This is a single article from *Research Matters: A Cambridge Assessment publication*. http://www.cambridgeassessment.org.uk/research-matters/
© UCLES 2011

Subject difficulty – the analogy with question difficulty

Tom Bramley Assistant Director, Research Division, Assessment Research & Development

Introduction

Concerns about differences in the difficulty of examination subjects are not new. Moreover, there have been considerable differences in opinion over i) how subject difficulty should be defined; ii) whether and how it should be measured (represented numerically); and iii) whether individual results in examination subjects should be adjusted to 'allow' for differences in difficulty as defined and measured in some particular way. See Newton (*in press*) for a review.

The purpose of this article is to explore in some depth one particular way of defining and measuring subject difficulty – a way that will be called the 'IRT approach'. This approach has been investigated in Australia in the context of university admissions by Tognolini and Andrich (1996) and in the Netherlands by Korobko, Glas, Bosker, and Luyten (2008), and has recently been advocated in the UK context by Robert Coe at the CEM centre in Durham (Coe 2008, Coe *et al*., 2008).

This article is structured as follows. First the IRT approach is briefly described. Then the analogy of using the IRT approach when the 'items' are examination subjects is explored. Next the task of defining difficulty from first principles is considered, starting from the simplest case of comparing two dichotomous items within a test. The thinking of Louis Guttman on scales and dimensionality is shown to provide a useful framework for understanding difficulty, and the link between Guttman and IRT is described. Finally, an alternative to the IRT approach, based on producing visual representations of differences in difficulty among just a few (three or four) examinations, is offered as an idea for future exploration.

Item Response Theory

Item Response Theory (IRT) is concerned with modelling the scores obtained on the *items*[1] *on a test*, rather than scores or grades obtained on a whole test (or on a composite of several tests). It (IRT) is not limited to educational tests – for example, it is quite widely applied in psychological testing more generally and in healthcare, but the educational context is the only one considered here. An overview of IRT can be found in Yen and Fitzpatrick (2006).

The organising concept of IRT is that of the 'latent trait' or continuum – an abstract line representing whatever the test concerned is supposed to be measuring. The most commonly used unidimensional IRT models contain a single parameter that represents person location on the trait (usually referred to as their 'ability') and one or more parameters characterising the item. In the simplest IRT model, the 1-parameter IRT model for dichotomous items, each item is characterised by a single parameter representing its location on the trait (usually referred to as its 'difficulty'). The 1-parameter model expresses the probability of a person with a given ability succeeding on (i.e. answering correctly) an item with a given difficulty as a function of the difference between ability and difficulty.

The 2- and 3- parameter IRT models for dichotomous items include extra parameters that represent 'discrimination' and 'guessing' respectively. The latter is often used for multiple-choice tests. IRT models for polytomous (i.e. multiple-mark) items also exist. These contain parameters representing the thresholds between adjacent score categories on the trait. In a multidimensional IRT model a person's ability is represented as an *n*-element vector rather than by a single number.

There are many reasons why IRT models are used, but the one of most relevance to this article is that (when the data fit the model) estimates of person ability and item difficulty on a common scale can be made when people have answered different subsets of items. This is the principle behind item banking and computerised adaptive testing (CAT), two of the main practical applications of IRT.

It is this feature of IRT that suggests it might have something to offer to the problem of comparing examination subject difficulty, because in most examination systems (and in particular for GCSEs and A levels) the examinees do not all take examinations in the same set of subjects. In applying the 'IRT approach' the different examination subjects have the role of different items in a test. A pass-fail examination could therefore be modelled analogously to a dichotomous item, and a graded test modelled analogously to a polytomous item.

The analogy with item-based IRT

The first issue that potentially weakens this analogy is the lack of clarity about the meaning of the trait when examination subjects are modelled with the IRT approach. When individual items are being modelled, as in 'normal' IRT, it might be something like 'maths ability' (for a maths test). The items in such a test will have been designed according to a specification setting out the criteria (e.g. topics and skills tested) that items must meet in order to be included in the test. In an IRT item banking/CAT scenario the items will also have been screened to check that they do in fact fit the model to a satisfactory degree. An important part of test validation (e.g. Kane, 2006) is to provide evidence of 'construct validity' – in other words that the items do conform to the definition of the trait and that their scores enter into the empirical relationships predicted by the theory of the trait.

However, there is no such deliberate design and validation in the case of examination subjects. The set of possible examinations on offer depends heavily on the cultural and political context of the examination system. In the case of A levels there are currently around 80 possibilities including subjects as diverse as Physical Education, English Literature, Accounting, Chemistry, Latin, Travel and Tourism, Music, and Critical Thinking. If these subjects can be located along a single unidimensional

1 In this report, the terms 'item' and 'question' are used interchangeably.

trait it might be called 'General Academic Ability' (Coe, 2008). While it is a bit optimistic to expect every single subject to be adequately represented on a single line, explorations of the data might reveal a subset of subjects that can more reasonably be thus represented. For example Coe (2008) found that by starting from a group of 37 large-entry GCSE subjects, removing ten that did not fit the model well, and then selectively adding in smaller-entry subjects he was able to include 34 subjects in his final model. Coe (ibid) presented a graph showing the relative difficulty of his set of 34 GCSE subjects: Latin, German, Spanish and French were the 'hardest'; Sport/PE, Textiles, Drama and Media Studies were the 'easiest'. Somewhat surprisingly (given the greater choice and fewer examinations taken, see below), Coe *et al*. (2008) found that only one of 33 large-entry A level subjects did not fit a unidimensional model.

A different approach is to use a multidimensional model splitting the subjects into more natural groupings either on an a priori basis (e.g. sciences, languages) or on the basis of investigating the statistical dimensionality of the data. This was tried by Korobko *et al*. (2008) using pre-university examinations taken in the Netherlands by 18 year olds (i.e. at a similar stage to A level students). They found that a unidimensional model did not fit the data nearly as well as a multidimensional model (which is not surprising), but more interestingly they found that some implausible results were obtained from the unidimensional model in terms of the 'expected scores' imputed to examinees for subjects they had not chosen to take. For example, average scores in French and German imputed to examinees who had mostly chosen science subjects were nearly as high as those actually achieved by examinees who had mostly chosen language subjects, despite the fact that these science students clearly appeared to have less 'language ability' than the language students on the basis of their scores on the (compulsory) examinations in Dutch and English. This apparent anomaly disappeared when a multidimensional model was used. Korobko *et al*. (ibid) produced tables showing the estimated grade point averages (GPAs) obtained from their models – that is, the average grades in each subject that would have been obtained if all students had taken each subject (interestingly, Latin came out as the 'easiest' subject, whichever model was used!). Nonetheless, the issue of the meaning of the trait and the interpretation of the 'difficulty' parameter still remains, regardless of how well the data fit any particular IRT model. This is discussed again later in this article.

A second issue that weakens the analogy with item-based IRT is that in most applications of IRT where different examinees have taken different subsets of items they have not had any choice in which items they take. For example, in a CAT the next item will be chosen by the computer according to its item selection algorithm, usually taking account of its current estimate of the examinee's ability plus any content coverage requirements. In on-demand testing where tests are constructed from a calibrated item bank there may be a variety of different test forms (versions) but no choice for the examinee in which form they answer. In contrast, for A levels especially, the examinees have enormous choice open to them in which subjects they take. If these choices are not independent of ability (and it would seem unrealistic to expect them to be) then it is not reasonable to assume that the modelled outcome on not-chosen subjects will be adequately predicted by the model. In statistics the 'missing data' literature (e.g. Rubin, 1976) deals with the circumstances under which the mechanism producing the missing data can be ignored. Korobko *et al*. (2008) tried to incorporate a model for the subject choice process into their IRT model:

> *Since the students can only choose a limited number of subjects, it is reasonable to assume that the probability of choosing a subject as a function of the proficiency dimension … is single peaked: Students will probably choose subjects within a certain region of the proficiency dimension … and avoid subjects that are too difficult or too easy.*
> (Korobko *et al*. 2008, p.144).

This assumption was not supported in a large-scale survey of A level students (Vidal Rodeiro, 2007) where liking for the subject and university/career plans were found to be more important than perceived difficulty as factors influencing subject choice. Nevertheless, it does represent an attempt to tackle the missing data issue. In fact, Korobko *et al*. (ibid) found that including a model for the missing data mechanism did not yield substantively different results, once multidimensionality had been modelled (see above).

A third, perhaps less important, difference between item-based IRT and subject-based IRT is that in the former the ability estimate of examinees will be based on the responses to a relatively large number of items – perhaps 60 dichotomous items, or 60 marks-worth of polytomous items. When a small number of subjects is chosen, in contrast, the ability estimate will be based on only a few 'items' (perhaps three to five in the case of A levels). The number of score categories per subject depends on the grading scale used – it is currently seven for A levels since the introduction of the A* category in 2010. Thus the 'maximum score' for an examinee taking three A levels is 21. Whilst this would not normally be considered a sufficient 'test length' for reliably estimating an individual's 'ability' this is perhaps not such a problem when the focus of the analysis is on estimating the difficulty parameters for the items (i.e. the subjects).

Definition of difficulty in the IRT approach

One of the reasons why debates about comparability, standards and subject difficulty have been so protracted and inconclusive is that those involved have often disagreed about the most appropriate definition of these and related terms. That there is this disagreement is of course recognised:

> *… much debate on the comparability of examination standards is at cross-purposes, since protagonists use the same words to mean different things. Within the educational measurement community we have both variants of this problem: the use of the same term to mean different things and the use of different terms to mean the same thing. … There seem to be almost as many terms as commentators.*
> (Newton, 2010, p.289)

Two recent articles by Newton (ibid) and Coe (2010) give thoughtful analyses of these definitional problems. Their arguments will not be repeated here, but one important insight of Newton's is the importance of distinguishing between definitions and methods:

> *An issue that has clouded conceptual analysis of comparability in England, perhaps the principal issue, is the failure to distinguish effectively between definitions of comparability and methods for achieving comparability (or methods for monitoring whether comparability has been achieved).*
> (Newton, 2010, p.288)

The 'IRT approach' as described in this article has been used as a method for monitoring whether comparability has been achieved, by retrospectively analysing examinations data. How was difficulty defined by the authors of the articles that have been described previously?

Korobko *et al*. noted that using GPAs results in

> *... systematic bias against students enrolled in more rigorous curricula ... A lower GPA may not necessarily mean that the student performs less well than students who have higher GPAs; the students with the lower GPAs may simply be taking courses and studying in fields with more stringent grading standards*.
> (Korobko *et al*., 2008, p.144)

While superficially this sounds very reasonable, without a precisely stated definition of what is meant by 'more rigorous' curricula or 'performs less well' or 'more stringent grading standards' there is the suspicion that a lurking circularity could cloud the interpretation of the findings. Nonetheless, it is clear from reading their text as a whole that for Korobko *et al*.: i) subject difficulty is whatever it is that is represented by the difficulty parameter in the IRT model, and ii) once scores (grades) on subjects not taken have been 'imputed' to examinees based on the parameters of the (best fitting) IRT model, the estimated average scores (grades) in each subject can legitimately be compared. To paraphrase, this is equivalent to saying that the rank ordering of examination subjects in terms of difficulty is the rank order by average grade in a hypothetical scenario where all examinees take all subjects. The IRT model is used to simulate this hypothetical scenario.

Coe (2007; 2008; 2010) has given far more consideration to the conceptual issues behind the use of IRT models for comparing examination subjects. Using the concept 'construct comparability' he argues that examinations can be compared in terms of the amount of some common construct implied by given grades. For example, when fitting a 1-parameter IRT (Rasch) model to GCSE subjects, the common construct is 'general academic ability'. If subjects are to be compared (for example on the basis of their difficulty parameters from an IRT model) then this comparison must be stated in terms of the common construct:

> *So rather than saying that maths is 'harder' than English we must say that a particular grade in maths indicates a higher level of general academic ability than would the same grade in English*.
> (Coe, 2008, p.613)

This approach allows Coe to make interpretations of observed statistical differences in subject grading outcomes without having to commit either to a particular definition of difficulty or of general academic ability, since both emerge from the IRT analysis. It also implicitly assumes that 'common construct' is synonymous with 'latent trait'.

Defining difficulty for items in a test

The previous section considered how difficulty has been defined (or its definition has been circumvented) by those employing an IRT approach to investigate difficulty of examination subjects. In this section the issue is approached from the other end – that is, by considering how difficulty has been defined at the item level.

Before IRT became widely used, the framework now known as 'Classical Test Theory' (CTT) was used to analyse data from educational tests. In many contexts CTT is still the preferred choice because in some respects it is conceptually more straightforward, and it is often simpler mathematically, both of which make it easier to explain to non-specialists.

The familiar index of item difficulty in CTT is the 'facility value', defined as the mean mark (score) on a question divided by the maximum possible mark. If the question is dichotomous, the facility value is also the proportion of examinees who answered correctly. Therefore, on a test consisting entirely of compulsory dichotomous items, if question 4 (say) has a higher facility value than question 7, this means that question 4 was answered correctly by more people than question 7. It seems completely uncontroversial to say in these circumstances that question 7 was more difficult than question 4. Because we are dealing with CTT, there is, or seems to be, no need to invoke a latent trait or construct. The qualifier 'for these examinees' might be added, but only in a context where it makes sense to consider the performance of other examinees who did not happen to take the test.

But there are complications possible even for this apparently simple case. First, what can be said if the difference does not hold for identifiable sub-groups? For example, suppose that more males answered question 7 correctly than question 4, but that the opposite was the case for females. In this instance it seems natural just to add the qualifier 'for females, but not for males' to the statement 'question 7 was more difficult than question 4'. A more interesting example is if the group of examinees is split into two groups, 'high scoring' and 'low scoring', on the basis of their overall test score. Now it is again possible for the order of difficulty of the two questions to be different in the two groups, but now adding the qualifier 'for high scorers on the test overall' *does* raise the question of what the test overall was measuring. This is because if question 4 and question 7 were included in a test with different items (but the same examinees) it is conceivable that their relative difficulty with respect to high and low-scoring groups could change.

A second complication with even this simple case is that it does not consider the individual patterns of performance on the two questions, as illustrated by Table 1 below.

Table 1: Question scores for three questions on an imaginary test taken by ten examinees

Person	*Q1*	*Q2*	*Q3*
1	1	1	1
2	1	1	0
3	1	0	0
4	1	1	1
5	1	0	0
6	1	1	0
7	0	0	1
8	0	0	0
9	0	0	0
10	0	0	1
Facility	0.6	0.4	0.4

According to facility values, Table 1 shows that Q1 is easier than both Q2 and Q3, and that Q2 and Q3 are equally difficult. But there is an interesting contrast between Q2 and Q3 in terms of their relationship with Q1. Every person either scored the same or better on Q1 than they did on Q2, whereas this does not hold for the comparison between Q1 and Q3.

Looking at it another way, if a two-item test were made up of the items Q1 and Q2 then knowledge of the total score on this test would

also be knowledge of which items were answered correctly – a person with a score of 2 out of 2 would have got both right, a person with 1 out of 2 would have got Q1 right and Q2 wrong, and a person with 0 out of 2 would have got both wrong. In contrast, on a 2-item test made up of Q1 and Q3, knowledge of the total score would not permit knowledge of which items were answered correctly.

The kind of relationship between Q1 and Q2 was formalised by Louis Guttman in his work on scalogram analysis (e.g. Guttman, 1944; 1950). In brief, a set of items forms a scale if the item scores[2] are a simple function of the scale scores. Guttman was well aware that achieving a 'perfect scale' was not likely in many practical contexts but found that 90% perfect scales (in terms of the reproducibility of the item scores from the scale score) were usable as efficient approximations of perfect scales. (It should be noted that scalogram analysis does not just apply to dichotomous items).

There are two reasons why Guttman's work on scalogram analysis is of interest from the point of view of the present article. The first is that he considered it to be a method for analysing *qualitative* data. It has become so natural for us to think of the data arising from testing as quantitative that we can sometimes lose sight of the fact that the 'raw data', as it were, usually consists of written answers to written questions. Where do the numbers come in? The mark scheme can be thought of as a coding scheme that assigns numerical values (usually integers) to examinee responses according to a certain rationale. One purpose of scalogram analysis is to discover whether the item level data (set of responses across the items) for each examinee can be represented by a single number. (In most examinations this would be the raw score obtained by adding up the scores on each item). If the questions form a scale in the scalogram sense then the scale (total) scores have a definite interpretation in terms of the item scores.

The second reason is that Guttman's starting point was definitions of the universe of attributes (e.g. items) and the population of objects (e.g. examinees) to be scaled. The universe of attributes is the concept of interest whose scalability is being investigated, conceived as the indefinitely large set of questions that could be asked on that concept. Items belong to the universe based on their content, not on statistical criteria. For example, the set of questions testing topics on a particular maths syllabus might define a universe whose scalability could be investigated. The population of objects could be examinees who have studied the appropriate course and prepared for an examination in it. The question of scalability then becomes a matter of empirical investigation that can be carried out on a particular sample of items and examinees. A scalable set of items is by definition unidimensional.

Guttman's approach, in my view, represents the closest thing to 'starting from first principles' in developing definitions of difficulty and comparability. For dichotomous items, if two items P and Q are from a scalable universe then item P is more difficult than item Q if some people (from a defined population) get item Q right and P wrong, but no-one gets Q wrong and P right. Unfortunately, extending even this simple definition to polytomous items runs into problems, as shown in Tables 2a and 2b.

The data for Q4 and Q5 in Table 2a meet the scale definition in that if a scale score is made (e.g. by summing the two responses) then the item scores are perfectly reproducible from the scale scores. Everyone scores at least as well on Q4 as they do on Q5, so Q4 could be said to be 'easier' than Q5.

Table 2a: Question scores for two questions on an imaginary test taken by ten examinees

Person	*Q4*	*Q5*	*Score*
1	2	2	4
2	2	2	4
3	2	1	3
4	2	1	3
5	1	1	2
6	1	1	2
7	1	0	1
8	1	0	1
9	0	0	0
10	0	0	0
Facility	0.6	0.4	

Table 2b: Question scores for two questions on another imaginary test taken by ten examinees

Person	*Q6*	*Q7*	*Total*
1	2	2	4
2	2	2	4
3	1	2	3
4	1	2	3
5	1	1	2
6	1	1	2
7	1	0	1
8	1	0	1
9	1	0	1
10	0	0	0
Facility	0.55	0.5	

However, in Table 2b, although the item scores are perfectly reproducible from the total score it is not the case that everyone scores at least as well on one item as the other. Perhaps the most that can be said is that it is easier to score 1 or more on Q6 than Q7, but easier to score 2 on Q7 than Q6.

This last example makes clear that even the ordering of two items by facility value is ambiguous for polytomous (multiple-mark) items. With a different assignment of scores to response categories, the order could change. For example, in Table 2b if the responses scored '2' were scored '2.8' then Q7 would have a higher facility value than Q6.

To summarise, Guttman's work on scalogram analysis provides a definition of unidimensionality and a definition of what it means for one item to be more difficult than another (for dichotomous items at least).

The link between Guttman and IRT

Unfortunately, item level data from real educational tests never conforms exactly to Guttman's pattern. But there is a strong connection between one particular IRT model, the Rasch model (Rasch, 1960), and Guttman's scale pattern (Andrich, 1985). The expected (i.e. modelled) scores from the Rasch model meet the ordering requirements of the Guttman pattern in that people with higher ability have higher expected scores on every item than people with lower ability, and people of all abilities are expected to score higher on a dichotomous item with a lower difficulty than on one with a higher difficulty. This is not necessarily true for other IRT models. It is also noteworthy that Rasch introduced the concept of

2 The item scores need not be numerical – they could represent responses of 'yes' or 'no' to attitude questions, for example.

'specific objectivity', the 'specific' part of which emphasised that the model only held within a specified frame of reference describing the persons and items, a parallel to Guttman's stressing the need for definitions of the universe of attributes and the population of objects whose scalability was to be investigated.

In fact, Guttman did recognise the concept of a quasi-scale – one where the item responses are not highly reproducible from the scale score but where the 'errors' occur in a gradient (Guttman, 1950), in a manner that seems to conform very closely to the pattern of misfit expected from a Rasch model. The significance of a quasi-scale is that the scale score can still predict an outside variable as well as any weighted combination of the individual item scores (as is the case with a perfect scale). The counterpart of this in Rasch analysis is that the total score is a sufficient statistic for estimating ability (Andersen, 1977) – this means that when the data fit the model there is no additional information about ability in the pattern of item responses. People who have attempted the same items and received the same total score will get the same ability estimate regardless of any differences in scores on the individual items.

This suggests that when data fit the Rasch model, it is possible to define difficulty (for dichotomous items) in a reasonably straightforward way: one item is more difficult than another if any arbitrarily selected person has a lower probability[3] of success on it than on the other item.

As with facility values, and as with the Guttman scale, there is no way round the inherent ambiguity of the concept of difficulty for polytomous items when analysed with a Rasch model. For example, the Rasch partial credit model (Masters, 1982) estimates difficulty threshold parameters representing the points on the latent trait where adjacent score categories are equally probable. There are different possible ways of using these threshold estimates to come up with a number representing 'overall difficulty'. For example, the average of the threshold estimates represents the point on the trait where the lowest and highest score categories are equally probable. Alternatively, it is possible to find the point on the latent trait where the expected score is equal to 'half marks' on the item. Because these are different definitions of difficulty, it would be possible for the ordering of two items to differ depending on which definition was used.

Of course, there is not necessarily any need to produce a number representing 'overall difficulty' – it may be more informative to make comparisons at each category. This was the approach taken by Coe (2008) in comparing relative difficulty of GCSE subjects by grade category. (See Andrich, de Jong and Sheridan, 1997; and Linacre, 2010, for a discussion of some of the issues involved in interpreting Rasch threshold parameters).

While the followers of Rasch seem keen to cite Guttman with approval, essentially regarding the Rasch model as a probabilistic form of the Guttman scale, it is not clear whether this approval was reciprocated. Guttman seemed to avoid using the concept of a latent trait. He also made the following comment about a conventional (CTT) item analysis:

> *This idea of scale construction is a most comfortable one: it is virtually guaranteed to succeed for the kinds of data concerned. I know of no instance in which all items were rejected. In other words, item analysis does not test any hypothesis of scalability. It assumes that scalability exists, and that its task is merely to cull out inappropriate items.*
> (Guttman, 1971, p.343)

3 The interpretation of probability in this context is beyond the scope of this article. See Holland (1990) for some discussion.

Rasch practitioners might feel that this criticism does not apply to them, because they are very keen to stress the primacy of the model over the data (e.g. Andrich, 1989; Wright, 1999), but without an a priori definition of the trait it is probably true in some cases in practice that misfitting items are culled and the resulting set of items provides the 'best' measure of an ill-defined concept. It could be argued that this is what happens when attempts are made to model subject difficulty with the Rasch model (e.g. Coe, 2008; Coe *et al*., 2008). Without starting from a definition of 'general academic ability' it is not clear what the estimated values of subject difficulty with respect to this variable actually mean.

Spatial representations of subject difficulty

For Guttman, it was clear that the dimensionality of the data was something to be discovered rather than imposed. If the empirical evidence showed that two items did not form part of the same unidimensional scale then 'not comparable' was a valid experimental finding. In the later part of his career he developed some of the methods that have become part of the field known as 'multidimensional scaling' or MDS (see, for example, van Deun and Delbeke, 2000). Very broadly speaking, the aim of this kind of analysis is to represent objects in the lowest dimensional space that preserves certain aspects of empirically discovered relationships between them. These relationships could be (for example) indices of similarity or of monotonicity. The final spatial representation might attempt to preserve actual differences in terms of these indices ('metric MDS'), or just their order ('non-metric MDS'). For Guttman, the purpose of these spatial representations was to test hypotheses (made in advance on non-statistical grounds) about how the objects would group into regions of the multidimensional space (see, for example, Schlesinger and Guttman, 1969).

A new direction for investigations of subject difficulty might be to explore such an approach. Given that two objects can always be represented in a single dimension, and generally n objects can be represented in n-1 dimensions, a very simple 2-dimensional example can be contrived by considering 3 subjects. There are several reasonable choices for an index of similarity. If there was no need or desire to maintain any connection with an IRT approach then the difference in mean grade achieved by examinees common to each pair of subjects could be used. This is the index of difficulty familiar from subject pairs analyses (see Coe, 2007, for a description of this and related methods).

However, to stay close to the spirit of Rasch it seems interesting to explore an index of difference that has a close connection with the Rasch model for dichotomous items. In this model, one way of estimating item difficulties is the paired method (Choppin, 1968) where an estimate of the difference in difficulty between any two items A and B is the logarithm of the ratio of the number of examinees succeeding on A and failing on B to the number failing on A and succeeding on B. In the context of examinations rather than items we could choose to make them dichotomous by defining success as 'grade x or above' and failure as 'below grade x'. In the example below the A grade has been chosen as the grade for x. The data have been invented for the purpose of the illustration.

Table 3a shows that 300 people got an A in Psychology but not in Biology, whereas only 50 people got an A in Biology but not in Psychology. On the index of difficulty we are using, Biology is thus log (300/50) ≈ 1.8 logits 'harder' than Psychology.

Table 3a: Biology and Psychology grade A

		Psychology	
Biology	Below A	Grade A	Total
Below A	900	300	1200
Grade A	50	200	250
Total	950	500	1450

Table 3b: English Literature and Biology grade A

		Biology	
English	Below A	Grade A	Total
Below A	400	20	420
Grade A	100	120	220
Total	500	140	640

From Table 3b we see that Biology is log (100/20) ≈ 1.6 logits 'harder' than English Literature, and from Table 3c we see that English Literature is log (160/100) ≈ 0.5 logits 'harder' than Psychology.

Table 3c: English Literature and Psychology grade A

		Psychology	
English	Below A	Grade A	Total
Below A	1100	160	1260
Grade A	100	200	300
Total	1200	360	1560

Because these three differences satisfy the 'triangle inequality'[4] in that the sum of any two differences is larger than the remaining one, it is possible to represent these results diagrammatically as in Figure 1 below.

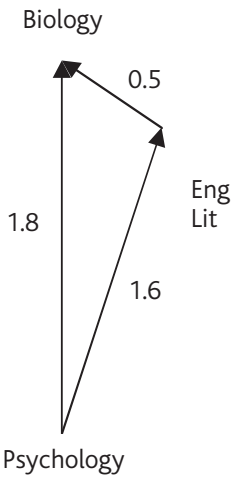

Figure 1: Visual representation of differences in difficulty when triangle inequality is satisfied

4 http://en.wikipedia.org/wiki/Triangle_inequality Accessed 12/04/11.

The length of the arrow represents the logit difference between any two subjects, and the head of the arrow points to the 'more difficult' subject. The closer the three points are to lying on a straight line with arrowheads pointing in the same direction, the more comparable they are as a triplet in terms of difficulty, in the sense that the direct comparison between two subjects is the same as the indirect comparison via a third subject.

Suppose, however, that instead of (1.8, 1.6, 0.5) the three logit differences had been (2.0, 1.5, 0.3). Then the triangle inequality would not have been satisfied and it would not be possible to represent the results as in Figure 1. An alternative depiction of such a scenario is shown in Figure 2.

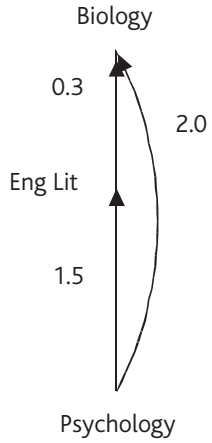

Figure 2: Visual representation of differences in difficulty when triangle inequality is not satisfied

As in Figure 1, the lengths of the arrows represent the logit differences and the heads of the arrows point to the more difficult subjects. (The curved line is part of a circle arc with the straight part as a chord).

With good graphical software it might be possible to represent differences between four subjects (i.e. as a 2D projection of the 'correct' 3D configuration). For higher numbers of dimensions the correct configuration would not be visualisable without either applying some data reduction technique to achieve the best lower dimensional solution according to some criterion, or producing several projections. This is an area for further research.

Conclusion

Using an IRT approach to investigate differences in difficulty among examinations relies on an analogy with using the same approach in its original context – differences in difficulty among items in a test. The software used for the IRT analysis is of course blind to where its inputs have come from and in this sense the outputs of the analysis can be subjected to the usual tests of reliability and model fit.

However, doing this places a greater burden on the analyst to interpret both the latent dimension of the IRT model and the difficulty parameter in that model. This article has shown that it is not entirely straightforward to define difficulty even in the simplest possible case of two dichotomous items in a test. The complications increase as we move to scenarios with polytomous items, scenarios with missing (not presented) items, scenarios with missing (not chosen) items, and finally to scenarios where whole examination subjects are treated as items and there is no a priori defined single trait (dimension) or traits.

This is not to say that an IRT approach is necessarily inadvisable or misleading – the results just need to be interpreted very carefully. It may even be one of the better approaches in cases where there is a pragmatic operational need to produce global rankings of examinees on the basis of overall attainment (as in Tognolini and Andrich, 1996). However, for investigations of differences among subjects, I suggest that it might also be worth going back to the principles first articulated by Guttman, and building up slowly from ground level, considering differences among just a few subjects and representing these visually – searching for stable patterns and always being prepared to accept that 'not comparable' is a reasonable outcome.

References

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, **42**, 1, 69–81.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In: N. Brandon-Tuma (Ed.), *Sociological Methodology*. 33–80. San Francisco: Jossey-Bass.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In: J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.), *Mathematical and theoretical systems*. 7–16. New York: North-Holland.

Andrich, D., de Jong, J.H.A.L., & Sheridan, B.E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In: J. Rost & R. Langeheine (Eds.), *Application of Latent Trait and Latent Class Models in the Social Sciences*. 59–70. New York: Waxmann Münster.

Choppin, B. (1968). Item bank using sample-free calibration. *Nature*, **219**, 870–872.

Coe, R. (2007). Common examinee methods. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. 331–367. London: Qualifications and Curriculum Authority.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education*, **34**, 5, 609–636.

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. (2008). *Relative difficulty of examinations in different subjects*. Report for SCORE (Science Community Supporting Education). Durham: CEM Centre, Durham University.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, **9**, 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In: S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction*. 60–90. Princeton, NJ: Princeton University Press.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, **36**, 4, 329–247.

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, **55**, 4, 577–601.

Kane, M.T. (2006). Validation. In: R. L. Brennan (Ed.), *Educational Measurement*. 17–64. Westport, CT: ACE/Praeger series on higher education.

Korobko, O.B., Glas, C.A.W., Bosker, R.J., & Luyten, J.W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, **45**, 2, 139–157.

Linacre, J.M. (2010). Transitional categories and usefully disordered thresholds. Online Educational Research Journal, **1**, 3. Retrieved from www.oerj.org 11/01/11.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 2, 149–174.

Newton, P.E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, **25**, 3, 285–292.

Newton, P.E. (in press). Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education: Principles, Policy & Practice*.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 3, 581–592.

Schlesinger, I.M., & Guttman, L. (1969). Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin*, **71**, 2, 95–100.

Tognolini, J., & Andrich, D. (1996). Analysis of profiles of students applying for entrance to Universities. *Applied Measurement in Education*, **9**, 4, 323–353.

van Deun, K. & Delbeke, L. (2000). Multidimensional Scaling. http://www.mathpsyc.uni-bonn.de/doc/delbeke/delbeke.htm Accessed 12/04/11.

Vidal Rodeiro, C. L. (2007). *A level subject choice in England: patterns of uptake and factors affecting subject preferences*. Cambridge Assessment report. http://www.cambridgeassessment.org.uk/ca/digitalAssets/114189_Survey_Report_-_Final.pdf Accessed 12/04/11.

Wright, B.D. (1999). Fundamental measurement for psychology. In: S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: what every psychologist and educator should know*. 65–104. Mahwah, NJ: Lawrence Erlbaum Associates.

Yen, W.M., & Fitzpatrick, A.R. (2006). Item Response Theory. In: R. L. Brennan (Ed.), *Educational Measurement*. 111–153. Westport, CT: ACE/Praeger series on higher education.