.J. and Everson H.T. 2009. Differential Item Functioning. Los Angeles: SAGE.

Paxton M. 2000. "A Linguistic Perspective on Multiple Choice Questioning." Assessment and Evaluation in Higher Education, 25:2, pp. 109-19.

Read J. 2000. Assessing Vocabulary. Cambridge: Cambridge University Press.

Samuelowicz K. 1987. "Learning Problems of Overseas Students: Two Sides of a Story." Higher Education Research and Development, 6, pp. 121-33.

Siegfried J.J. and Kennedy P.E. 1995. "Does Pedagogy Vary with Class Size in Introductory Economics?" The American Economic Review, 85:2, pp. 347.

Smith C. 2011. "Examinations and the Esl Student- More Evidence of Particular Disadvantages." Assessment and Evaluation in Higher Education, 36:1, pp. 13-25.

Smith N.F., Wood L.N., Gillies R.K., and Perrett G. 1994. "Analysis of Student Performance in Statistics." Education Research Group

Swaminathan H. and Rogers H.J. 1990. "Detecting Differential Item Functioning Using Logistic-Regression Procedures." Journal of Educational Measurement, 27:4, pp. 361-70.

Swope K.J. and Schmitt P.M. 2006. "The Performance of Economics Graduates over the Entire Curriculum: The Determinants of Success." The Journal of Economic Education, 37:4, pp. 387-94.

Teresi J.A. 2006. "Different Approaches to Differential Item Functioning in Health Applications. Advantages, Disadvantages and Some Neglected Topics." Medical Care, 44:11 Suppl 3, pp. S152-S70.

Teresi J.A. and Fleishman J.A. 2007. "Differential Item Functioning and Health Assessment." Quality of Life Research, 16:1, pp. 33-42.

Walstad W.B. and Robson D. 1997. "Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics." The Journal of Economic Education, 28:2, pp. 155-71.

Walstad W.B. 1998. "Multiple Choice Tests for the Economics Course," in Teaching Undergraduate Economics: A Handbook for Instructors. W.B. Walstad and P. Saunders eds. New York: McGraw-Hill, pp. 287-304.

Watts M. and Lynch G.J. 1989. "The Principles Courses Revisited." The American Economic Review, 79:2, pp. 236.

Watts M. and Schaur G. 2011. "Teaching and Assessment Methods in Undergraduate Economics: A Fourth National Quinquennial Survey." The Journal of Economic Education, 42:3, pp. 294-309.

Zajonc T. 2009. "Openirt." Harvard University: Cambridge, MA.