ESTIMATING THE DIFFICULTY OF A'LEVEL EXAMINATION SUBJECTS IN UGANDA

Connie V. Nshemereirwe

University of Twente, The Netherlands

c.v.nshemereirwe@utwente.nl

In order to gain access to institutions of higher learning in Uganda, including universities, all students sit a national examination at the end of A'Level, the scores of which determine their selection into various institutions of higher learning, including university. For most university degree programmes, entry is determined based on the A'Level scores irrespective of subject, essentially implying that the same scores in the different subjects are comparable. In order to investigate this comparability, a generalised partial credit item response model was fit to the A'Level examination results data for the years 2009 and 2010. Science and non-science subjects were hypothesised to load on two separate dimensions of the latent ability scale, and subject difficulty and discrimination parameters were estimated. It was found that science subjects were relatively more difficult than humanities and language subjects, and that they also provided the largest amount of information, although this was for the higher end of the ability scale. Some other subjects like Art and Kiswahili were not only relatively easier, they also provided very little information on the ability scale underlying the other subjects. These findings bring into question the comparability of scores in the different subjects at A'Level, and if student ability based on examination performance can be better represented by integrating information on difficulty levels.

Key words: Subject Difficulty; A'Level Examinations, Subject Comparability; University Selection, Generalised Partial Credit Model; Multidimensional Item Response Theory.

INTRODUCTION

The Ugandan pre-university education system follows a 7-4-2 system: seven years of primary school, four years of lower secondary or "Ordinary Level (O'Level)" and two years of upper secondary or "Advanced Level (A'Level)". To advance from one level to the next, students must sit and pass a national examination which is centrally developed and administered by the Uganda National Examination Board (UNEB). At the end of primary school, pupils are examined in four subjects, all of which are compulsory: English, Mathematics, Social Studies and Basic Science and Health Education. At the end of O'Level, students sit ten to twelve subjects that they may choose from thirty-six in total; of these, five are compulsory: English Language, Mathematics, Physics, Biology and Chemistry. At A'Level, students may choose three subjects out of a possible twenty-six, and an additional compulsory subject called General Paper. These twenty-six subjects are categorised as shown in Table 1, and the UNEB gives the following guidelines on subject choice:

> Candidates are advised to avoid selecting more than one subject from groups that are normally timetabled together. […] candidates are particularly advised to avoid combining Science subjects with Arts subjects, e.g. Sciences with Languages, Physics with Geography, etc. (UNEB, 2010)

Table 1: A'Level Subject Categories

I. GENERAL PAPER (COMPULSORY) **II. HUMANITIES** P210 History P220 Economics P230 Entrepreneurship Education P235 Islamic Religious Education P245 Christian Religious Education P250 Geography **III. LANGUAGES** P310 Literature in English P320 Kiswahili P330 French P340 German P350 Latin P360 Luganda P370 Arabic **IV. MATHEMATICAL SUBJECTS** P425 [Pure] Mathematics S475 [Subsidiary] Mathematics	**V. SCIENCE SUBJECTS** P510 Physics P515 Agriculture: Principles and Practice P525 Chemistry P530 Biology **VI. CULTURAL SUBJECTS AND OTHERS** P615 Art P620 Music P630 Clothing and Textiles P640 Foods and Nutrition **VII. TECHNICAL SUBJECTS** P710 Geometrical and Mechanical Drawing P720 Geometrical and Building Drawing P730 Woodwork P740 Engineering Metalwork

(Source: UNEB, 2013)

UNIVERSITY SELECTION

The minimum requirement for university entry in Uganda is two principle passes, or a score of between *A* and *E* in at least two A'Level subjects taken at principle level. Additional entry requirements differ between academic programmes as well as universities. Some academic programmes have more restrictive entry requirements, like engineering and medicine, but many other academic programmes have open requirements. The majority of university students in Uganda are enrolled at public universities, where a system of weighting is used at selection. A'Level subjects categorised as essential for a given university academic programme receive a weight of three, the relevant subjects receive a weight of two, and any other subject not categorised as essential or relevant receives a weight of one or a half. However, in programmes which have more open subject requirements, these weights are simply applied to the subjects in which students score the highest grades. Table 1.2 shows the entry requirements for three very popular academic programmes at public universities: Bachelor of Information Technology, Bachelor of Business Administration, Bachelor of Development Studies and Bachelor of Laws. It can be observed that subject requirements become more and more open until students who apply to enter the bachelor of development studies or the bachelor of laws can be admitted with any A'Level subject combination. The question this raises, however, is whether all subject scores are interchangeable, and can be thought to represent similar academic ability.

Table 2: Entry Requirements for Four Academic Programmes at Public Universities

Programme	"Essential" (receives a weight of 3)	"Relevant" (receives a weight of 2)
B. Information Technology (BIT)	Two best done of Maths, Economics Physics, Biology, Chemistry, Literature, Geography, Entrepreneurship, Technical Drawing, Fine Arts	One better done of the remaining A'Level subjects
B. Business Administration (BBA)	Economics and one better done of the remaining A'Level Subjects	Next better done of the remaining A'Level Subjects
B. Development Studies (BDS)	Two best done of all A 'Level Subjects	Third best done of all A' Level Subjects
Bachelor of Laws	Two best done of all A 'Level Subjects	Third best done of all A' Level Subjects

(Source: Joint Admissions Board, 2012/2013)

THE CONCEPT OF "SUBJECT DIFFICULTY"

Subject difficulty as a concept is rather controversial. On one hand, the observation that certain subjects generally have higher pass rates than other subjects appears to indicate that some subjects are relatively more difficult than others; on the other hand, it can be argued that pass rates may be a result of other factors intrinsic to the education system such as less qualified teachers in some of the subjects, or intrinsic to students themselves, such as varying levels of motivation (i.e. more motivated students tend to choose certain subjects), rather than a characteristic of the subject itself. Additionally, there is a possibility that grading practices in some subjects are simply more stringent than in others. Finally, it can also be argued that scores in different subjects may indicate different dimensions of ability in the first place, rather than a uniform dimension that underlies all subjects, and that therefore no sensible comparison can be made between them.

Aside from comparison of subjects to one another at the same sitting, another issue of contention is comparability of examination scores across time. Public confidence in the school system is often shaped by whether performance is improving or not, judging from pass rates. Unfortunately, this sets up a situation where an increase in the proportion of students passing raises concerns that examination standards are falling (examinations are easier or have been compromised), and when pass rates drop, this raises concerns that standards in schools are falling. William (1996, in Coe, 2010) has described the dilemma that school systems and examination boards face in this case as a "heads I win, tails you lose" situation.

Current Views of Subject Comparability

In considering subject comparability, it may be useful to start with reviewing the process of grade allocation itself. In examination systems such as Uganda's, an A'Level grade scale, such as A-F, is applied across all subjects, and the grade boundaries agreed upon by a panel of subject matter experts. Care is taken to decide on these grade boundaries in such a way as to maintain some kind of comparability between the letter grades from year to year. According to Newton (2005), these kinds of panels may also make use of statistical information on candidate performance in previous years, as well as technical information regarding mark distributions for the particular sitting, so as to arrive "comparable" grade boundaries. This process of judgemental grade boundary allocation or "linking" is meant to enable fair decision-making, such as university selection, for students sitting the same subjects from year to year.

The purpose of national examinations, however, is not only for selection for the next level, but also to provide data to enable the monitoring of schools and education systems. In this case, it is also necessary to be able to determine the actual achievement levels of students from year to year; that is to say, the knowledge and skill levels in each subject so as to judge progress. In Uganda, the UNEB uses a combination of criterion and norm referencing to arrive at grade boundaries, and these two methods of viewing performance reflect the two main views on "comparability" as well, namely *performance* comparability and *statistical c*omparability.

Performance comparability of any two subjects concerns judging difficulty based on the degree of challenge each subject presents students. This challenge may be in terms of complexity, skill level or knowledge required to score the same grade in each subject. The main difficulty with this conceptualisation of difficulty is the fact that complexity and skill levels cannot be directly observed and therefore must be inferred, making this comparability method problematic (Coe, 2010). Further, different knowledge and skill sets may be necessary for the different subjects, and then how can a judgement be made on which is the more "difficult"?

Statistical comparability circumvents this problem by only relying on defining a standard as the relative chances of success that candidates have in different subjects. Coe (2010) puts it as follows: "Two subjects are of comparable standard if the same grades are equally likely to be achieved by comparable candidates in each" (p275). A statistical conceptualisation of comparability, however, takes no account of the quality or content of the examinations, which, depending on the use to which the comparability is to be put, may be problematic as well. Nevertheless,

For purposes of the current study, a statistical comparability view is appropriate because the focus is on the use of a simple average of A'Level subjects scores for selection for university. That is to say, scores in the A'Level examinations are used as a basis to qualify students by ranking them, rather than as an indication of specific skill and knowledge levels. In that case, it is more useful to apply statistical comparability, and a more detailed description of the methods involved in this is presented in the next section.

Statistical Comparability of Subject Scores

Coe *etal* (2007) gives a summary of the statistical methods employed in the comparison of subject scores. These include:

Common Examinee Linear Methods – the best known of these methods is Kelly's method (1976), which estimates the difficulty of a subject based on all candidates who have taken that particular subject along 5

with any other. Kelly's method involves the solution of simultaneous equations, which allows the average performance of each subject to be used in the computation of the subject difficulties of all the other subjects in an iterative process that repeatedly corrects for ―difficulty of each subject until the differences between corrected subject scores is zero.

Latent Trait Methods – these are methods that rely on Item Response Theory (IRT), which takes the view that not all items in a test give the same amount of information about the ability of a student. Some items are more difficult, and even though two students get the same number of items correct, a student's ability depends on *which* questions s/he got correct. The idea is that the probability of a person answering a given item correct is a mathematical function of the difference between the ability of a person and the difficulty of that question. Given the responses of a number persons on a set of items, therefore, the "difficulty" of items can be simultaneously estimated along with person "ability" using an iterative maximum likelihood procedure which assigns an ability to a person that best matches their response pattern given the difficulty of the items. The difficulty of items and the ability of persons can then be represented on the same "latent trait" scale, with persons higher up on the latent scale having a higher probability of answering more difficult items correct. In estimating subject difficulty, latent trait models take the individual subjects to be items, and the subject scores to represent the response pattern by each student on these items (subjects).

Latent trait models have an advantage over Common Examinee Linear Methods like Kelly's in that they allow for the interval between subject scores in terms of difficulty to vary. In other words, the difference between a score of A and B need not be equal to the difference between a score of B and C; similarly, the distances between scores in different subjects need not be the same, so that the distance between a score of A and B in History can differ from the distance between a score of A and B in Chemistry. Another advantage of Latent trait models is that depending on the particular model employed, it is possible to determine the extent to which subject can be represented on by single underlying dimension or more than one dimension, and to test which explanation best fits the observed data. In this way, it is possible to examine the extent to which subjects can indeed be compared to one another.

Statistical Comparability – some criticisms

Coe (2008) outlines some criticisms of statistical approaches, such as the basic incomparability of subjects in general, and the fact that performance is affected by many other factors besides "difficulty". Further, the analysis of subject "difficulties" for different subgroups, such as males and females, may result in different difficulties, and that even the method of statistical analysis itself matters as different methods tend to give different results. Coe (2008) maintains, however, that statistical differences are still interpretable within the context of a linking construct as long as all inferences are confined to that linking construct. The important consideration, then, is the identification of a plausible linking construct.

Construct Comparability: An Integrated View of Subject Comparability

Given the shortcomings of both performance and statistical views of comparability, Newton (2005) proposes a third, integrated view, which he terms as *construct* comparability. This view of comparability takes the position that it is inadvisable to infer any sense of equivalence based on a statistical comparison of scores on a combination of subjects; rather, "comparison" can only translate the scores in these different subjects to another scale which expresses the extent to which the scores measure the same *construct.* Inferences about the scores so-linked can therefore only be made with reference to this construct. It should be noted that this construct is not identical to any of the constructs being measured by individual tests, and that no such inference should be made (Newton, 2005). Coe (2008) goes further to

say that in comparing subject scores, it can only be said that a given score in a subject indicates a lower level of the linking construct than the same score in another subject. Take for instance comparing scores in Mathematics and English: while these two subjects clearly represent different abilities, it is still reasonable to say that a high score on both may be indicative of a more general academic ability. In placing the scores in these two subjects on a scale of academic ability (the linking construct), it can then be said that a high score in one subject represents a higher level on the linking construct than the same score in the other subject. That being said, careful thought and consideration must go into defining this linking construct, and then "*made explicit for all users and stakeholders"* (Newton, 2005, pp 111, emphasis in original) so as to avoid invalid inferences.

Subject Comparability of A'Level Examinations in Uganda: A Linking Construct

Depending on the purpose to which the scores in national examinations are put, therefore, a linking construct can be proposed. For instance, A'Level examination results in Uganda are the basis for university entry; as such, a linking construct such as university "potential" can be proposed so that the scores in different subjects can be compared based on such a scale. This is especially applicable for those university degree programmes that do not impose any limitations on the A'Level subjects required for admission, but even for those that do, such as Engineering and Medicine, a construct such as "scientific ability" can also be used to place scores in subjects such as Mathematics, Physics, Chemistry and Biology on the same scale. In other words, subjects that are strikingly different can be scaled separately and then the aggregation made thereafter so that students who choose "difficult" subjects are not disadvantaged.

ESTIMATING A'LEVEL SUBJECT DIFFICULTY IN UGANDA: A METHODOLOGY

Item Response Theory

Item Response Theory (IRT) is a general statistical theory which attempts to relate the performance of an individual on an item to the ability measured by that item (Hambleton & Jones, 1993). In contrast to traditional testing where a person's ability is inferred from a total score, IRT uses the information on the individual's responses to every item. IRT rests on three assumptions: a) items measure a uniform underlying trait (unidimensionality); b) a response on one item is not dependent on the response to another item on the same test (local independence); and c) That the relationship between a person's response and their ability can be mathematically modelled by a logistic function (Hambleton & Jones, 1993).

In general, IRT modelling proceeds by analysing the responses of a large number of individuals to a given number of items with the aim of estimating the ability level associated with a given response pattern. In this process, two parameters are commonly estimated: item *difficulty, b*, and item *discrimination, a*. Item difficulty, *b*, represents the ability level at which there is a 50-50 chance of scoring in a given category, and in this way can *locate* the item difficulty on the same scale as person ability, Θ (theta). Once items have been calibrated, a particular response will indicate the same Θ value no matter who attempts the question, which is a distinct advantage of IRT because item parameters are not tied to a particular population - this property of IRT is known as *invariance*. The Θ scale itself runs from negative infinity to positive infinity, and is often scaled by fixing the zero point at the population mean, with each unit change in the value of theta being equal to a change in ability represented by one standard deviation in the population.

Secondly, it is usually also possible to model how well a given item discriminates between individuals with a different latent trait ability. An item has high discrimination if it can detect a small difference in the level of ability between persons based on their response; in other words, if the probability of a given

response was plotted against ability levels, a highly discriminating item would have a steeper slope since the difference in probability of that response at low levels of ability would be quite different from that of individuals with a higher level of the latent trait. A flatter slope would signify that the probability of a given response does not change much between persons of low and high ability (Baker, 2001). The idea of discrimination is parallel to that of factor loadings in factor analysis; an item which has a high discrimination can be thought of as loading heavily on the underlying latent trait, and can measure the ability levels of different individuals more precisely. It should be noted that an item may have high discrimination only in a small part of the ability dimension; for instance, an item may be very well suited to differentiate individuals at the upper end of the ability scale but have little discriminatory power at the lower end since most of the individual would score in the lowest category on that item. This gives IRT an advantage in testing because it is sometimes desirable to discriminate between individuals of a similar level, an advantage that is put to full use in computer adaptive testing. Once item difficulties and discriminations have been computed, an individual's response pattern places him/her on an ability scale, which is on the same scale as item difficulties.

Within the IRT framework, various models have been developed to deal with different test formats and to meet different assumptions. Students in Uganda may obtain a grade of *A, B, C, D, E, O* or *F* in the national A'Level examinations, with A being the highest grade and F being the lowest. Each students takes examinations in three or four subjects; in order to model student performance using IRT, each subject can be thought of as an item with seven score categories to represent the seven possible grades. Since there are more than two possible scores categories for each subject (or item), then modelling the relationship between student responses and subject difficulty requires a model developed for polytomous items.

IRT Models for Polytomous Items

These models are divided into two major categories – those for items where the response categories are ordered (ordinal), and those where the response categories are in no particular order (nominal). In the present case, if the ordering is certain, i.e. $A>B>C>D>E>O>F$, then the one of the ordinal models would be suitable; however, if this ordering cannot be assumed in advance and one wants to test the hypothesis that $A>B>C>D>E>O>F$, then a nominal response model is more appropriate. In the present case, the ordering was not assumed in advance; further, it was of interest to not only estimate the difficulty of items but also their discrimination, and the most suitable model for this was found to be the Generalised Partial Credit Model (Muraki, 1992). The GPCM is also particularly suitable for the modelling of A'Level subject difficulty because it allows items to have different numbers of score categories; At A'Level, there are some subjects where no one scores *A*, or where no one scores *F,* so that subjects end up with a different number of score categories, so that items end up having a different number of score categories.

Generalised Partial Credit Model (GPCM)

Difficulty in the GPCM is conceptualised as the *threshold* where the probability of scoring in the adjacent category is more likely; as such, threshold values are estimated for all adjacent categories so that more than one difficulty, or threshold, parameter is estimated for every item. It can be imagined that as ability increases, the probability of scoring in a lower category decreases as the probability of scoring in the adjacent category increases. Put another way, the probability of scoring in the lowest category, for instance, is always dropping with increasing ability since the probability of scoring in any other category is also rising at the same time and the total of probabilities always equals one. At some point, the probability of scoring in an adjacent category becomes higher than that of scoring in the lowest category, and the point at which these two curves cross marks the threshold ability or difficulty where the chances

of scoring in either category are equal. Figure 1 represents the category response curves for an item with five response categories k = 1 to k = 5. For this particular item, the ability level that is needed to "cross" the threshold between category one and category two, or the point at which the probability of scoring in the adjacent category becomes higher than scoring in the lowest category, is around Θ = -1.5; the next threshold occurs at approximately Θ =-1.2, and then the next one, where responding in category k=3 becomes less likely than scoring in category k=4 occurs at ability level around Θ = 1.4., with the last threshold being located at closer to Θ = 1.8.

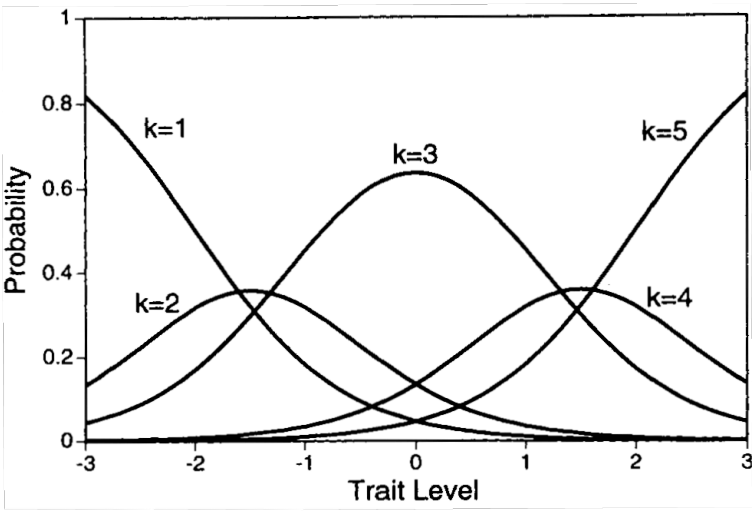

Figure 1: Probability of scoring in different categories for a polytomously scored item. Adapted from "Confirmatory Factor Analysis and Item Response Theory: Two approaches for exploring measurement invariance" by S. P. Reise, K. F. Widaman, & R. H. Pigh, 1993. *Psychological Bulletin 114* (3), 552-566. Copyright 1993 by the American Psychological Association. Adapted with permission.

In other words, the threshold parameter is that value of θ where scoring in the adjacent category is more likely.

Item Information.

In IRT, the item difficulty parameter locates it on the Θ scale, and the discrimination parameter describes its loading or steepness on the underlying latent scale; however, in order to interpret these parameters in any meaningful way, it is necessary to inspect the information functions for each of the items (Muraki, 1993). Item information is essentially an expression of how precisely a given item estimates the ability parameters of individuals responding to it; this precision is indicated by the variance of those estimates, and item information is equal to the reciprocal of that variance. If responses on an item lead to quite a precise estimate of the ability parameters, then the variance of those estimates will be low and information will be high; if, on the other hand, the estimates have a high variance, such an item provides little information on the latent trait (Baker, 2001).

For polytomous items, item information functions may be unimodal or multimodal, depending on the distance between the threshold difficulty parameters of adjacent categories; if it is large, then the information will drop in the Θ range between them (Muraki, 1993). Figure 2a and 2b show the item category response functions of two different items together with their item information functions. The

first item has the following item parameters: a = 1.304, b = (-1.289, 0.292, 0.381, 1.252, 2.026, 2.735), and the second has the following item parameters: a = 0.647, b = (-3.751, -1.794, -1.600, 0.476, 1.894, 3.301). Both items can be scored from 0-6, and total item information is shown by the thick dashed line. The information curves show that item 1 provides more information about the underlying trait than item 2. Further, the peak of the item information indicates where along the latent trait this item provides the most information: for item 1 that is around Θ = 0.7 and for item 2 is around Θ = -1.5. This is consistent with the *a* and *b*-values of the two items since item 1 has a higher discrimination and average difficulty than item 2, which leads to the expectation that item 1 will provide more information, and at a higher value of Θ than item 2.

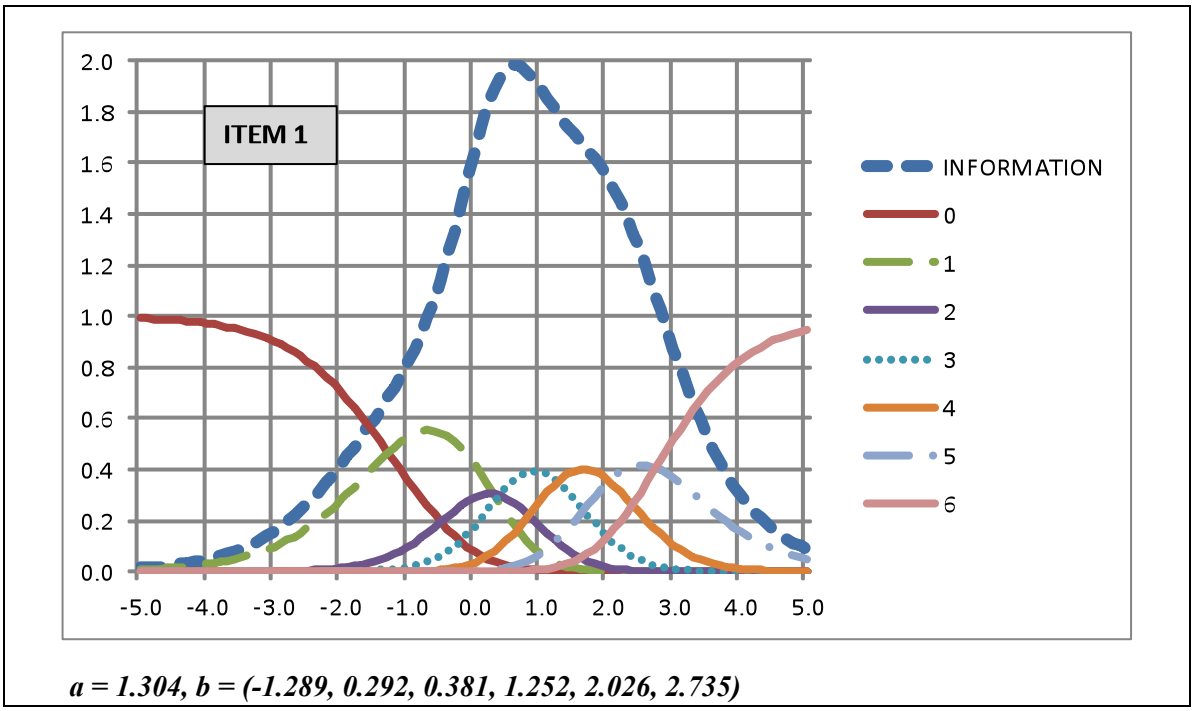

Figure 2(a): Item category response functions for a polytomous item with high information

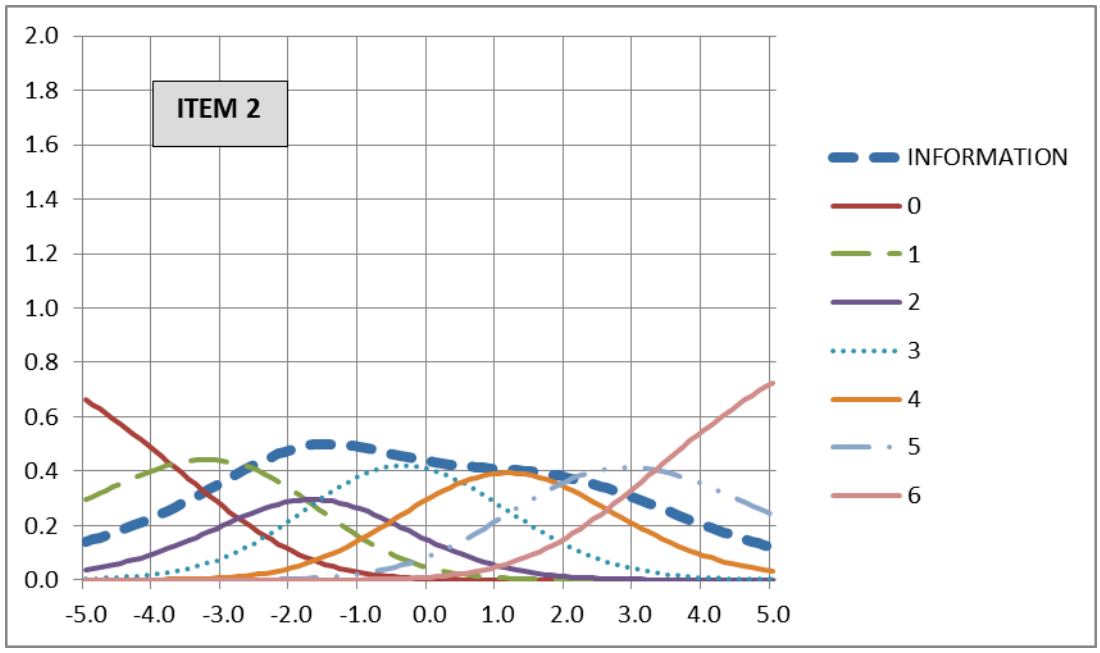

Figure 2(b): Item category response functions for a polytomous item with low information.

Multidimensional Item Response theory (MIRT).

One of the basic assumptions of IRT is that the test items whose parameters are being estimated all measure a single underlying trait; however, it is easy to imagine a situation where the probability of success depends on more than one ability, such as word math problems where success depends on the respondent's ability to comprehend the language, as well as know the applicable mathematical principles to solve the problem. In modelling such items, assuming unidimensionality of the underlying trait would lead to inaccurate parameter estimates, in which case multidimensional IRT (MIRT) is a more suitable analytical framework (Ackerman, Gierl, & Walker 2003). In the present study, multidimensionality was suspected based on the large differences in performance in science subjects compared to the humanities at A'Level in Uganda. As such, a two dimensional latent ability was explored.

ESTIMATING SUBJECT DIFFICULTY IN THE UGANDA NATIONAL A'LEVEL EXAMINATIONS

Findings

Using the MIRT computer program (Glas, 2010), the GPCM was fitted to the data as a unidimensional model and a 2-dimensional model (Sciences/All the rest). The science dimension was made up of physics, mathematics, chemistry, biology and agriculture. It turned out that the 2-dimensioanl model fit the data best, and that the two dimensions were shown to be fairly distinct, with a correlation of only 0.66.

Discrimination parameters

The data analysed were from the 2009 and 2010 A'Level examination sitting, and Figure 3 shows a plot of the *a*-parameters (discrimination parameters) for each of the 16 subjects analysed, for the 1and 2-dimensional models for the data from 2010. The a-parameters obtained from each analysis indicate how well scores on a given subject discriminate between students with a different ability on the given dimension; as such, subjects with high values of discrimination provide more information on the ability of students than subjects with low discriminations. In this case, most of the science subjects were on the high end of the scale, while some of the languages and fine art were on the lower end. This means that scores in these subjects are not able discriminate between students with regard to ability as well as the ones on the high end. The rest of the subjects lie somewhere in the middle range.

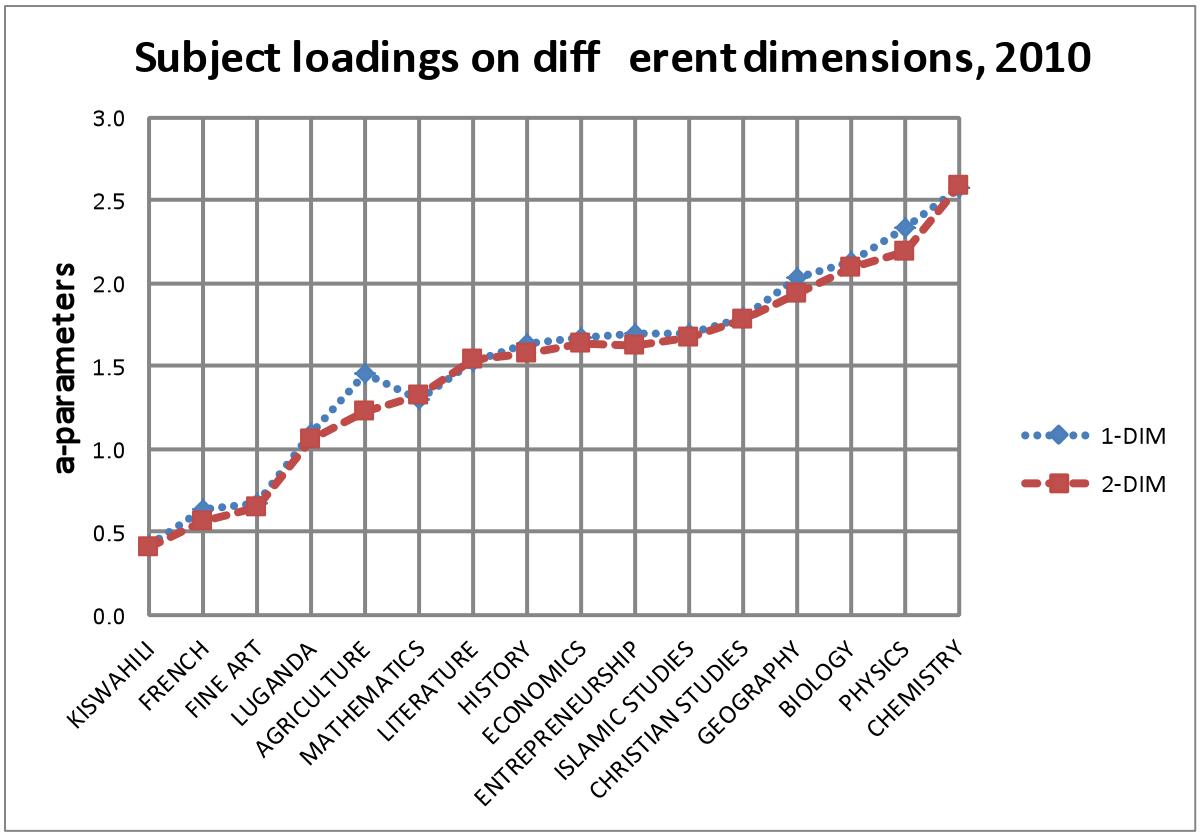

Figure 3: Subject Loadings (as indicated by subject discrimination or a-parameters) for the 1 and 2-dimensional models.

Difficulty Parameters

The subject difficulty was estimated by treating student scores on each subject as though it was the score on an item that could be awarded a mark between 0 and 6. The GPCM estimates *threshold difficulties, b,* which represent the values at which the probability of a student with a given Θ scoring in the adjacent category, say *D*, is higher than that of scoring in the present category, say *E*. However, since threshold difficulties differ so much between and within subjects, a comparison of subject difficulty based upon them is rather difficult; as such, an average of the threshold difficulties for each subject was computed, and in Figure 4 a plot of the relative subject difficulties for the two years is shown. The general trend

shows that the local languages Kiswahilli and Luganda have the lowest relative difficulty, while the four science subjects mathematics, physics, chemistry and biology have the highest relative difficulty.

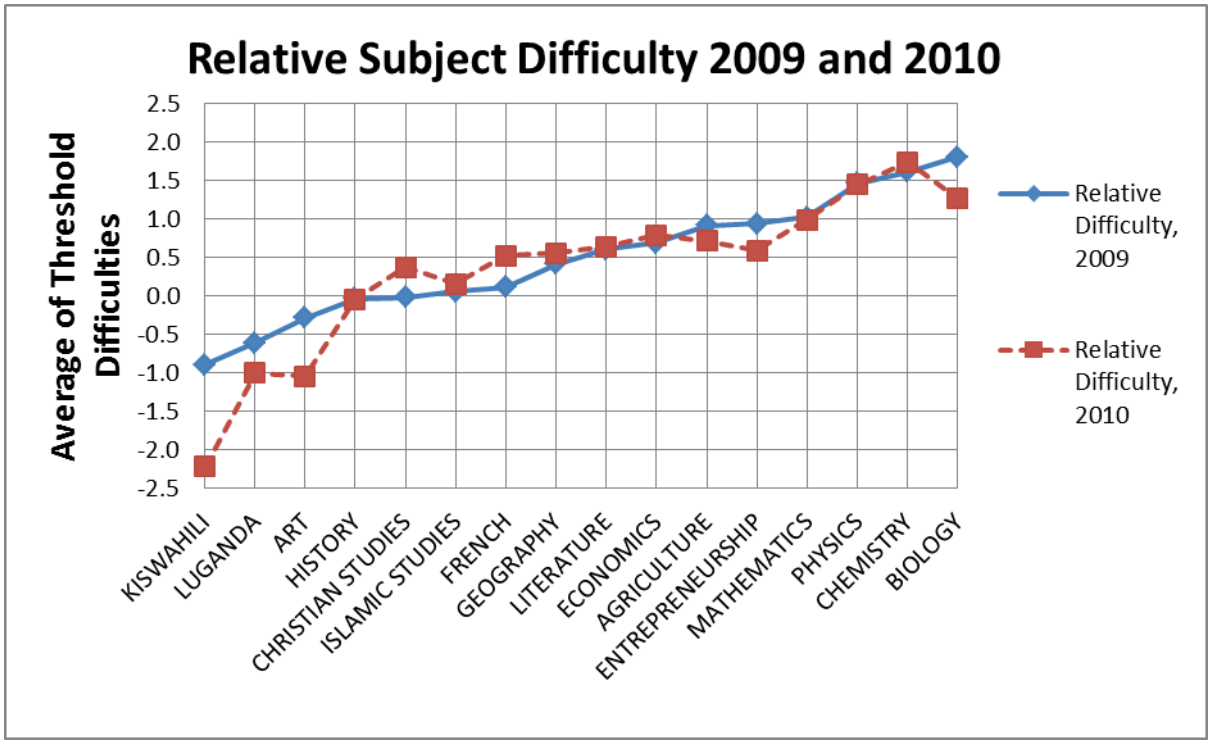

Figure 4:Relative subject difficulty, A'Level national examinations, 2009 and 2010.

Item information

In practical terms, estimates of item difficulty and discrimination parameters are most useful if we know the ability levels about which they give the most information. For a start, one may expect that high relative *a* and *b*-values mean that such items discriminate best at the higher end of the ability spectrum, while those with high *a* but low *b*-values discriminate best at the lower ability spectrum. However, for polytomous items such as the ones that were analysed in this study this is not straightforward because threshold categories behave different for different items. In this case, plots of the item information functions (IIF) are more informative, and these are shown for three different items in Figure 5. In order to show the relative amount of information provided by different items, all the IIFs are plotted to a value of information equal to 5.0, except for chemistry which goes to 7.0; information is indicated by the thick broken line.

The top panel shows a subject, Art, that gives almost no information about the underlying trait measured by the other subjects, and the bottom panels two show items which give more information about the underlying trait. History provides the most information at slightly below the average performance of students in 2009. Economics, mathematics and geography also gave a moderate amount of information. Finally, chemistry gives the most information at about 1.5 standard deviations above the average. Physics and biology gave similar amounts of information, but chemistry also tuned out to be bi-modal, with a smaller peak at just below $\Theta = 0$, meaning it also discriminates enough at that ability level to provide some information.

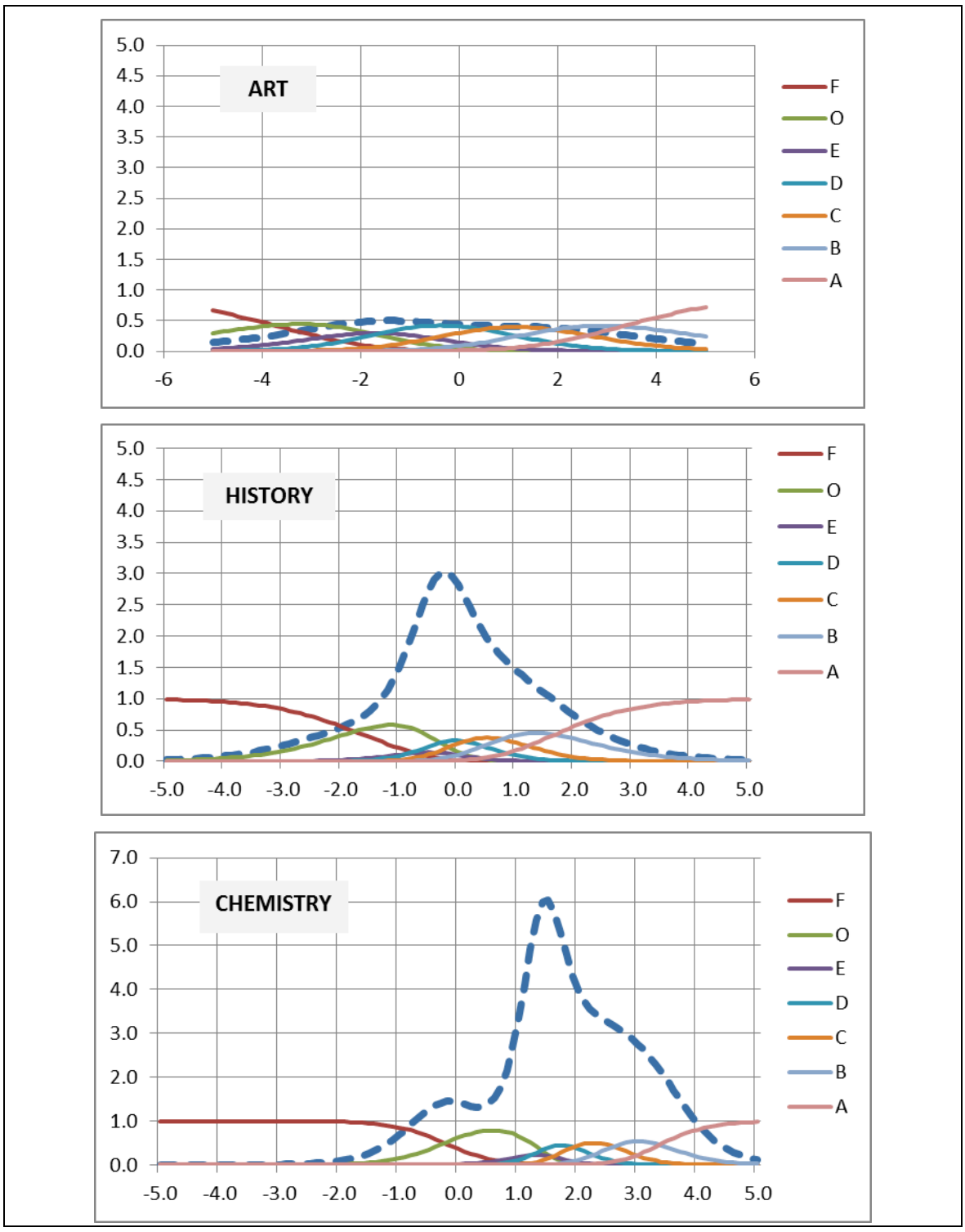

Figure 5: Item Information functions for three items in the A'Level national examinations, 2009

DISCUSSION AND CONCLUSION

Most degree programmes offered at universities in Uganda admit students almost sorely based on their scores in the national examinations at the end of the upper or advanced level of secondary school (A'Level). A weighting methodology is employed based on the subjects deemed most relevant for a given degree programme, and for the most part this is done for courses like engineering and medicine, but in many other cases the highest weight is applied to the best performed subject instead. It turns out that the majority of students take humanities and language subjects at A'Level, and these also have the highest pass rates. On the other hand, degree courses like development studies and business administration have very broad admission criteria, and so the majority of enrolled students took humanities and language subjects at A'Level. Assuming that universities want to enrol the students of highest ability, however, it may not be valid to assume that all subject scores are interchangeable, or that they represent a similar general ability.

The study described in this paper was aimed at estimating the subject difficulty of the sixteen most commonly chosen subjects at A'Level in Uganda so as to give a more accurate picture of the extent to which scores in these subjects can be compared. Using data from two A'Level examination years, 2009 and 2010, a modelling method based on item response theory (IRT) was utilised, and it was assumed that the sixteen subjects were different to such an extent that scores on them represented two separate ability dimensions: a science and a non-science dimension. The science dimension was represented by the subjects of biology, chemistry, physics, agriculture and mathematics, and the non-science dimension was represented by some humanities subjects like economics and geography, and some language subjects like Kiswahilli and Luganda. The generalised partial credit IRT model was fit to the data using the program MIRT (Glas, 2010). The two-dimensional model turned out to have best overall fit to the data, with the correlation between the two dimensions found to be about 0.65. This was low enough to support the likelihood that scores in science subjects represent a separate ability dimension

Modelling the subject difficulty in this study revealed that the science subjects, on the whole, have the highest relative difficulty (averaged over score categories), and that they also generally have the highest discrimination values. Aggregating the information provided by the difficulty and discrimination parameters, it was found that scores in subjects like Fine Art gave very little information about the 2-dimensional ability trait measured by the rest of the subjects, and that the little information they gave was at the lower end of the ability scale. Subjects like history, economics, mathematics and geography, on the other hand, gave a moderate amount of information around the middle ranges of the ability scale, while the sciences (biology, chemistry and physics) gave the highest amount of information, but more within the higher ability range of the scale. The exception was chemistry, whose information curve was bi-modal, such that it also gave a moderate amount of information within the middle ability range.

The findings of this study are in line with what is generally thought of the difficulty of science as compared to non-science subjects, but also provide a way to compare non-science subjects to one another. The subjects with some of the highest pass rates like Art and the local languages, appear to measure something different from what is measured by the other subjects, and yet they are assumed to be comparable to them. In the absence of a mechanism to compare subjects, universities understandably have to rely only on raw scores in different subjects; however, the findings of this study provide an alternative way of regarding the A'Level grades of applicants at selection, so as to improve the quality of students enrolled and make the selection process more fair.

REFERENCES

Ackerman, T. A., Gierl M. J., & Walker, C. M (2003). Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice* 22:3, 37–51

Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Coe, R. 2008 C*omparability of GCSE examinations in different subjects: an application of the Rasch model*, Oxford Review of Education, 34:5, 609-636, DOI: 10.1080/03054980801970312

Coe. R. 2010 *Understanding comparability of examination standards*. Research Papers in Education. 25: 3. 271 — 284

Coe, R., Searle, J., Barmby, P., Jones, K. & Higgins, S. 2007 *Relative difficulty of examinations in different subjects.* Report for SCORE (Science Community Partnership Supporting Education) (Durham, Curriculum, Evaluation and Management Centre, Durham University).

Glas, C. A. W. (2010). *Preliminary manual of the software program Multidimensional Item Response Theory (MIRT).* Retrieved from University of Twente website: http://www.utwente.nl/gw/omd/Medewerkers/temp_test/mirt-manual.pdf

Hambleton, R. K. & Jones R. W. 1993 *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development.* Educational Measurement: Issues and Practice 12(3), 38–47 DOI: 10.1111/j.1745-3992.1993.tb00543.x

Kelly, A. 1976 A study of the comparability of external examinations in different subjects. *Research in Education*, 16, 37-63.

Muraki, E. (1993) Information functions of the generalised partial credit model. *Applied Psychological Measurement 17*, 351-363.

Newton, P. E. 2005 *Examination standards and the limits of linking,* Assessment in Education*,* 12(2), 105–123. DOI: 10.1080/09695940500143795

UNEB. 2010: Registration of Candidates for 2010 UCE and UACE Examinations. Retrieved from : http://www.uneb.ac.ug/index.php?link=Guidelines&&Key=Secondary (4.10.2013)

UNEB. 2013 Uganda Advanced Certificate of Education (UACE) Examinable Subjects. Retrieved from http://www.uneb.ac.ug/index.php?link=Syllabus&&Key=A (4.10.2013)