Multivariate representations of subject difficulty

Tom Bramley Research Division

Introduction

The aim of this study was to explore multidimensional ways of representing similarities and differences in the grade distributions of different A level subjects, in contrast to the more familiar unidimensional ways which are often interpreted as revealing differences in subject 'difficulty'. Of particular interest was whether an a priori scheme for classifying the subjects would be reflected in the multidimensional configurations.

Debate over whether some examination subjects are 'harder' than others has been around for a long time. Newton (2010) has previously noted the importance of distinguishing between definitions of comparability, and methods for monitoring whether it has been achieved. Statistical methods for monitoring inter-subject comparability have been both used and criticised (see Coe 2007, 2010). Where statistical methods are used, the aim is usually to produce a *single* ranking of subjects according to an indicator of difficulty. For example, the report by Coe, Searle, Barmby, Jones, and Higgins (2008) contained many tables showing the results of such rankings from research exercises carried out in the 1970s, 1980s, 1990s and 2000s using a variety of different statistical methods. Their conclusion (for A levels) was that although different methods did give slightly different results (rankings of subjects by difficulty), the differences between methods were much smaller than the differences in difficulty between subjects that the methods revealed: "the argument that the different methods do not agree is not a convincing reason to use none of them." (Coe *et al*., 2008, p.89).

Focusing on one particular (class of) statistical method, the 'Item Response Theory (IRT) approach', Bramley (2011) explored the analogy between item difficulty (which IRT methods were developed to model/measure) and subject difficulty, concluding that using IRT methods for the latter places a greater burden on the analyst to interpret the meaning of the latent trait and the difficulty parameter in the IRT model than is the case for 'normal' use of IRT. This is mainly because examined subjects at a particular level (e.g. A level) form a largely *ad hoc* collection, in contrast to the set of items on a particular examination which have been designed to assess a syllabus and cover a range of topics and difficulties with a target population of examinees in mind. In view of this, Bramley *(ibid)* suggested exploring ways of representing subject difficulty graphically, without aiming to produce a single overall ranking of subjects by difficulty.

The study reported here followed up that suggestion by applying the technique known as multidimensional scaling (MDS)[1] to data from OCR A levels taken in June 2011. The MDS results were compared with those from two unidimensional methods (the Kelly method and the Rasch method) that give a single ranking of subjects in terms of difficulty.

1. The ideas behind MDS have been developed independently by different researchers in different places and consequently there is a variety of terminology in use.

Classifying A level subjects

The A level subjects were classified in advance into categories in order to see if the location of subjects in the unidimensional rankings or the MDS representations corresponded to these a priori classifications. There are obviously many different ways in which A level subjects could be categorised, all of which would be to some extent arbitrary. For example, the list of 'academic disciplines' (not A levels) on Wikipedia[2] has the following high-level groupings:

- Humanities (e.g. History, Philosophy, Performing Arts)
- Social sciences (e.g. Economics, Psychology, Anthropology)
- Natural sciences (e.g. Physics, Chemistry)
- Formal sciences (e.g. Computer sciences, Mathematics, Logic)
- Professions and applied sciences (e.g. Agriculture, Law, Engineering).

The problem with the above list is that it is more appropriate for university disciplines than A level subjects. Languages would only appear indirectly as 'Linguistics' or 'Literature' within the humanities, whereas they seem to form a more definite category of A level subject.

Taking a Facet Theory approach (e.g. Borg & Shye, 1995) to producing a categorisation scheme would require identifying a rule or rules by which a given A level could be unambiguously allocated to a category. Following discussion with colleagues of various categorisations currently in use, and given an aim to have some fairly uncontroversial and intuitive categories, the categorisation in Table 1 below was used for this research.

The STEM classification seemed fairly self-explanatory, even though no rule was created. Problem cases were Geology (classified as STEM), Psychology (classified as a Humanity) and Applied Science (classified as Applied).

2. http://en.wikipedia.org/wiki/List_of_academic_disciplines (Accessed 13/03/14).
3. Science, Technology, Engineering and Mathematics – a grouping often used in media reporting.

Table 1: Classification of A level subjects into categories

Category	*Rule*	*Examples*
STEM3		Maths, Physics, Computing
Humanities	Knowledge, skills & understanding expressed mainly through extended writing	English Literature, Classics, Media Studies, Psychology.
Languages	Require learning some of the vocabulary and grammar of a second language.	Latin, French, Spanish, Turkish.
Expressive	Knowledge, skills & understanding expressed mainly through performances or artefacts	Music, Design and Technology, Art and Design, Performing Arts.
Applied	Knowledge, skills & understanding lead more directly to jobs or job-related further courses.	Accounting, Health & Social Care, Applied ICT, Law.

The list of subjects in the Humanities category seemed reasonable enough, although the classification rule itself would not be good enough to unambiguously allocate subjects to the category.

The Languages category was also fairly straightforward, although it requires assuming that the first language of A level examinees is English (which was classified as a Humanity for this research). For some language A levels (e.g. Turkish) it seems possible that a significant proportion of native speakers may take the A level, but this does not so much cast doubt on the validity of the classification rule, but on the validity of inferences made about the relative difficulty of some language A levels using statistical methods.

The Expressive categorisation was more problematic in that Design and Technology could perhaps also fit in the STEM or Applied categories, and that in some cases it is perhaps doubtful whether knowledge, skills and understanding are expressed *mainly* through performances and artefacts (as opposed to through written responses).

The Applied category was relatively straightforward on the assumption that subjects with the word 'applied' in their specification title are indeed intended to lead more to jobs or job-related further study than to academic study, as per the classification rule for this category.

Unidimensional representations

Kelly method

The Kelly method (Kelly, 1971) is a relatively straightforward way of deriving rankings of subjects by difficulty. It is used by the SQA to obtain rankings of Scottish Highers. The method is described in technical detail by Coe (2007). Basically, the output of the method is a difficulty rating for each subject which can be interpreted as the adjustment that should be made to the (numerical[4]) grades in each subject in order that, on average, examinees achieve the same average adjusted grade in their other subjects that they achieve in any particular subject. A positive value therefore indicates a more difficult subject (defined by this method as one in which examinees on average obtained lower grades than in the other subjects they took).

The analysis used a sub-set of 33 of the OCR A level specifications from the June 2011 examination session. For subjects with more than one specification, the one with the larger entry was retained. Specifications with fewer than 400 examinees taking at least one other OCR A level were dropped, with the exception of German and Spanish, which were retained so that the category of Languages would be better represented. Table 2 shows the Kelly difficulty ratings of these 33 subjects, colour coded by higher-level category. The change in rank position (out of 33) from 2010 to 2011 is also included.

Inspection of Table 2 shows that Kelly difficulty rating was related to category, with (in general) STEM subjects and Languages being more difficult, Expressive and Applied subjects being easier, and Humanities generally in the middle, with the exceptions of General Studies and Critical Thinking being more difficult, and Sociology and Media Studies being easier. A plausible explanation for the relative difficulty of General Studies is motivation – if examinees do not try as hard or prepare as well for this exam then it will appear harder. Similarly Critical Thinking may suffer from both motivation effects, and a lack of teaching time and teaching experience (see Black, 2009). The stability of the ranking from 2010 to 2011, as shown by the fact that no subject changed by more than three places in the overall ranking, is indirect evidence of within-subject standard maintaining from year to year.

4. Letter grades are converted to numbers on an interval scale: A*=6, A=5, ... E=1, U=0.

Table 2: Difficulty ratings (Kelly method) of 33 OCR A level specifications in June 2011

	Category	*Assessment Name*	*Difficulty from 2010*	*Change*
1	STEM	Further Mathematics	0.95	=
2	Humanities	Critical Thinking	0.74	=
3	Humanities	General Studies	0.60	+1
4	STEM	Physics A	0.49	-1
5	STEM	Chemistry A	0.46	=
6	STEM	Biology	0.23	=
7	Languages	Classics: Latin	0.19	+1
8	Languages	French	0.18	+2
9	Expressive	Music	0.12	+3
10	Languages	German	0.11	-3
11	STEM	Computing	0.06	-2
12	STEM	Mathematics	0.04	-1
13	Languages	Spanish	0.02	=
14	Applied	Applied ICT	-0.02	=
15	Humanities	Economics	-0.09	=
16	Humanities	History A	-0.18	+2
17	Humanities	Government And Politics	-0.21	-1
18	Humanities	Classics: Classical Civilisation	-0.25	-1
19	Humanities	Geography	-0.27	+1
20	Humanities	Psychology	-0.31	=
21	Humanities	English Literature	-0.33	-1
22	Humanities	Religious Studies	-0.35	-3
23	Applied	Physical Education	-0.43	+2
24	STEM	Geology	-0.48	-1
25	Applied	Law	-0.49	-1
26	Applied	Business Studies	-0.62	+1
27	Expressive	Performance Studies	-0.65	+2
28	Applied	Health And Social Care	-0.66	-2
29	Expressive	Design And Technology: Product Design	-0.67	-1
30	Humanities	Sociology	-0.85	=
31	Expressive	Art And Design: Fine Art	-0.95	=
32	Humanities	Media Studies	-1.01	=
33	Expressive	Art And Design: Photography – Lens And Light-Based Media	-1.37	=

Rasch method

The Rasch method for comparing subject difficulty is also described in Coe (2007) and Bramley (2011). The Rasch model characterises persons and items (here, A level specifications) by a single number that can be taken as representing their location on the overall construct that is being measured by the items. In this case, the overall construct has to be interpreted as something like 'general academic ability' (Coe, 2010).

The Rasch Partial Credit model (PCM) (Masters, 1982) was fitted to the A level data, which instead of the usual examinee × item matrix contained examinees on the rows but A level specifications in the columns, with the data being the numerical grade obtained by the examinee in that specification. The matrix was large and contained mostly missing data (as examinees took at most five A levels). The data was analysed with the FACETS program (Linacre, 1987).

Although the input is identical to that for the Kelly analysis, the Rasch analysis is more complex in that an explicit model is fitted, and parameters are estimated for the thresholds between each grade category. There is therefore no single difficulty of an item estimated with the PCM, although it is conventionally taken as the mean of the threshold parameters. An interpretation of this mean value is that it is the ability level at which obtaining a grade in the bottom (U) or top (A*) categories is equally likely (Linacre, 2005). Higher values therefore indicate more difficult subjects, but the logit (log odds) scale is less readily interpretable than the Kelly output (which is in terms of numerical grades).

As Coe *et al.* (2008) found, the Kelly and Rasch results were very similar. The correlation was 0.90, which rose to 0.96 when outliers were excluded.[5] Unlike the Kelly method, the Rasch method also produces an indication of how well each person and item (here, A level subject) has fit the model. An 'overfitting' item or person is one whose observed responses conform more closely to the model than expected, given its probabilistic nature, whereas an 'underfitting' or 'misfitting' item or person is one whose observed responses confirm less well to the model than expected. A 2-dimensional representation of the Rasch results can thus include both difficulty and fit[6], as shown in Figure 1 for the 33 subjects.

Figure 1 shows that as well as being more difficult, STEM subjects tended to overfit the Rasch model. The Languages, and the subjects classified as Expressive tended to underfit (misfit) the model. Interestingly the two most difficult Humanities subjects (General Studies and Critical Thinking) also had large values for the misfit indicator, supporting the earlier conjecture that factors other than general academic ability might have affected the observed grades in these subjects. The Humanities and Applied subjects generally seemed to fit the model reasonably well.

5. These outliers were specifications containing grade categories with no examinees in them (usually U or A*). The Rasch analysis 'collapses' such empty categories when they occur, thus changing the meaning of some of the parameters and hence of their average value.

6. The fit statistic shown in Figure 1 is the infit-z statistic output from FACETS. Negative values indicate overfit, positive values misfit.

Multidimensional representations

Although no one denies that the results of unidimensional representations such as those of the Kelly or Rasch methods have produced stable outcomes over a long period of time, there is much more disagreement over the interpretation, utility and implications of such results. Detailed discussion can be found in Coe (2007) and Newton (2010). The main purpose of the present study was to explore whether anything might be gained from setting aside the potentially inflammatory search for a single ranking of subjects by difficulty and looking for other ways to summarise or characterise the same underlying data (i.e. the grades obtained by OCR A level examinees in each specification). The technique explored was multidimensional scaling (MDS).

MDS is actually a set of techniques that have the common aim of representing indices of similarity or dissimilarity between a set of objects as a spatial configuration of points, where the points represent the objects, and the distances between points in the configuration reflect relationships between the indices of similarity or dissimilarity. See Kruskal and Wish (1978) or Borg and Groenen (2005) for detailed explanations of MDS concepts, formulas, applications and issues. There are many choices that have to be made when carrying out an MDS analysis, including:

- the function relating similarities/dissimilarities to distances in the MDS configuration;
- which index of similarity or dissimilarity to use;
- the dimensionality of the configuration;
- whether to give equal weight to all the data, or more weight to some points and less to others.

For all the analyses, a non-metric (i.e. ordinal) function was specified – this imposes the fewest constraints on the analysis. The aim of non-metric MDS is to preserve rank-order relationships as far as possible in the MDS configuration. For example, if (according to the index of

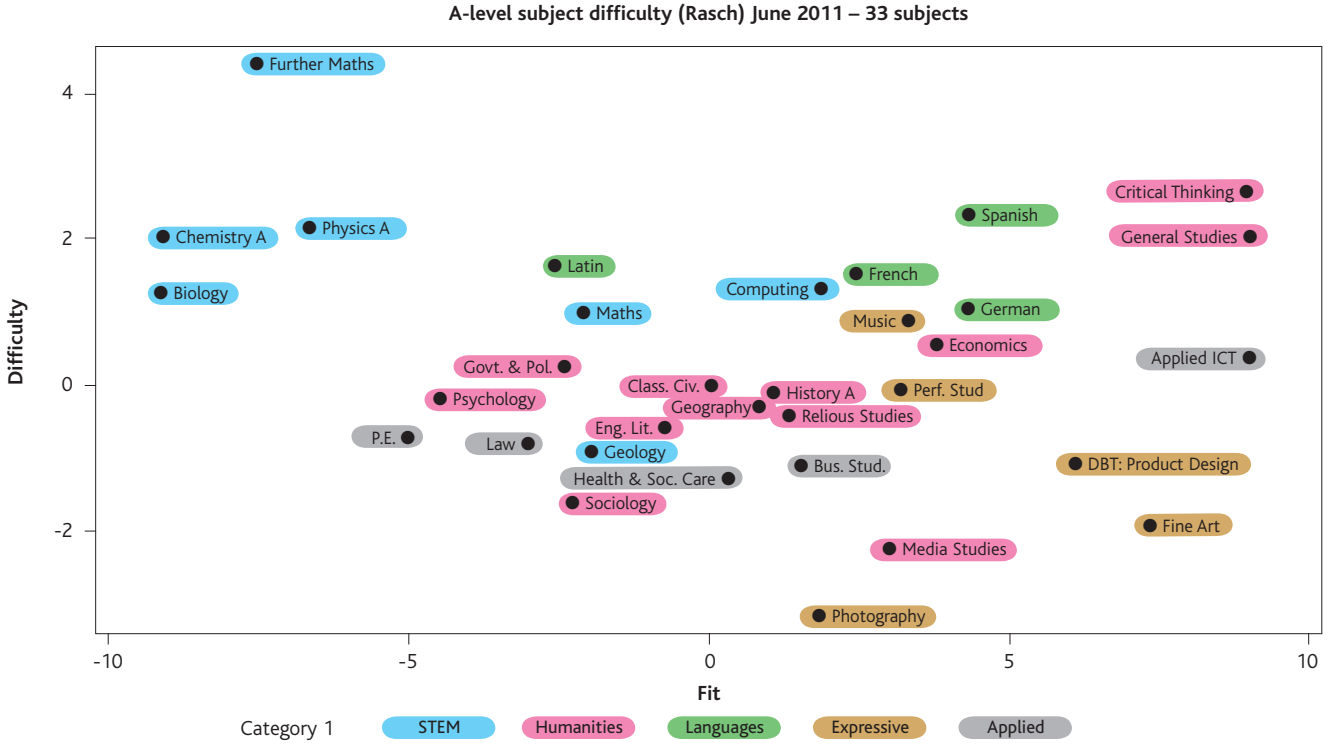

Figure 1: Plot of Rasch average difficulty v fit.

similarity used) subject P is more similar to subject Q than subject R is to subject S, then in the MDS configuration, P will be as close or closer[7] to Q than R is to S.

Three indices of similarity between pairs of subjects were explored here. First, the absolute (unsigned) difference between mean numerical grade obtained by common examinees. The (signed) difference between mean grades is familiar from subject-pairs analyses (e.g. Forrest & Smith, 1972), the idea being that if a group of examinees obtains (on average) a higher grade in subject P than in subject Q, then subject P is less difficult than subject Q. For the analysis here, however, only the size of the difference between each pair of subjects was preserved, not the direction of the difference.

The second index was the proportion of common examinees obtaining exactly the same grade in subject P as subject Q (denoted here as P_0). Clearly, the higher this index the more similar it can be argued the two subjects are – but it does not address difficulty *per se* because it does not take into account what grades were obtained by common examinees who did *not* get the same grade. It is in principle possible for two pairs of subjects (PQ and RS) to have the same value for P_0, but for the majority of common examinees in one (say PQ) who did *not* get the same grade in P as Q to get a better grade in P than Q, whereas in the other (RS) for common examinees who did not get the same grade to be equally distributed among those who had got a better grade in R than in S and those who had got a better grade in S than in R.

The third index of similarity was Guttman's coefficient of monotonicity μ_2 (Guttman, 1977). This is essentially an ordinal correlation coefficient, ranging from -1 to +1. The formula is given in the appendix. It takes its maximum value of +1 when an increase in one variable is always associated with an increase (or no change) in the other. As with P_0, it does not address difficulty – two subjects could have a perfect monotonic correlation between the grades of common examinees, but the grades obtained in subject P might be systematically higher (or lower) than in subject Q. The μ_2 coefficient has been the index of similarity favoured by many practitioners of Facet Theory because it requires no assumptions of interval-scale measurement or of linear relationships.

Solutions for 1 to 4 dimensions were investigated in each case, to gain a feel for how much information was being lost by reducing the dimensionality. It seemed sensible to give more weight in the analyses to indices of similarity from pairs of subjects with large numbers of common examinees, on the assumption that common examinees from such subject pairs were more likely to be representative of the general examinees in those subjects, and the view that it was in general more important to give weight to the larger-entry subjects. The software used to run the analyses was the PROC MDS procedure in SAS 9.2. The default options were used[8].

1. Similarity in mean grade of common examinees

The two-dimensional MDS solution had a value for the 'stress' (badness of fit) statistic just under 0.20. The value of 0.20 is given by some sources as a rule of thumb for an acceptable or adequate fit, although most sources emphasise that (as with any complex statistical method), rules of thumb are often misleading as stress can be affected by a number of factors, such as the number of points being represented and error in the proximities. However, there is agreement that the main purpose in exploratory MDS is to arrive at an interpretable visual representation, which means that in practice usually only 2- and 3-dimensional solutions are considered. For the other indices of similarity (see later) the 2-D stress was above 0.2, so 3-D representations were considered for those.

It can be seen from Figure 2 that the location of points along Dimension 1 happened to correspond closely to the ordering of difficulty

7. This is the 'weak monotone' function most commonly used (Borg & Groenen, 2005, p.40).

8. See http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mds_sect004.htm for a description of these default options.

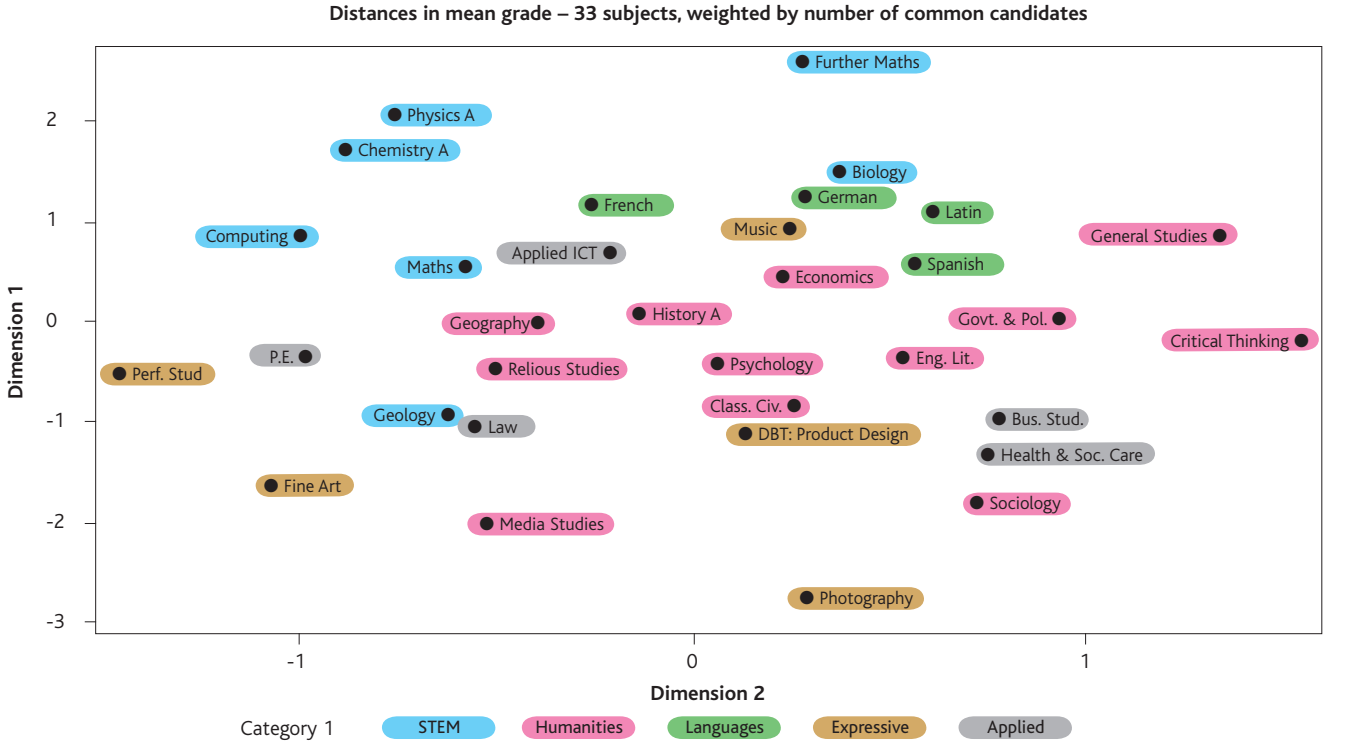

Figure 2: Two-dimensional non-metric MDS representation of differences in mean grade.

from the Kelly and Rasch methods ($r \approx 0.9$). Dimension 2 however did not correlate highly with the fit to the Rasch model ($r \approx 0.2$), as can be seen by comparing Figure 2 with Figure 1. Of course, the axes in an MDS solution have no intrinsic meaning – it is the distances between points that are relevant. (In other words, the configuration in Figure 2 could be rotated or reflected without affecting the fit of the solution). Nonetheless, it is still reasonable to look for any interpretable directions across the configuration so the relationship with difficulty is interesting, particularly since it emerged without 'telling' the software which subject in each pair had the higher mean grade. The STEM subjects, the Languages and to a lesser extent the Humanities do seem to group together in Figure 2, suggesting that within these groupings, differences in mean grade of common examinees were more similar than across groupings. There is no obvious pattern for the subjects classified as Applied, but for the Expressive subjects there is a tendency for them to be on the edge of the configuration, suggesting greater differences between these subjects and the others.

2. Similarity in percentage of common examinees with the same grade

The 2-D solution had a stress value of around 0.24, but the 3-D solution had a value around 0.16, suggesting that three dimensions were needed to adequately preserve the relationships between similarity of P_0 values. The 3-D representation is shown in Figure 3 below.

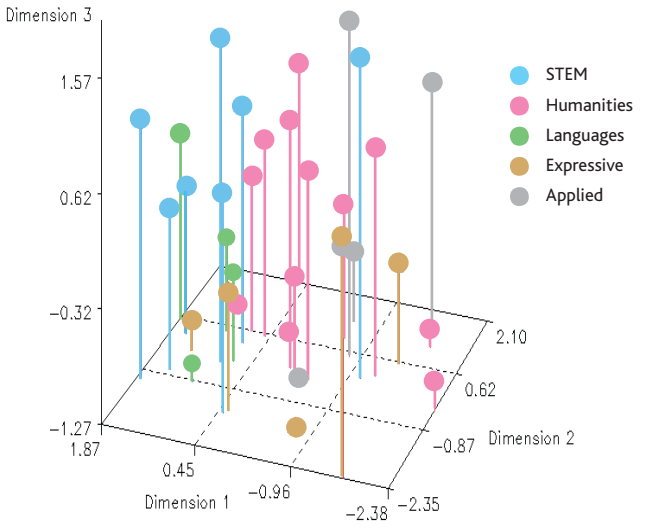

Figure 3: Three-dimensional non-metric MDS representation of differences in P_0

In Figure 3 the STEM (except Geology) and Language subjects seem to group well, and the Humanities reasonably well. The Expressive and Applied are less clearly grouped but still closer (by eye) than a random allocation of points. Interpreting static 3-D representations on a 2-D surface is not easy, so rotating the graph (possible with most modern graphics software) can make it easier to look for patterns.

3. Similarity of coefficient of monotonicity between grades of common examinees

The 2-D solution had a stress value of around 0.26, but the 3-D solution had a value around 0.18, suggesting that three dimensions were needed to adequately preserve the relationships between similarity of μ_2 values. There was some discernible clustering by group, but not quite as clear-cut as for the P_0 similarity index depicted in Figure 3.

Discussion

The MDS analyses have shown that 2- or 3-D representations of aspects of the raw data (the A level grades obtained by OCR examinees) do highlight groupings of the subjects in terms of categories that can be identified prior to analysis. Although there were differences in the patterns across the MDS representations using different indices of similarity, at a broad level the same findings were observed – that is, STEM subjects, Languages and Humanities clustering together fairly well in the representations, Expressive and Applied subjects less well.

Did increasing the dimensionality beyond the usual one (interpreted as 'difficulty') yield new insights? Unfortunately the difficulties in interpretation remained, and this is an intrinsic feature of the data at hand: examinees choose a small and very non-random subset of the possible subjects. Twenty seven pairs of subjects had no common examinees, and 167 pairs (of the 528) had fewer than ten. Only seven pairs had more than 1,000 common examinees, and these all involved STEM subjects and General Studies. Clearly it is impossible to create the 'ideal' situation where all examinees take all subjects. We can therefore never know whether some pairs of subjects would have higher (or lower) indices of similarity if more examinees had taken both of them.

The MDS methods do not of themselves permit an interpretation of the dimensions of the configuration – it is the distances between the points that should be interpreted. Nevertheless, it can be hard to resist the temptation to look for a 'difficulty' dimension, given the stable Kelly and Rasch findings. It was interesting that one of the dimensions in all three of the MDS analyses seemed to be fairly closely related to unidimensional difficulty, given that only the first of the three indices of similarity was directly related to difficulty. However, the input for calculating all three indices of similarity was essentially the same – the 7×7 cross-table of grade in subject X against grade in subject Y with cells of the table containing the number of examinees containing the corresponding pair of grades in the two subjects, as shown in the example in Table 3.

The first index of similarity, absolute difference of mean grade, only uses the information in the margins of the table: abs [(257×6 + 472×5 ... + 112×1) – (240×6 + 455×5 ... + 92×1)]/1763.

The second index, P_0, the proportion of examinees achieving the same grade in both, only uses the information in the shaded blue top-left to bottom-right diagonal and the overall number of common examinees: (161 + 269 +... + 14)/1763. The third index, Guttman's μ2, takes into account the frequencies in each cell of the table, as shown in the second formula in the appendix. Although the different indices therefore use different aspects of the table, they are not independent. For example, if Physics were graded more leniently, more examinees would move into the cells above and to the right of the shaded diagonal. This would increase the mean grade difference and decrease P_0 (assuming that more examinees would move out of the shaded diagonal than into it). Larger differences in difficulty are therefore likely to correspond to lower values of P_0, and hence it is perhaps not surprising that one direction in the MDS configurations correlated well with unidimensional difficulty.

Future work could verify that the Rasch and Kelly results continue to show a stable pattern, and explore whether the 2- and 3-D MDS configurations also show stability. If there is a clearly identifiable 'background of stability' this could prompt investigations of any subjects that appear to be moving against the stable background – for example this might signify changing entry patterns, or changing grading standards.

Table 3: Cross-tabulation of grades obtained by examinees taking both Physics and Chemistry in June 2011

		Chemistry							
		A (6)*	*A (5)*	*B (4)*	*C (3)*	*D (2)*	*E (1)*	*U (0)*	*Total*
Physics	*A* (6)*	161	86	9	1	0	0	0	257
	A (5)	74	269	115	12	2	0	0	472
	B (4)	5	89	198	71	14	2	0	379
	C (3)	0	10	95	123	51	12	0	291
	D (2)	0	1	15	72	100	22	8	218
	E (1)	0	0	3	14	39	42	14	112
	U (0)	0	0	0	1	5	14	14	34
	Total	240	455	435	294	211	92	36	1763

However, the index of similarity for subjects with small entries is always likely to be unstable and therefore the relative positioning of such subjects is likely to fluctuate. Another extension of this work could be to try other categorisations of subjects to see if there are some that lead to cleaner/sharper delineations of regions in the resulting spatial representations. To stay within the spirit of Facet Theory there would ideally need to be a rule or principle by which the categorisations could be applied.

In conclusion, the MDS representations of A level subjects according to various indices of similarity derived from the joint grade distributions of common examinees are interesting, but perhaps have too many difficulties attached to their interpretation to be worth pursuing. It may ultimately be easier to interpret trends and patterns in the indices of similarity directly, perhaps via consideration of cross-tabulations of pairs of subjects like those shown in Table 3 above. If a visual representation of a large number of subjects is desired then in my opinion the 2-D plot of the Rasch results (Figure 1) is the most informative, because it both allows a specific interpretation of 'difficulty', but also clearly shows the caveats in terms of the large differences in fit – which also are systematically related to subject groupings.

References

Amar, R. (2005). *HUDAP mathematics. 3rd edition*. Jerusalem: The Hebrew University of Jerusalem Computation Authority.

Black, B. (2009). *Introducing a new subject and its assessment in schools: the challenges of introducing Critical Thinking AS/A level in the UK*. Paper presented at the Association of Educational Assessment-Europe Conference, 7th November 2009, Malta.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: theory and applications*. New York: Springer Verlag.

Borg, I., & Shye, S. (1995). *Facet Theory: form and content*. Thousand Oaks: CA: SAGE.

Bramley, T. (2011). Subject difficulty – the analogy with question difficulty. *Research Matters: A Cambridge Assessment Publication, Special Issue 2: comparability*, 27–33.

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*. (pp.331–367). London: Qualifications and Curriculum Authority.

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*. Durham: CEM Centre, Durham University.

Forrest, G. M., & Smith, G. A. (1972). *Standards in subjects at the Ordinary level of the GCE, June 1971*. Occasional Publication 34. Manchester: Joint Matriculation Board.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26*(2), 81–107.

Kelly, A. (1971). The relative standards of subject examinations. *Research Intelligence, 1*(2), 34–38.

Kruskal, J., & Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.

Linacre, J. M. (1987). FACETS (Version 3.67.1): www.winsteps.com.

Linacre, J. M. (2005). The partial credit model and the one-item rating scale model. *Rasch Measurement Transactions, 19*(1), 1009.

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.

Appendix: Guttman's coefficient of monotonicity, μ_2

$$\mu_2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j||y_i - y_j|}$$

where x_i is the numerical grade of examinee i on exam X, y_i is the numerical grade of examinee i on exam Y, and N is the total number of common examinees.

or, from an $n \times m$ cross-tabulation of frequencies on X and Y:

$$\mu_2 = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{m-1}\sum_{l=k+1}^{m}(f_{ki}f_{lj} - f_{li}f_{kj})(\xi_i - \xi_j)(\psi_k - \psi_l)}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{m-1}\sum_{l=k+1}^{m}(f_{ki}f_{lj} + f_{li}f_{kj})|\xi_i - \xi_j||\psi_k - \psi_l|}$$

where ξ_i is the numerical grade of the ith category of exam X, ψ_k is the numerical grade of the kth category of exam Y, and f_{ki} is the number of examinees obtaining grade category k on exam X and grade category i on exam Y.

The above formulas are taken from the reference manuals for the HUDAP software package (Amar, 2005).