Using Item Response Theory (Rasch Model) to investigate the existence of bottlenecks at the various levels of the P-scales.

Francis Ndaji and Peter Tymms
Curriculum, Evaluation and Management Centre, University of Durham, UK

The P-scales performance criteria, first published by the DfES/QCA in 1998 describe attainment levels for pupils that achieve below age-related expectations. The levels described range from P1 through P8 to Level 2 of the national curriculum.

In this study, Item Response Theory has been used in determining whether there is a bottleneck at P8, and indeed, at any level of the P-scales. This is in response to the suggestion by some teachers that there is a bottleneck at P8 because a good number of their pupils apparently found it very difficult to progress from P8 to the next higher level. Item Response Theory has been applied to this study. The study shows that there is no bottleneck at P8 or any level of the P-scales. The Partial Credit Rasch Model seems to work for tests of the levels description variety, such as the P -scales.

Introduction

The P-scales (Performance Scales) criteria were first published in 1998 by QCA/DfES. The scales describe attainments below level 1 of the national curriculum and applied to English (three strands), Mathematics (three strands) and Personal and Social Development (PSD) (three strands). The criteria for English and Mathematics were presented as 8-point performance scales (P1-P8) below level 1, and three additional levels, namely, levels 1, 2, and 3 of the National Curriculum. PSD was presented as 15-point performance scales(1).

Prior to the introduction of the P-scales in 1998, schools used the code W (working towards level 1) during the collection of statutory end of Key Stage test/task results to describe the attainments of pupils working below level 1 of the National Curriculum(2). Pupils for whom the tests/tasks were judged inappropriate were excluded from the statutory assessment altogether, leading to the loss of important information. Another disadvantage of using the code W is that it would not permit the exact attainment or progress of pupils to be obtained because W would not tell how

far below level 1 the pupil has achieved. Therefore a new system of assessment was required that would enable teachers measure the attainments and progress of their pupils who are working below level 1 of the National Curriculum. It was in response to that need that QCA and DfES introduced the P-scales.

The P-scales criteria were revised in March 2001 and published in a booklet titled *Supporting the Target Setting Process: Guidance for effective target setting for pupils with special education needs*(3). One of the highlights of the revision was the sub-division of the three lowest levels P1, P2 and P3 of the cognitive scales into six levels, namely, P1(i), P1(ii), P2(i), P2(ii), P3(i) and P3(ii). This sub-division of P1, P2, and P3 was carried out in order to increase the sensitivity of the scales at these lowest levels of attainment and progress. Other important features of the revised criteria were the introduction of Science (four strands) and the presentation of PSD as one scale of 8-point performance scale under a new name PSHE (Personal, Social, and Health Education). Hence the P-scales criteria currently describe attainments below level 1 in English (three strands: Speaking & listening, Reading and Writing), Mathematics (three strands: Using & applying mathematics, Number, and Shape, space & measures), Science (four strands: Scientific enquiry, Life processes and living things, Materials and their properties, Physical processes) and PSHE. English, Mathematics and Science also describe attainment levels up to level 2 of the National Curriculum. A minor revision was carried out in June 2004 (available on QCA website), and QCA will be issuing moderation materials and advising schools on moderation procedures in Spring 2005.

Since 1999 QCA has organised the annual collection and analysis of P-scales assessment data through the CEM Centre, University of Durham. The aims of the annual data collection are (a) to collect enough data in order to present a national picture of the performance of pupils working below age-related expectations, and (b) from the resulting dataset, prepare feedback for schools to help in their self-evaluation and target setting. The pupils involved were 5-16 years old that were classified as having one or a combination of special educational needs. Performance data based on the P-scales criteria were collected from special schools and mainstream schools with special units, and for the purpose of analysis were re-coded into numbers as follows, P1(i)=1, P1(ii)=1.5, P2(i)=2, P2(ii)=2.5, P3(i)=3, P3(ii)=3.5, P4=4, P5=5, P6=6,

P7=7, P8=8, L1c=9, L1b=10, L1a=11, L2c=12, L2b=13, L2a=14, L3=15, L4=16,L5=17,L6=18. There has been a statutory requirement(4) on all schools since 1998 to set performance targets in relation to national expectations for raising attainment. The P-scales data collection exercise and the resulting feedback have been perceived to be very useful to schools in that regard. The importance of the feedback to schools is captured in the following comments taken from many sent in by headteachers.

(a) *The feedback has assisted the school in setting objective-led targets and assessing progress for children in the Early Years Reception class and Year 1/Year 2 class.*

(b) *The school has been using the levels extensively since their initial publication and CEM Centre data (the feedback) greatly helps in the continuing evaluation of their use and the school's progress.*

(c) *Participation in the QCA P-scales data collection project at Durham University promises to be extremely useful in monitoring the efficacy of all teaching and learning at our school. It will furnish us with tangible evidence of individual pupil progression, assist in the setting of realistic targets for cohorts of pupils……..the data will afford key insights into whole school performance over time and provide us with an effective mechanism for the process of self-evaluation and hence school improvement.*

Although the P-scales data collection and analysis have been successful as attested by schools, there have been suggestions by teachers that a bottleneck exists between level P8 of the P-scales and level 1 of the national curriculum. This suggestion arose as a result of the observation by the teachers that some of their pupils apparently found it very difficult to progress from P8 of the P-scales to the next level which is level 1 of the National Curriculum.

This study was therefore motivated by the need to ascertain whether or not a bottleneck exists at P8 or indeed, at any level of the P-scales.

Method

The measurement model (Rasch Model) proposed by a Danish Mathematician, George Rasch(5), in 1960 has been applied to this study. The theory was based on the assumption that a mathematical relationship (model) can be found that would explain the relationship between the ability of a person attempting test items (person parameter), one or more characteristics of the items (item parameters) and the likely outcome of the attempts. The model may be used for dichotomous questions, usually scored right or wrong and multi-choice objective questions (6). However, equating and other types of errors could occur if the model and data do not fit(7). Further benefits of IRT methods may not be realised and information derived from the model may be unreliable unless the model adequately fits the data. Therefore it is necessary to determine whether the P-scales data fit the Rasch model before taking the study any further.

The P-scales data have been analysed using the software called WINSTEPS.

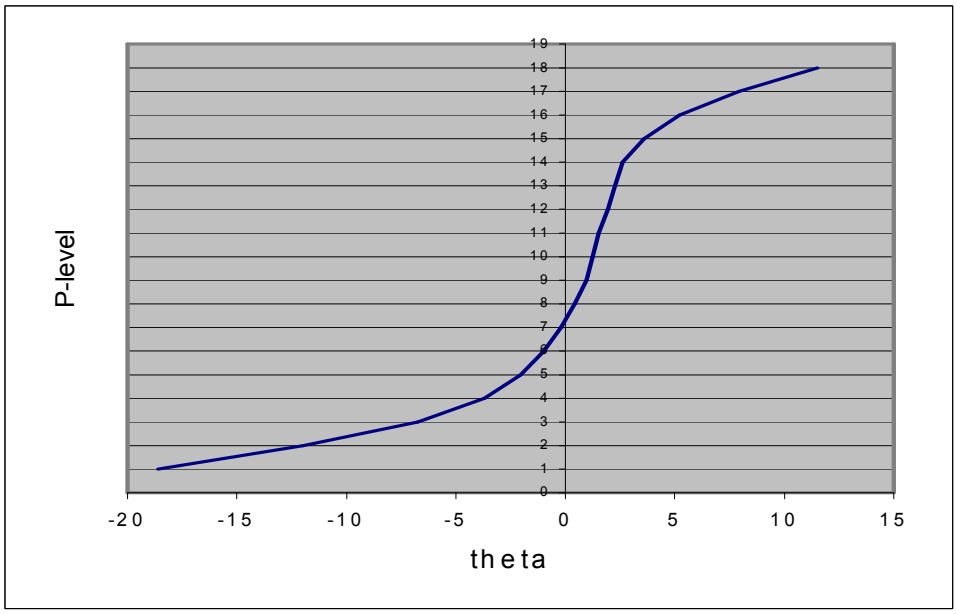
If the data and model fit, the item parameter will be person-independent and the person parameter will be item-independent (6). To test for the person-independence of the item parameters the item difficulties obtained from the scores on Writing by two groups of pupils (males and females) were subjected to an Independent samples t-test and the means of the item difficulties were found to be statistically equivalent ($t_{(14)}=0.16$, $p=0.870$) despite the fact that the means of the raw scores for the boys and girls were statistically different. Also, a very high correlation (correlation coefficient = 0.98) was found between the item difficulties obtained for the two groups of pupils - indicating that the two sets of item difficulties do not vary greatly. These show that item difficulty is independent of the two groups taking the test, suggesting that the Rasch model can also be applied to tests consisting of level descriptions such as the P-scales.

Results and Discussion

Estimated Item Characteristic Curves (ICC)

The analysis of the P-scales data was carried out using the Partial Credit Rasch Model after re-coding the assessment scores into numbers ranging from 1 to 18 as described previously. The Partial Credit Rasch Model was chosen based on the assumption that all pupils who score below 18 on a given subject have given partially correct answers for the subject. So the program was run with ten items formed from each of the P-scales. Each item had 18 marks available to it. Winsteps allows the analyst to see the link between the estimated ability of pupils (theta) and the score on the item (P-level) using the Estimated Item Characteristic Curve.

Figure 1: Estimated Item Characteristic Curve obtained from the P-scales data set.



The Estimated TCC, *Figure 1*, from the P-scales data shows the ability required by a pupil to achieve a given level of the P-scales.

Investigating for the existence of bottlenecks on the various levels of the P-scales

A bottleneck would exist at a given level of the P-scales if for a given subject a substantial number of the pupils examined remain at that level of attainment in that subject for a number of years without any progress or with only little progress over a long period of time. This situation will occur if a large measure of ability is required to progress from the level at which the bottleneck occurs to the next higher level and would be expected if very difficult tasks are required to be accomplished in order to progress to the next level. Therefore if bottlenecks are to be prevented, it is important that tasks in level descriptions (such as the P-scales) are designed in such a way that equal measures of ability are required to progress from one level to the other over the whole range of levels.

Figure 2: A section of the Estimated Test Characteristic Curve obtained for Number

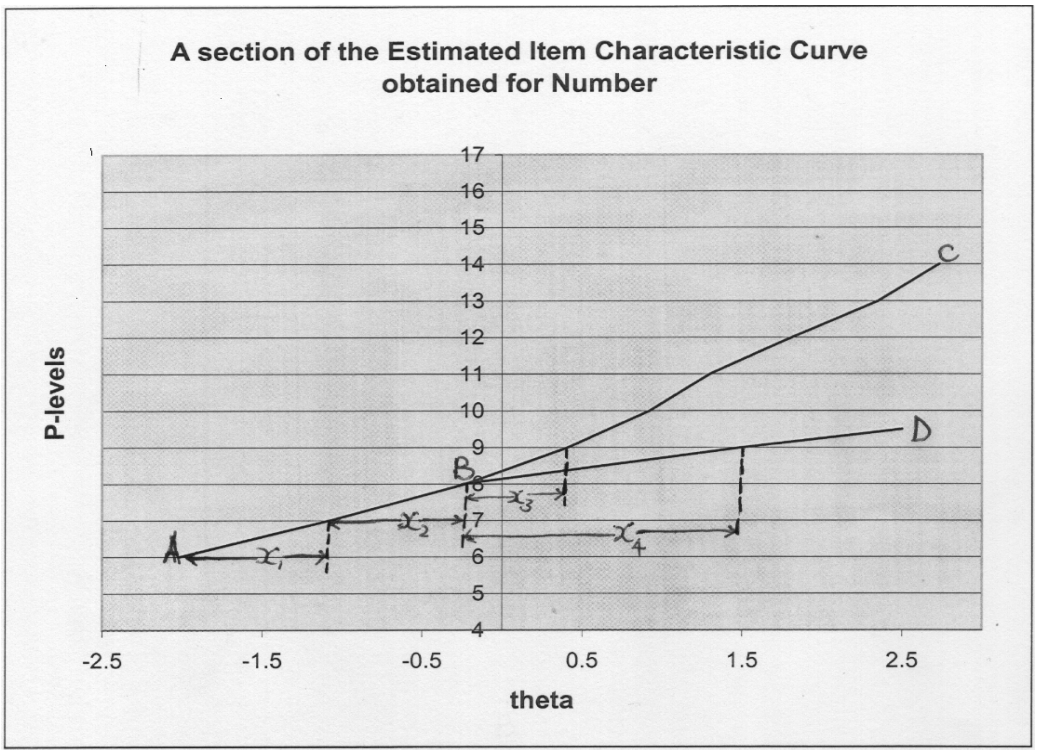
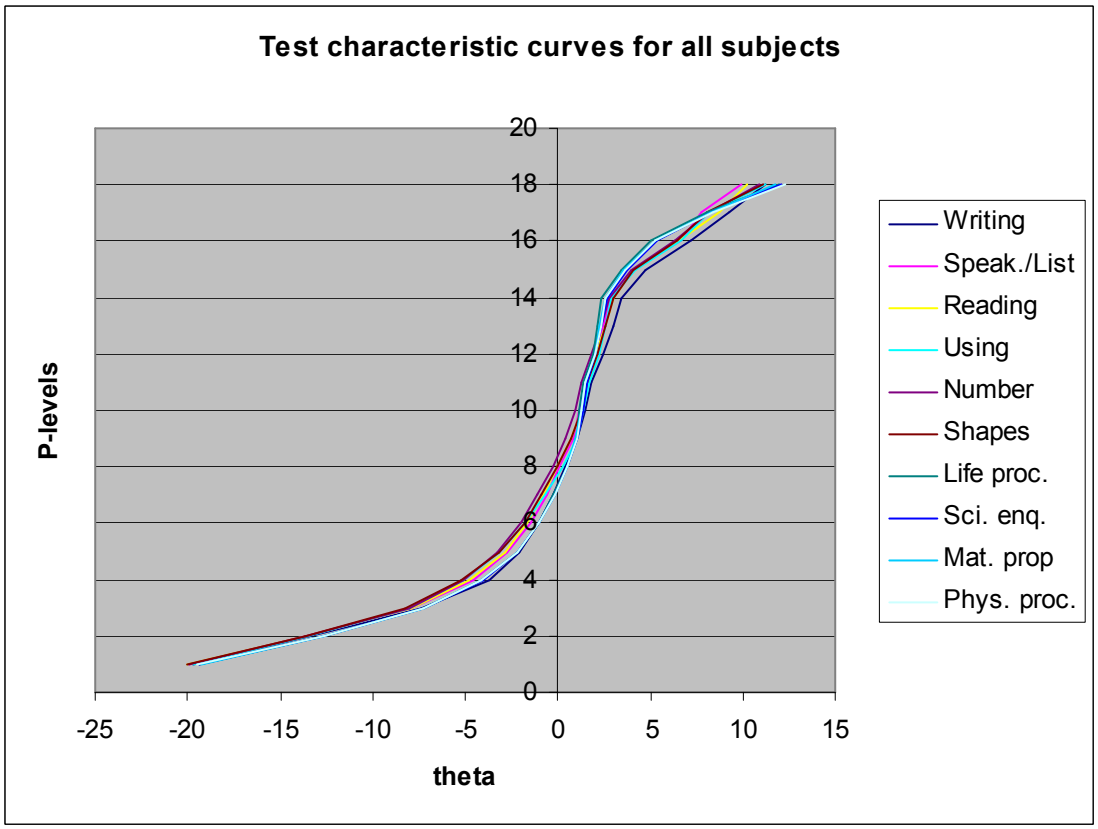


Figure 2 shows a section of the Estimated Item Characteristic Curve (ICC) from the mathematics strand Number. The Estimated ICC shows the abilities required to attain various P-levels. In *Figure 2* it is apparent that section ABC, covering level 6 to level 14 (level 2a of the National Curriculum) of the ICC is nearly a straight line. Therefore the lengths marked x_1, x_2, and x_3 which represent the abilities required to progress from levels 6 to 7, 7 to 8, and 8 to 9 respectively, are equal. Therefore equal

measures of ability are required to progress from one level to the other within that region that includes progress from level 8 to 9 which teachers thought was a problem for pupils. If there was a bottleneck at level 8 we would expect the shape of the graph to be ABD instead of ABC, and in that case the section marked x_4 would give the measure of ability required to progress from 8 to 9. Obviously x_4 is greater than each of x_1, x2, and x_3. Therefore a greater measure of ability would be required to progress from 8 to 9. This would give rise to a bottleneck.

However, the ICCs from the cognitive scales of the P-scales data, shown on *Figure 3*, all have the same shape, and section ABC of the ICC shown in *Figure 2* was taken from one of them, Number. Therefore no bottleneck exists between levels 8 and 9 of any strand of the P-scales.

Figure 3: Test Characteristic Curves for all the subjects described by P-scales



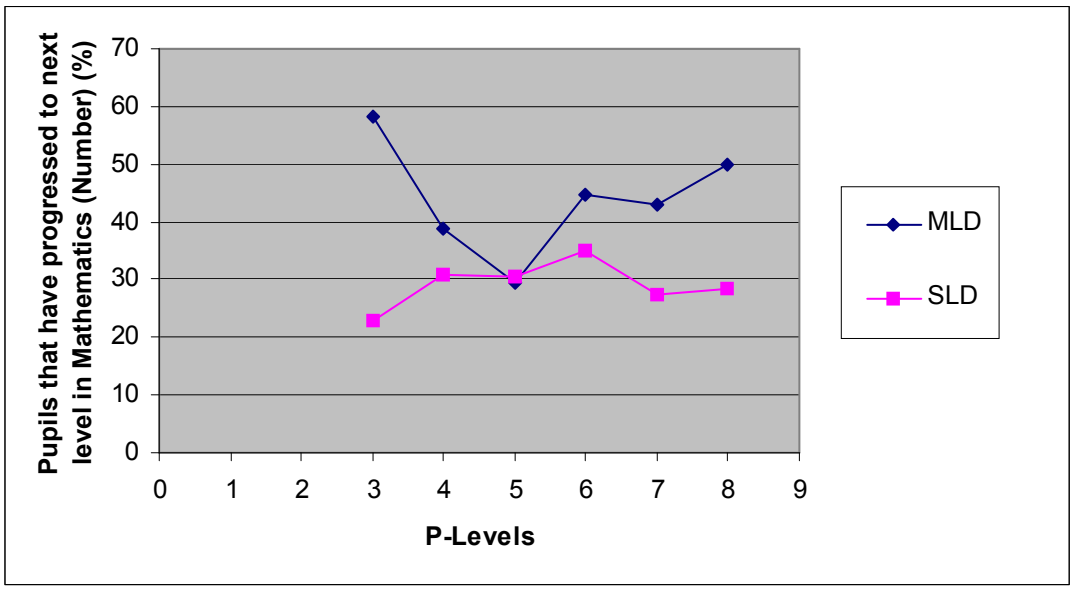
Comparison with classical analysis

Table 1: Distribution of pupils across the three learning difficulty categories

Learning Difficulty	Number of pupils
MLD	9366
SLD	8548
PMLD	2790

A classical approach to the problem of verifying whether a bottleneck exists or not at level P8 of the P-scales was adopted by calculating the percentage of pupils that have progressed from one level of the P-scales to higher levels in various subjects to see whether the values diminished at P8. *Figure 4* shows a graph obtained from Mathematics (Number) but is typical for other subjects of the P-scales. *Figure 4* shows that the percentage of pupils progressing from P8 to the next level of the P-scales (in Mathematics (Number)), did not decrease when compared to the percentages obtained for the preceding levels. This was found to be true for pupils classified as having SLD and MLD. Therefore, the results of the classical analysis support that of the IRT method to confirm that no bottleneck exists at level P8 of the P-scales.

Figure 4: Graph illustrating the results of a classical approach to investigating whether a bottleneck exists at P8 of the P-scales.

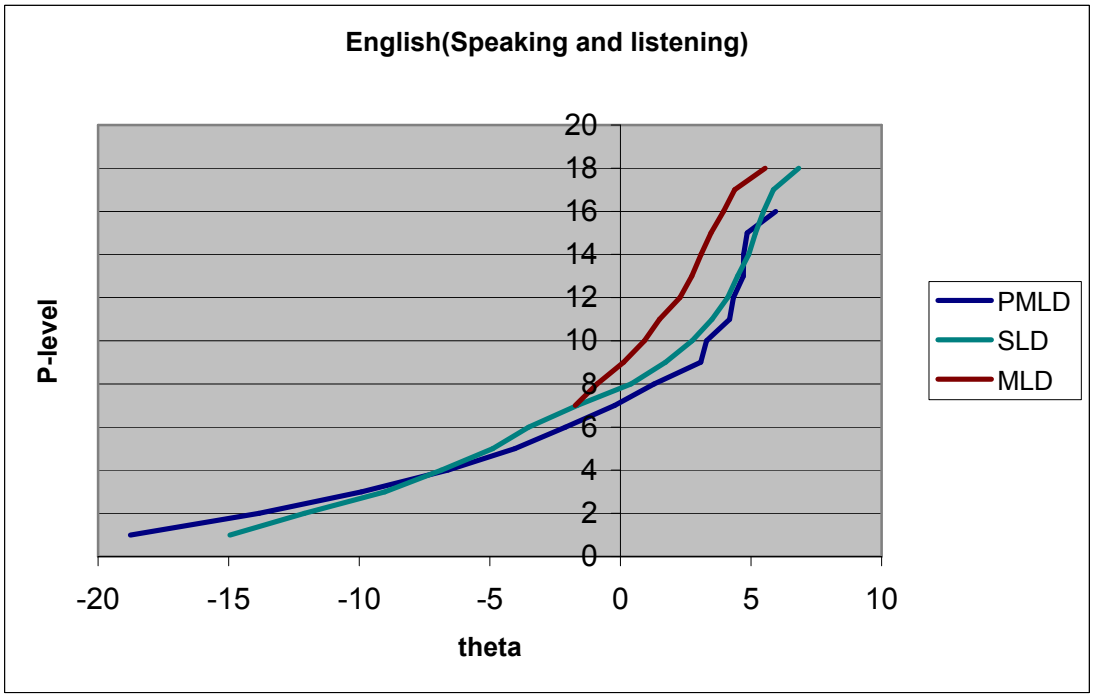


Problem with applying the Partial Credit Rasch Model to the P-scales data

Although the P-scales data has been shown to fit the Rasch Model in a previous section, there appears to be a problem when the abilities of pupils classified as having three different learning difficulties are compared.

In the Estimated Item Characteristic Curves obtained from the P-scales data on two subjects, English (Speaking/Listening) and Mathematics (Number), for the three learning difficulty categories, namely, Profound and Multiple Learning Difficulties (PMLD), Severe Learning Difficulties (SLD), and Moderate Learning Difficulties (MLD), *Figures 5(a)* and *5(b)* respectively, it would be expected that all pupils possessing a certain ability would achieve a given level of the P-scales in any subject irrespective of the learning difficulty under which they have been categorised.

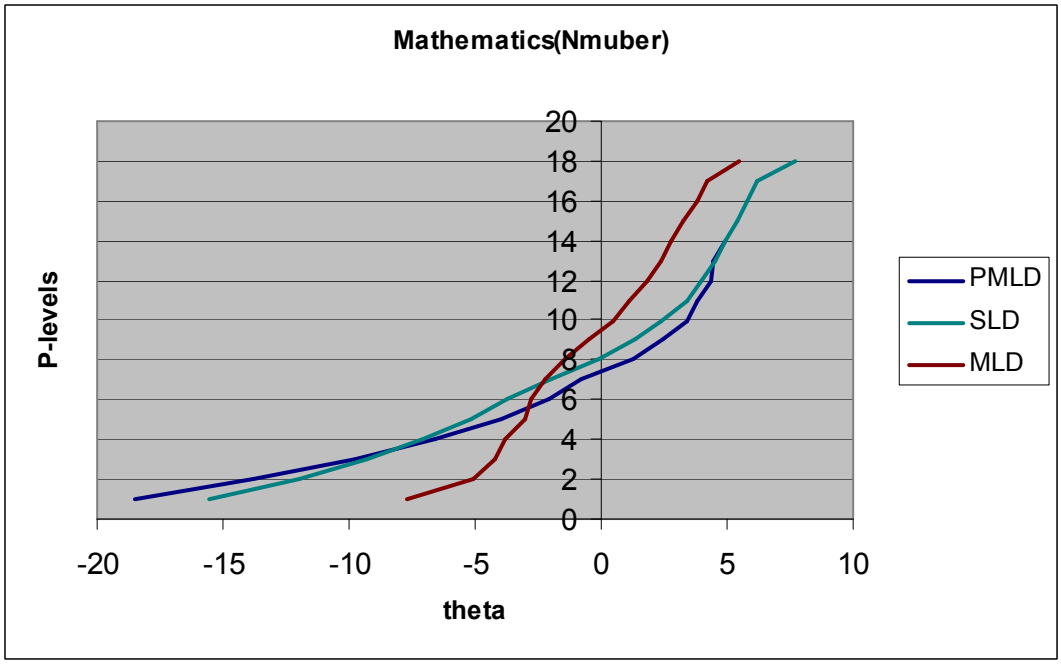
Figure 5(a):Item Characteristic Curves obtained for English (Speaking/listening) for three learning difficulty categories.



This is almost true when comparing attainment data for pupils classified as having PMLD and SLD, because as shown in *Figures 5(a) and 5(b)* the abilities required to

achieve the P-scales levels are not very different for these two groups of pupils. However, the situation is different when MLD pupils are included in the comparison. When the item difficulties for MLD pupils are compared with those of SLD and PMLD the ICCs suggest that the ability a child would possess in order to achieve a given level of the P-scales would depend on which learning difficulty they have been categorised, with PMLD pupils needing higher abilities than SLD and MLD pupils to achieve between P5 and level 2c of the National Curriculum in Speaking and Listening. This observation is unexpected since the P-scales data fits the Rasch model. The reason for the anomalous behaviour of the P-scales data when MLD pupils are included needs further investigation.

Figure 5 (b): Item Characteristic Curves obtained for Mathematics (Number) for three learning difficulty categories.



Conclusion

This study suggests that the Rasch model can be applied to the P-scales. It is also clear from the resultsof this study that there is no bottleneck at P8 or indeed any level of the P-scales. This study has also raised a question about the compatibility of the P-level awards to pupils with different classifications of learning difficulties.

References

1. The collection of performance data for pupils working significantly below age- related expectations. Published by the QCA, 1999.
2. Data Collection and Analysis of P-scales Data. Report submitted to the QCA by CEM Centre, University of Durham, December 2003.
3. Supporting the Target Setting Process (revised March 2001). *Guidance for effective target setting for pupils with special educational needs.* Published by QCA/DfES, March 2001.
4. Setting Targets for pupils with special educational needs. Office of Standards in Education, February 2004.
5. Rasch G. Probabilistic Models for some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.
6. Wilmott A. S and Fowles D.E., The objective interpretation of test performance. *The Rasch Model applied.* NFER Publishing Company Ltd., 1974.
7. Divgi D.R., Does the Rasch Model work for multiple choice items? Not if you look closely. Journal of Educational Measurement. Vol 23, No 4, 1986, pp 283-298.
8. Kwan S.K., The application of measurement theory to tests in mathematics. Ph.D Thesis, University of Durham, 2003.
9. Rasch G., An individualistic approach to item analysis. In P.F. Lazarsfeld and N.W. Henry Eds. Readings in Mathematical Social Science. Chicago:Science Research Associates, 1966, pp 89-107.