ASSESSING ATTAINMENT OF BLOOM'S COGNITIVE LEVELS USING TESTLETS AND MULTI-CATEGORICAL IRT

Helen Chow Min Kim
Seletar Institute
Ministry of Education
Singapore


Abstract
This study attempts to demonstrate how alternate forms of assessment involving a diagnostic approach could be used in the context of IRT applications. The items were written in clusters of 3 based on a narrowly defined common content area in the area of microeconomics. Each cluster consists of items testing progressively higher Bloom's cognitive levels. Due to context effects, the item clusters cannot be analysed using standard 1-, 2- or 3-Parameter Logistic models. Instead, the Rating Scale Model is used. The study demonstrates in detail, the manner in which examinees respond to items testing different cognitive levels based on item thresholds. Diagnostic and remedial implications for the use of testlets are discussed.

Introduction and Rationale of the Study

Within the classroom there is a great deal of individual variability as regards the extent of mastery of educational objectives in the cognitive domain. Lyon & Gettinger (1985) examined the differences in time by 7th and 8th graders for learning for three types of learning tasks from Bloom's taxonomy of educational objectives: knowledge, comprehension and application, and found large differences in the mastery rates. Their study reported a clear trend toward decreased speed and depth of mastery from knowledge to higher-order tasks, a result which is consistent with the theory underlying Bloom's taxonomy.
that each class of learning in the taxonomy is more complex than the previous one(s).
The above-mentioned individual differences in rates of learning could be partly attributable to differential exposure to task-types. That is, schools may provide an optimum of experience with lower-order (knowledge) tasks but increasingly less exposure to higher-order tasks. Under these conditions, students would likely become efficient learners of knowledge or recall tasks but remain less efficient at mastering higher-order tasks with which they are less familiar.

Of further concern to the educator is the precedence of current external assessments for accountability over other educational purposes; this has reportedly led to, among others, a de-emphasis on higher-order thinking skills (Suarez & Gottovi, 1992). In Singapore, student performance in the GCE 'O' and 'A' level examinations now serve the purpose of identifying successful and less successful schools as implied in the annual school ranking exercise. As reported by Suarez and Gottovi (1992), negative consequences (such as poor school ranking and adverse publicity) for poor performance give rise to teachers teaching to the tests and coaching students in test-wiseness; less effort is devoted to developing higher-order thinking in students. The study recommended alternative forms of assessment for improving educational quality.

In view of research evidence, an important pedagogical consideration is the assessment of learning objectives with greater emphasis on the measurement of higher-order cognitive skill mastery. This calls for a criterion-referenced form of measurement in which a student's performance would not be compared (or referenced) to the performance of other students; rather it would be compared to descriptions of what the student could or could not do with respect to a well-defined domain of cognitive behaviour. This allows individual students to be rewarded on the basis of their attainment relative to the objective of instruction, not their standing relative to their peers. It also permits specific identification of a student's areas of non-mastery; that is, it provides formative evaluation. For example, this criterion-referenced approach increases the teacher's awareness of individual student's level of attainment of educational objectives and thus provides information for instructional planning such as identifying students for remedial or supplementary work in enhancing higher-order skills.

Bloom's Taxonomy and the assumption of cumulative hierarchy

Gilbert (1992) suggested that teachers use a taxonomy as a template for writing items that stimulate students to think at higher cognitive levels. Bloom's Taxonomy was one of those recommended as questions based on this taxonomy had been found effective in improving students' cognitive skills. Also noted was that a mixture of questions at various levels of the taxonomy may result in the greatest learning at higher levels.1
The major legacy of Bloom's Taxonomy has been in its definition of hierarchical levels of student learning which can be used to determine the extent to which educators emphasise both lower- and higher-order thinking behaviours. For this reason, this taxonomic model of learning has become extensively used by curriculum designers and educators. If one considers that even the University of Cambridge Examinations Syndicate uses the Taxonomy as the test blueprint for the GCE examinations, its importance to the Singapore teacher in understanding and using it in classroom assessments becomes clear. (For example, the setters of the multiple-choice component in GCE 'A' Level Economics employ a table of specifications developed in accordance with Bloom's Taxonomy.)

A common reason for incorporating the taxonomy into test specifications, is to ensure that 'higher order' cognitive processes are assessed. Bloom's Taxonomy assumes that cognitive functioning can be represented by a hierarchical structure from the lowest level of functioning (Knowledge) to higher, more complex levels, such as Application and Analysis.

Given the widespread employment of Bloom's Taxonomy in formulating objectives in a wide range of curricular areas in the classroom, there has been, understandably, much interest and frequent attempts at validating (or otherwise) the Taxonomy's inherent assumption of cumulative hierarchy of complexity in its structure. Over the decades since the publication of Handbook I, a plethora of research work had been initiated to validate the Taxonomy or its variations and to verify the existence of its asserted levels. They have yielded mixed results.

One underlying assumption in the Taxonomy is that the higher the complexity of the task, the more difficult it is. However, several researchers have found that difficulty and complexity of tasks need not necessarily correlate. Indeed, certain Knowledge tasks may be more complex than certain Analysis tasks. For example, the Knowledge-level task of reciting Marc Antony's approbrious speech after Julius Caesar's death in Shakespeare's Julius Caesar is far more difficult than an Analysis-level analogy question such as those found in some graduate school admission tests (Seddon, 1978).

At least one other researcher has postulated that a more satisfactory classification would take the form of six parallel taxonomic categories in place of Bloom's linear model (Ormell, 1974). Yet others (Madaus et al., 1973) proposed that the Taxonomy was better represented by a Y-shaped structure, in which the stem of the Y began at Knowledge and continues to Comprehension and from there, one branch goes to Application and then to Synthesis while the other branch goes to Analysis. Some philosophers of education have contended that Knowledge should not be separate from the other cognitive levels. Their argument is that Knowledge in any important sense presupposes intellectual skills. In fact, whether knowledge of specifics, as narrowly defined in the Taxonomy, should even be considered an
educational objective, has been questioned. Indeed, there is no shortage of critics who find fault with the order or even the distinction between the taxonomic levels.

Efforts to validate the hierarchical structure of the Taxonomy can be no better than the precision of classification of the test items used to collect the data. In this study, raters or judges, were schoolteachers who had been trained in the Taxonomy. (Indeed, for them, Bloom’s Taxonomy was the only one which had been taught to them as trainee teachers in the Institute of Education.) Each had, for more than a decade, written items using specification guidelines based on the Taxonomy. Despite their familiarity with Bloom's definitions, the raters did not agree in many instances. Also, if the same behaviour or test item can be placed into two or more categories, the possibility of achieving a clear hierarchy is lessened substantially. Unsuccessful item categorisation (as evidenced by inter-rater disagreement) would render any attempt at verifying the hierarchy, futile, in any case.

To compound the problem further, there is the gap between items as written and items as experienced by the test taker. Inter-judge (dis)agreement aside, one still cannot know the mental processes of the test taker in encountering an item. What may have been written at the Analytical level could well turn out to present itself merely as a Knowledge task to the test taker if the test taker had been over-exposed to items of a very similar nature or format. In other words, there lies the inherent uncertainty about the exact mental processes used when test takers answered the items.

On the flip side, some research efforts have supported the hierarchical structure of the Taxonomy. Notably, Kropp and Stoker’s (1966) findings showed a generally decreasing mean performance on successive levels of the Taxonomy, thus supporting the hypothesis of increasing difficulty. In the discussion of their results, however, they conceded that the assumption of a correlation between item complexity and item difficulty is not without its flaws. Further support emerged from a reanalysis of the Kropp and Stoker data by Smith (1970); in particular, the categories Comprehension, Application and Analysis were consistently ordered, as predicted, although they failed to confirm Knowledge as the first level in the hierarchy.

A later reanalysis of the Kropp and Stoker data by Madaus et al. (1973) yielded the Y-structure referred to earlier, weakly supporting the hierarchy. The validity of the Y-structure was given some support in the work of Benson et al. (1992) which examined the value of the Taxonomy for assisting academic mentors when guiding students in writing literature reviews. This study revealed a lack of differential ratings between Analysis and Synthesis scores on items designed to tap the said levels.

Kreitzer and Madaus' (1994) summarised other work based on the Kropp and Stoker data which gave rise to similar supportive findings on the hierarchy assumption, particularly the work of Hill and MacGaw (1981).
There were some investigations, they summarised, which did not employ the Kropp and Stoker data, and yielded results supporting the hierarchical structure; yet others did not. As Seddon (1978) put it, no one has been able to prove nor disprove the existence of the psychological properties of the Taxonomy. The research record today is no different. Lack of concurrence among researchers over the past four decades, does not therefore, in any way, change the Taxonomy's status as a useful tool in the eyes of many educators. Perhaps, the exercise to discredit or uphold the Taxonomy should take a back seat, as it shows up an excessively serious view of the approach taken by the authors of the Taxonomy.

For many teachers, the Taxonomy facilitated the construction of criterion-referenced tests in the subjects that they teach. These teachers are not at all interested in whether the taxonomic categories are hierarchical nor are they unduly worried about whether one category is more difficult than another. They are interested in ensuring that important cognitive skills are in fact covered sufficiently in the tests. The use of the taxonomic categories increases efficiency in writing test items. That is, being able to place an objective into a particular category enables the test designer to know what type of items are likely to be appropriate. Obviously, the taxonomy is not without its flaws. Equally obvious however, is the important role which it has played and is still playing, in prompting much thought and discussion about the purpose of education and the methods of assessing its accomplishment.

The IRT Measurement Model

This study is based on the measurement philosophy inherent in Item Response Theory (IRT). IRT was chosen over the traditional approach to measurement, called the Classical Test Theory (CTT) because of its many advantages:

- Differences in item and person strengths are estimated independently of one another, i.e. item parameters and person parameters are separable. This separability ensures objectivity in measurement.
- Persons and items in the measurement instrument are posited on the same linear continuum. Items and persons are measured on an interval scale with a common unit, the logit, derived from a function of the probability of a correct answer, given person and item parameters. The logit correct equals the difference between person ability and item difficulty.
- The calibration procedure is independent of the sample to which the test is administered, i.e. it is invariant over the population. The measures of persons are also 'test-free', i.e. it does not matter which selection of items is used to estimate them.

Although IRT consists of a family of models, this study will only employ two models, the Rasch (1980) model and the Rating Scale (Andrich, 1978) model. In the Rasch model, items are characterised as
differing from one another with respect to difficulty only and all items are assumed to possess the same capacity to discriminate among examinees. Since in this model, the examinee’s response is categorised as either correct or incorrect, the Rasch model is known as a binary or dichotomous model.

However as testlets are used in the later part of the study, an IRT model which makes use of polytomous data and assesses information from all item options is also required. For this reason, the Rating Scale (RS) model is also employed.

The use of testlets

A recent measurement technique involving the use of item clusters or testlets that tap higher-order thinking, has evolved. The items in the testlets are designed to test various higher-order thinking skills (HOTS) and are therefore of different difficulty levels. In this form of testing items are inter-dependent and reliability would be overestimated should the items be treated as separate or unrelated. The purpose of employing testing units that are larger than the test item is to reduce item dependency. However, there would still be dependency between testlets (Wainer & Lewis, 1990).

Several studies have been carried out on testlets, including testlet reliability (Sireci et. al., 1991), testlet validity (Wainer et al., 1992) and testlet-based computerised mastery testing (Sheehan & Lewis, 1992). However, there is still a need to make IRT and testlets more meaningful to the teacher interested in assessing HOTS.

This study attempts to apply IRT in assessing learning outcome proficiency in 'A' Level Economics. Design of the instrument is based on Wainer & Kiely's (1987) definition of the testlet as a cluster of items related to a common content area and consisting of a set of predetermined response paths. In this study, a testlet is defined as a group of 3 hierarchical items testing three levels of Bloom's Taxonomy of Educational Objectives.

Method

An example of a testlet designed for this study is given below.

Level 1 : Knowledge and/or Comprehension (K/C)
An inferior good is one whose income-elasticity of demand Ais less than unity.*B is negative.
Cdecreases when income rises. Dincreases when income rises.

Level 2 : Application (AP)
As a result of an increase in income from $1000 to $1200, my expenditure on food rises by 10%. My coefficient of income elasticity
of demand on food is
A2B1*C0.5D0.2

Level 3 : Analysis (AN)
The demand function for good A is given by the following function :
QA = 100 - PA + 2PB - 0.2Y
where
QA = quantity demanded of good A
PA = price of good A
PB = price of good B
Y = income

From the expression for the demand function, it is possible to conclude that
Agood A is a complement to good B.
*Bgood A is an inferior good.
Cgood A is a Giffen good.
Dgood B is a normal good.

Each testlet consists of 3 items arranged hierarchically, each item directed at a specific educational objective in Bloom's taxonomy (cognitive domain).
In the preliminary stage of the study, a 50-item instrument, clustered into 12 testlets, each containing not less than 3 items, was administered to 436 students of 'A' Level Economics. The items were scored individually and submitted to a Rasch analysis. Their classical facility (p-values) and discrimination indices were also obtained. These were used to select 10 testlets of 3 items each for the final analysis. Care was taken to ensure that items in each testlet tested the 3 levels of Bloom's Taxonomy specified earlier. Among the 20 items which were rejected were those which for one reason or another, did not fit the Rasch model (e.g. some items tested concepts which had not yet been taught in one school) or had very low classical discrimination indices. There were also a number of items which had to be dropped as they did not perform consistently across schools. For example, some items which had been written to tap a higher-order cognitive skill (Analysis) had very high p-values in one school. It turned out that students in that school had been drilled in and exposed to items of a very similar nature. What had intended to be Analysis (or Level 3) items had become mere recall (Level 1) items.
Examinee responses (1=correct, 0=wrong) for the final 30 items were submitted to a binary factor analysis using the MicroFACT computer programme. Item scores were then linearly summed for testlet scores. Factor analysis based on polychoric correlations was carried out on the testlet dataset to check for unidimensionality. The dataset was
fitted into Andrich's (1978) Rating Scale model using the QUEST computer programme, which employs a generalised form of Masters' (1982) Partial Credit Model in the estimation process:

where (n is the proficiency level of person n,
wij is the score assigned to step j in item i to allow for estimation where null categories are involved, and
(i and (ij characterise the difficulty or location parameter of item i.
The following constraints are applied:

If the ( parameters are constrained so that (1j = (2j = (3j = ..., then the Partial Credit Model becomes the Andrich Rating Scale Model.

Results

The response patterns (and the percentage of examinees with such response patterns) is shown in Table 1. As the items in each testlet are arranged hierarchically the four expected scoring patterns of the testlets are:

<table>
<thead>
<tr>
<th>K/C</th>
<th>Item 1</th>
<th>API</th>
<th>Item 2</th>
<th>AN</th>
<th>Item 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>i)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>all items incorrect</td>
</tr>
<tr>
<td>ii)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>01st corr; 2nd and 3rd incorrect</td>
</tr>
<tr>
<td>iii)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>01st, 2nd correct; 3rd incorrect</td>
</tr>
<tr>
<td>iv)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>all items correct</td>
</tr>
</tbody>
</table>

Table 1. Response patterns and percentage responding for 10 testlets

<table>
<thead>
<tr>
<th>Pattern 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>1.2</td>
<td>7.0</td>
<td>6.0</td>
<td>11.0</td>
<td>8.1</td>
<td>3.9</td>
<td>3.2</td>
<td>2.5</td>
<td>8.6</td>
</tr>
<tr>
<td>100</td>
<td>2.3</td>
<td>14.3</td>
<td>16.0</td>
<td>24.1</td>
<td>11.8</td>
<td>11.2</td>
<td>16.8</td>
<td>12.9</td>
<td>12.4</td>
</tr>
<tr>
<td>110</td>
<td>17.2</td>
<td>29.9</td>
<td>33.3</td>
<td>18.9</td>
<td>23.0</td>
<td>15.2</td>
<td>33.3</td>
<td>41.1</td>
<td>27.6</td>
</tr>
<tr>
<td>111</td>
<td>52.6</td>
<td>17.5</td>
<td>21.5</td>
<td>15.2</td>
<td>23.0</td>
<td>46.5</td>
<td>25.7</td>
<td>27.8</td>
<td>22.4</td>
</tr>
<tr>
<td>73.2</td>
<td>68.7</td>
<td>76.9</td>
<td>69.2</td>
<td>78.7</td>
<td>79.1</td>
<td>84.4</td>
<td>71.0</td>
<td>78.6</td>
<td></td>
</tr>
<tr>
<td>001</td>
<td>4.7</td>
<td>3.3</td>
<td>2.1</td>
<td>5.8</td>
<td>3.7</td>
<td>3.9</td>
<td>2.3</td>
<td>1.6</td>
<td>4.0</td>
</tr>
<tr>
<td>010</td>
<td>1.6</td>
<td>7.9</td>
<td>8.6</td>
<td>10.0</td>
<td>9.7</td>
<td>3.9</td>
<td>6.9</td>
<td>6.410</td>
<td>3.6</td>
</tr>
<tr>
<td>101</td>
<td>11.6</td>
<td>15.0</td>
<td>9.0</td>
<td>11.9</td>
<td>9.3</td>
<td>11.1</td>
<td>19.3</td>
<td>5.310</td>
<td>10.3</td>
</tr>
<tr>
<td>011</td>
<td>8.8</td>
<td>5.1</td>
<td>3.5</td>
<td>4.0</td>
<td>8.4</td>
<td>2.8</td>
<td>2.8</td>
<td>2.3</td>
<td>4.4</td>
</tr>
</tbody>
</table>

N=436

Aberrations in the response patterns (001, 010, 101, 011) are possible due to guessing or other random responses. From Table 1, it is noted
that departures from the expected response patterns account for less than one-third of examinees' responses. Such aberrations in the response pattern raise suspicions about the assumption of cumulative hierarchy in Bloom's Taxonomy. Table 2 shows the classical facility (p-values) and discrimination Indices for each item. Within each testlet, the p-values decrease, as expected, as they are arranged in ascending taxonomic order, from Level 1 (K/C) to Level 3 (AN). At a cursory glance, the information from both tables point to the apparent existence of a hierarchy within testlets in the instrument but on closer examination, such a conclusion becomes suspect. It is noteworthy that for testlets 1 and 6, the p-values for the 2nd (Level 2, AP) and 3rd (Level 3, AN) items are very close, indicating a possibility that, for these testlets at least, Madaus' (1973) et.al. proposition of a Y-structure for the Taxonomy works better.

Table 2. Classical item p and d values

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Testlet no.</th>
<th>Item no.</th>
<th>Testlet no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>27</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>21</td>
<td>7</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>27</td>
<td>9</td>
</tr>
<tr>
<td>27</td>
<td>8</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td>28</td>
<td>8</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>29</td>
<td>8</td>
<td>30</td>
<td>10</td>
</tr>
</tbody>
</table>

N=436

Factor analysis based on tetrachoric correlations on the responses to the 30 items yielded a two-factor solution with a minor second factor for which 2 items (no. 20 and 29) loaded significantly. (Refer to Table 3.) However, these items also loaded more heavily on the first factor. The variance explained by the first factor was 21.54 while that explained by the second factor was only 0.83, leading to a conclusion that the response matrix is, by and large, unidimensional.

Table 3. Factor loadings based on tetrachoric correlation matrix of the 30 items

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Testlet no.</th>
<th>Item no.</th>
<th>Testlet no.</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>16</td>
<td>6</td>
<td>-0.94</td>
<td>-0.01</td>
<td>-0.89</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>17</td>
<td>6</td>
<td>-0.90</td>
<td>-0.13</td>
<td>-0.86</td>
<td>-0.01</td>
</tr>
</tbody>
</table>
item 3  -0.89  -0.05  item 18  -0.97  -0.03  
item 4  -0.78   0.04  item 19  -0.84  0.00  
item 5  -0.72  -0.23  item 20  -0.68  -0.47*  
item 6  -0.91   0.04  item 21  -0.97  -0.15  
item 7  -0.86   0.02  item 22  -0.92   0.11  
item 8  -0.70  -0.10  item 23  -0.71   0.08  
item 9  -0.91   0.10  item 24  -0.91   0.11  
item 10 -0.77  -0.06  item 25  -0.86   0.27  
item 11 -0.72  -0.27  item 26  -0.68   0.08  
item 12 -0.89  -0.01  item 27  -0.95   0.05  
item 13 -0.88   0.04  item 28  -0.80   0.15  
item 14 -0.77  -0.18  item 29  -0.67   0.46*  
item 15 -0.97   0.11  item 30  -0.96  -0.10  

Variance explained  Factor 1 : 21.54  Factor 2 : 0.83
* Significant loadings on both factors.

Upon conversion of the item response matrix into the testlet response matrix, the data was submitted to another factor analysis, this time, based on polychoric correlations, and a one-factor solution emerged, confirming the unidimensionality of the testlet response matrix.

Figure 1.  Item Estimates (Thresholds)
all on all (N = 436 L = 10 Probability Level=0.50)

A Rasch analysis was performed on the dataset. The test was found to be fairly well targeted (Average Ability = 0.57 logits; Average Difficulty = 0.00 logits) and reliability of the test was estimated at 0.90. As can be seen from Figure 1, notwithstanding a certain degree of overlap among the three Levels in the test, most items testing K/C (Level 1) were found at the easy end of the continuum and most items testing AN (Level 3) at the difficult end, with those testing AP (Level 2) found in between these two extremes.

The dataset was then submitted to a Rating Scale Analysis. The item fit statistics for each testlet range from 0.87 for testlet 6 to 1.12 for testlet 2, falling well within the acceptable limits. Infit and outfit statistics (mean squares) were close to 1 while their mean t-values were close to zero (refer to Table 4), implying an excellent fit of the data to the model.

Table 4.  Rating Scales Analysis of the testlet dataset

Table 4 also shows the average difficulty estimates for each testlet and the tau-values for the items within each testlet. (The k is the location of the kth category in each testlet relative to the testlet's scale value. They are also known as thresholds and are themselves,
ordered because they separate the ordered categories.)
The least difficult testlet is testlet 1 because it has the lowest scale value (-0.88). It tests the most basic concept taught at the earliest point in the Theory of Demand and Supply, which is the demand curve or schedule for a commodity. The item cluster or testlet with the highest scale value (0.55) is testlet 4 and it tests the concept of price-elasticity of demand, one which is relatively more abstract and quantitative in nature. Price-elasticity of demand, taught midway through the topic, often remains the most difficult to grasp for many students who are less mathematically inclined. Tau differences for each item cluster vary considerably between items. The difficulty of crossing the Level 1 (K/C) threshold depends on the complexity of the first items in each item cluster or testlet. The first items in testlets 4 and 7 are much easier compared to the 2nd items in the same testlets. This implies that crossing the Level 1 threshold to Level 2 requires a higher level of competence for examinees attempting these testlets. Similarly, since testlets 2 and 8 have large differences between the 2nd and 3rd items, crossing the Level 2 threshold to Level 3 requires a higher level of competence for testees attempting these testlets. In contrast, the tau-differences between the Level 2 and Level 3 items for testlet 6 is very small, implying a relative ease of crossing from Level 2 to Level 3 in this testlet.

With reference to figure 1, testlet 4 has even its 2nd item having a scale value which is higher than the third item in testlet 1. This brings to mind what was said earlier on page 2 in this paper about the lack of correlation between complexity and difficulty of task levels in Bloom's Taxonomy.

Insert Figure 2 about here

Figure 2 shows that the taus (or thresholds) corresponding to the 3 item categories (or levels) in each testlet, are clearly discernible. Level 2 appears to serve as a bridge between Level 1 (K/C) and Level 3 (AN).

The Rating Scale analysis of the cognitive-level-attainment data is summarised in Figure 2. The variable is laid out on the horizontal axis running from lower-order cognitive level attainment on the left to higher-order on the right. It is marked off in logits. Figure 2 shows the difficulties of the three levels from (K/C) to (AN) for each testlet. It can also be used to read off any student's most probable response to any of the 10 testlets, and so, provides a detailed interpretation for every score on this test. The following are interpretations of three students whose scale value lie in different positions on the continuum.
Jill (raw score=8; ability=-1.16, error=0.44) Jill’s performance band lies to the left of Level 2 (AP) and Level 3 (AN) questions and therefore, there is a very low probability of her responding correctly to any of them. In fact, there is only about a 50% likelihood of her answering Level 1 items correctly. She is a likely candidate for remedial lessons aimed at strengthening her skills at Level 1 and then perhaps later if she is successful at shifting her performance band further to the right, she could be coached in Level 2 questions. For the time being, AP and AN questions are way beyond her.

Jane (raw score=15; ability=-0.01, error=0.39) Jane is expected to respond incorrectly to the third item in each testlet i.e. items 3, 6, 9, 12, 15, 21, 24, 27 and 30, all of which test Level 3 (AN) but because she is to the right of the 1st step in all the testlets, we expect her to respond correctly to these items which test Level 1. There is about a 50% chance of her responding correctly to Level 2 items. That being so, we can infer that she is likely to have attained Level 1 (K/C) in Bloom's Taxonomy and is building up her skills in Level 2 (AP). Consequently, remedial lessons can be given to her to help her achieve Level 2. Level 3 is quite beyond her at this stage of her learning.

John (raw score=22; ability=1.15, error=0.44) John is very likely to answer correctly all the Level 1 and Level 2 items in each testlet but his chances of success on level 3 items are not as good. He may only get about half of the Level 3 items correct. This implies that he has already attained the two lower-order cognitive levels in Bloom's Taxonomy and is currently developing his competencies at the AN level. He should be given more practice in answering Level 3 questions.

Conclusion

This paper illustrates how an IRT measurement model can aid the teacher in locating students' attainment of Bloom's Cognitive Levels in a topic in 'A' Level Economics. Such information is useful in the planning of appropriate teaching and learning strategies so as to facilitate faster and greater attainment of educational objectives, with the existing level of competence as a starting point. Apart from the diagnostic and remedial implications the results of the study can be employed in the construction of criterion-referenced tests. Lam and Foong (1996) demonstrated this possibility in their study which estimated SOLO proficiency levels. Item scores for each level could be plotted against students' proficiency levels. With the aid of such a plot, then the students can be said to have attained, for example, Level 1 competence if they have correctly answered a pre-determined number of Level 1 items, and so on.
The main purpose of this study was not the verification (or otherwise) of the linearity of Bloom's Taxonomic categories. Nevertheless, a further consideration of Rating Scale Analysis is the possibility of detailed output which can provide teachers with a means of identifying students with aberrant response patterns (e.g. answering correctly for more difficult questions but responding incorrectly to easy questions) for further diagnosis (Cheung et. al., 1990). Whether such departures from the expected response patterns serve to debunk the assumption of cumulative hierarchy in Bloom's Taxonomy and lend credence to alternative propositions for the Taxonomy's structure, notably, that of Madaus et.al.(1973), requires further investigation, using a larger sample size and also, perhaps a larger number of testlets.

References

1 Gilbert (1992) p.45
1