**Paper to be presented to the 33rd IAEA Annual Conference,
Baku, Azerbaijan, September 16-21, 2007**

Jeffrey Searle and Neville Hallam

**Curriculum and Evaluation Management Centre (CEM Centre),
Durham University, England.**

Title of Paper

Response Times for Mathematics Items in Computer Adaptive Testing in Secondary Schools

Jeffrey Searle
Jeff Searle is a research associate in the Education Evaluation Group at the CEM Centre. Before joining the CEM Centre, Jeff taught mathematics and computing in secondary schools and in a college of further education. He is particularly interested in mathematics education and its assessment, and is a senior examiner in mathematics.

Neville Hallam
Neville joined the CEM Centre in 2001 as a research associate/computer programmer and has interests in the use of ICT in education, for both classroom teaching and school data analysis

Response Times for Mathematics Items in Computer Adaptive Testing in Secondary Schools

1 Introduction – why assess?

There are many reasons for assessing students' achievement in mathematics. Assessment situations vary on a continuum from "high stakes" public examinations, to "low stakes" class based tests and assessment of individual students. In public examinations the emphasis is on measuring achievement at the end of a course of study, and the results usually have career implications in terms of further study for students, particularly those aged 16 or 18 in England after sitting GCSE and A-level examinations respectively. Other assessment situations, which may be termed "high stakes", are the Key Stage tests in England, in which students are assessed in English, mathematics and science at the ages of 7, 11 and 14. These tests provide a measure of progress of the students and also enable comparison of the aggregated performance between the schools the students attend. Results of assessment might also be used in identifying special needs, in allocation of resources and in research.

In England "high stakes" tests are generally paper based, and the students taking these tests are encouraged to show their working out; that is to write down their thinking that leads to what they believe to be the answer or solution to a mathematics problem. The students thinking, as represented by their working out, is assessed as well as their final answer. It is quite possible for a student to score marks yet get a wrong answer to a question. However, there is no way of judging accurately how long a student spent answering a particular question which is information that might be useful in an overall assessment of a student's achievement.

"Low stakes tests" are usually internal to a school or particular class or student although results might be shared with parents and other interested parties. Such tests might include a class test on a recently studied topic, a half termly test acting as a review of topics studied in the last few weeks or end of year examinations. These tests provide teachers and their students with various levels of measure of progress in developing mathematical skills and understanding of concepts within the subject. Such tests might be given to individual students or groups of students or whole year group depending on their purpose. Teachers may use assessment information to initiate remedial work or plan the next stage of study, parents usually want to know how their son or daughter "is doing", and school management may use such information in managing classes.

Another use to which assessment tests are now commonly used is base line testing; that is where a student's performance can be compared with their performance from earlier years and a measure of their progress can be given in value added terms.

Computer based assessment is appropriate for some of these "low stakes" assessment situations. In recent year computer based assessment has been developing as an assessment tool (Tymms 2001; Gardner *et al* 2002; Ashton *et al* 2003; Lilley and Barker 2003; Russell *et al*, 2003; Tymms *et al* 2004; He and Tymms 2005) Computer based assessment has several advantages over conventional paper and pen based assessment. No printing of test papers is required, and providing there is sufficient IT infrastructure and technical expertise within a school, they are

easy to administer. Feedback in terms of results and their analysis can be supplied relatively quickly. The disadvantage is that generally there is no record of how a student got to his or her answer, but their thinking time, or working out time, can be recorded; that is the time between first viewing a question and submitting a response at the key board.

2. The Computer Adaptive Baseline Test developed at the CEM Centre, Durham University

Response time has been an area investigated by several researchers (Hornke 2000; Bridgeman and Kline 2004; Moshinsky and Rapp 2004; Chang et al 2005), although these have generally focused on tertiary level education. In this paper we investigate the response times of students aged 11 to15 to questions, (usually referred to as items), used in a computer adaptive baseline test (CABT) in mathematics that has been developed at the Curriculum Evaluation and Management Centre at Durham University. The test uses as its basis, algorithms dependent on measures from item response theory and in particular the 1 parameter model known as the Rasch model in which the performance of a student is determined by both his or her ability and the difficulty of the questions he or she is answering. The ability of a student and difficulty of a question are both measured on the same scale in units of logits. (see, for example, Bond and Fox, 2007).

The test is adaptive in that it attempts to assess the level of ability of a student by posing questions to him or her close to the difficulty when he or she will have a 50% chance of attaining a correct answer. As such, it is believed that that the CABT provides a test that can challenge students aged between 10 and 17.

The items used in the CABT have been trialled and calibrated for difficulty using students from UK schools. The items attempt to be curriculum free in that they are not written for any particular course of study but include questions on numeracay, algebra, shape and space and handling data.

The 2006 version of the CABT item bank comprises 515 items ranging in difficulty from –6.25 logits to 5.35 logits and comprises 274 multiple-choice items and 241 free response items.

The response times investigated in this paper have been obtained from 81444 students in years 7, 8 ,9 and 10 in English secondary schools, so are aged between 11 and 15 years. The students took the tests under standardised conditions. The questions the students answer are selected by the algorithm, which selects items at the estimated level of difficulty from the item bank, so that as students progress through the tests the software chooses items that are commensurate with their ability as determined by previous responses and converges on the student's ability. In the CABT convergence to students ability is generally achieved after he or she has answered 15 or so questions.

Table 1 shows the number of students in the data set and the number of responses recorded. Typically a test would involve approximately 20 questions for each student and would take about 20 minutes.

Table 1 Number of students and number of responses

Students					Responses		
	all	boys	girls		all	boys	girls
year 7	37858	17969	19889		734011	350953	383058
year 8	4475	2244	2231		86239	43378	42861
year 9	8018	4356	3662		151665	82559	69106
year 10	31093	15096	15997		587492	285870	301622
totals	81444	39665	41779		1559407	762760	796647

All questions in the CABT were attempted by some of the students in each of the year groups. The two easiest questions (-6.2, -5.8 logits) were the least answered questions with 301 and 282 responses respectively.

The full ability range was found in each year group as shown in Table 2

Table 2 The ability range of the students in logits

	minimum	maximum	Mean	std deviation
year 7	-8.11	6.18	-0.79	1.39
year 8	-8.08	6.15	-0.41	1.54
year 9	-8.05	6.98	0.44	1.68
year 10	-8.02	7.02	0.28	1.68

For ease of handling the data the ability of the students and difficulty of the questions were categorised. Thus for ability, category 1 is students of ability from –8.5 to –8.0 logits and category 31 is students of ability of 6.5 to 7.0 logits.

Figure 1 shows the distribution of ability over each of the year groups.

Figure 1 Year 7 Year 8

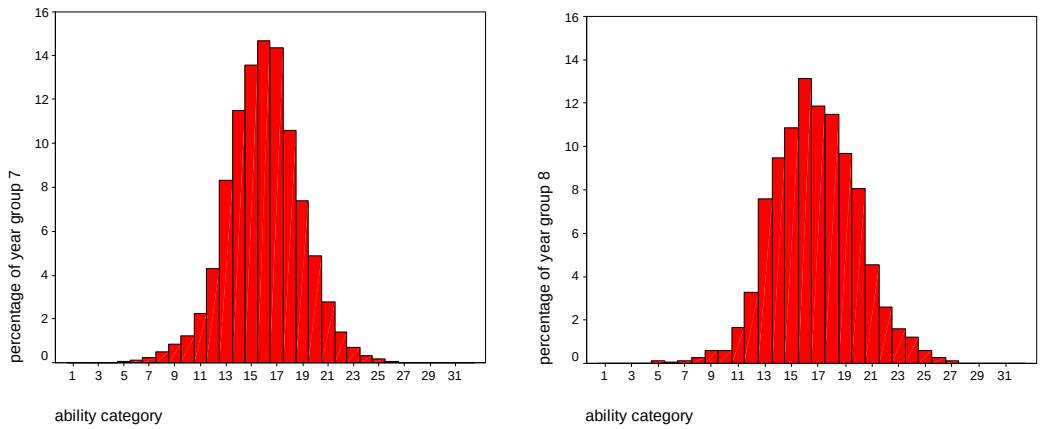
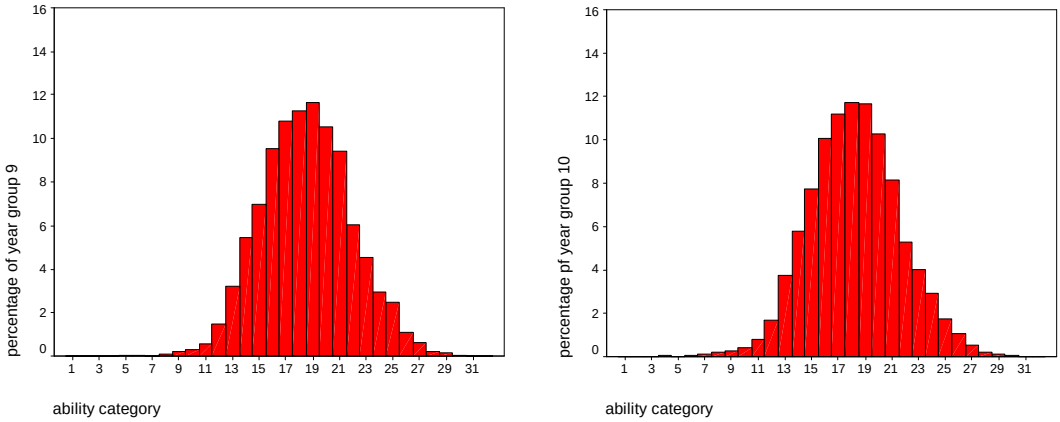


Figure 1 Year 9 Year 10



It is notable from Figure 1 that there is a small increase in the general ability of the students as they get older.

Figure 2 shows the distribution of the questions answered by percentage of students in each year group. Category 1 is questions of difficulty from –6.5 to –6.0 logits and category 24 is questions of difficulty from 5.5 to 6.0 logits.

Figure 2 Year 7 Year 8

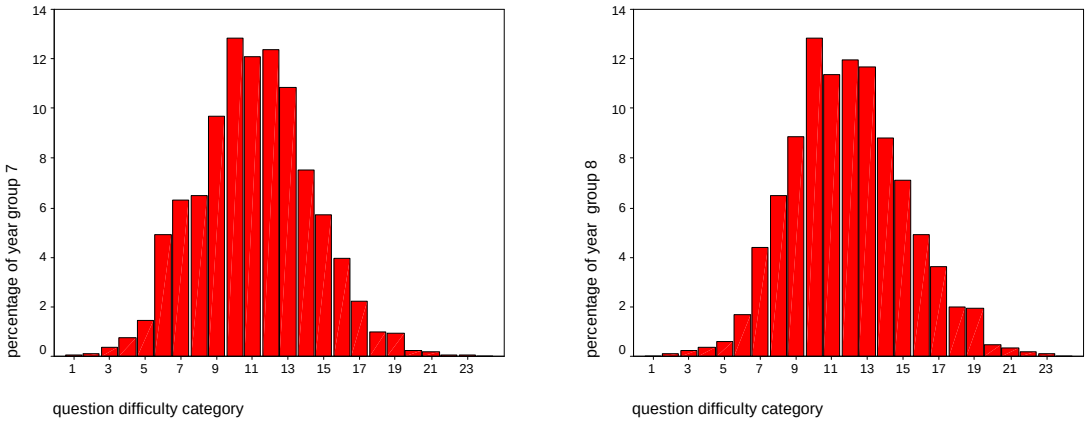
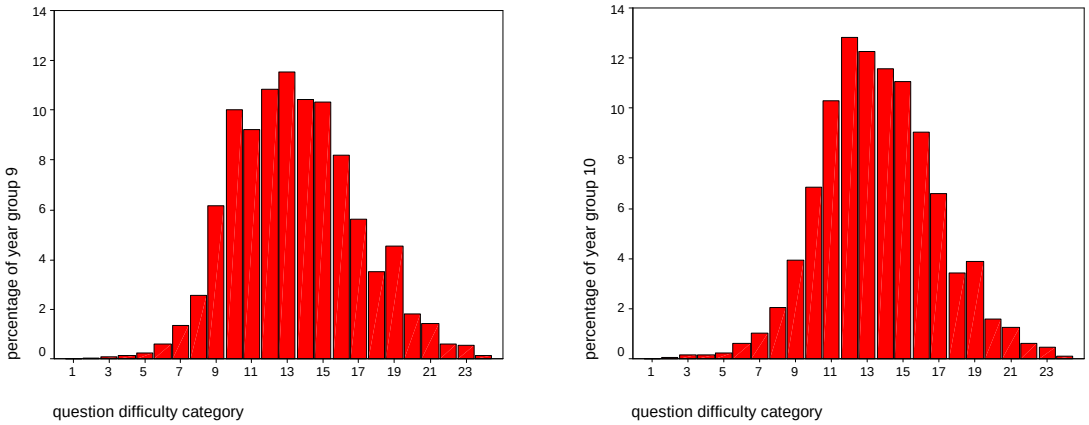


Figure 2 Year 9 Year 10



It is notable from Figure 2 that there is a general small shift towards the more difficult questions as the students get older.

3. Analysis of Response Times

We consider the mean response time and how this varies with the gender and age of the students and whether questions were answered correctly or incorrectly.

We look first at the overall mean response times for the categories of gender, age and whether a student has an answer correct or incorrect. This is shown in Table 3. The numbers in the brackets are the corresponding standard deviations.

Table 3 Overall Mean Times

	responses	All	girls	boys		right	wrong
year 7	734011	33.65	34.89	32.30		31.27	36.72
		(29.2)	(30.0)	(28.3)		(27.1)	(31.4)
year 8	86239	33.52	34.54	32.51		31.51	36.08
		(29.3)	(29.6)	(29.0)		(27.3)	(31.4)
year 9	151665	37.14	37.73	36.64		35.11	39.89
		(32.0)	(32.1)	(31.9)		(30.0)	(34.3)
year 10	587492	37.41	38.04	36.75		36.48	38.49
		(31.8)	(31.9)	(31.8)		(30.4)	(33.4)
all students	1559407	35.40	36.31	34.45		33.57	37.67
		(30.6)	(30.9)	(30.1)		(32.5)	(28.8)

Table 3 shows the overall mean values for these categories. These mean values suggest that on average older students spend a little longer, 3 or 4 seconds, in responding to items. However, Figure 2 indicates that the proportion of students in years 9 and 10 who attempt the more difficult questions is higher than for years 7 and 8; the older students may be taking more time to answer the more difficult questions.

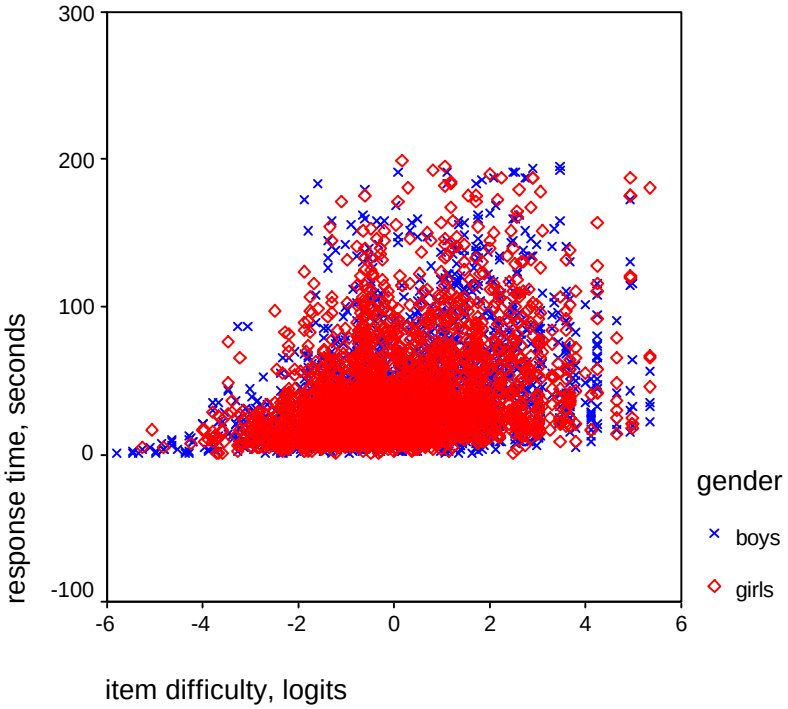
It is also seen in Table 3 that students invest a few seconds more on average in getting an item wrong. These mean values also indicate that girls take longer to respond than boys, but only marginally so.

All differences in the mean values are significant (ANOVA) but for samples of this size that is to be expected. What is also notable is the standard deviation is of the same order of size of the mean values themselves, indicating considerable spread in the data. This is illustrated in Figure 3.

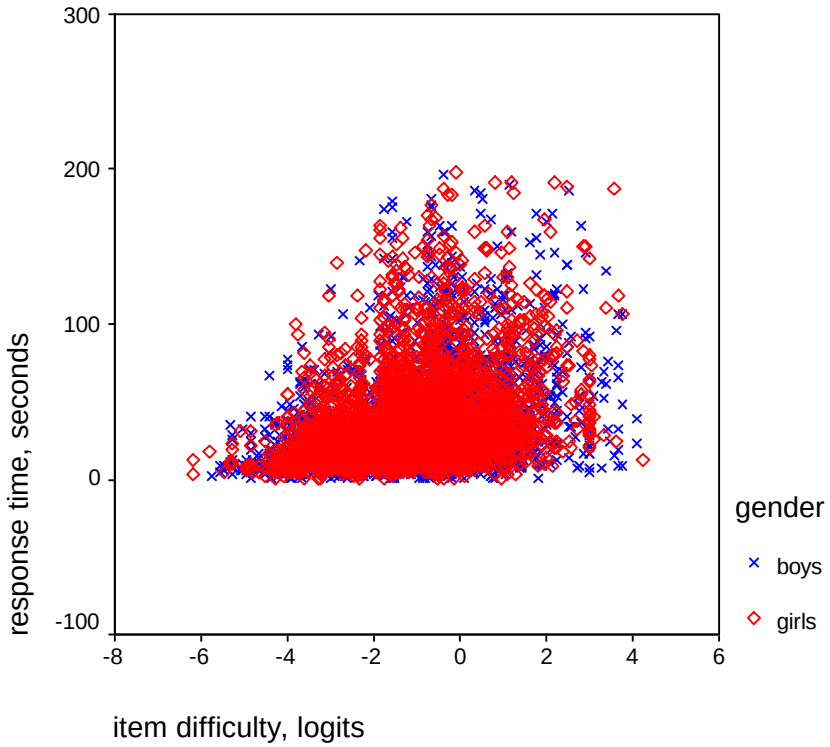
Figure 3, shows the scatter in the data for a 1 percent sample of students from year 7 and from year 10.

Figure 3

Year 7



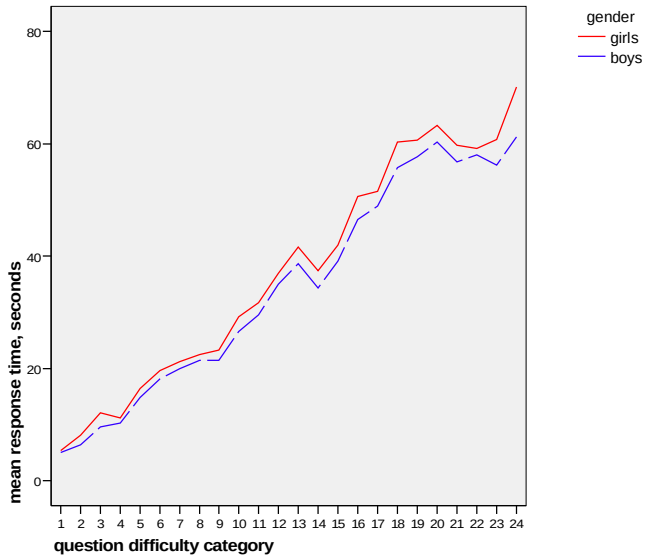
Year 10



As expected there is wide scatter in the data but with the bulk of the data for both year groups between question difficulties of 3 and –3 logits, and response times of less than 100 seconds. There is weak correlation between the question difficulty and the time taken to respond (r=0.3).

We now consider variation in the mean response time with the question difficulty categories. Figure 4 shows the variation in mean response time with the question difficulty categories for boys and girls.

Figure 4 Variation in mean response time with gender.



The trend in the graphs follow what we would expect, in that as the questions get more difficult, the students on average took longer to answer them. It is notable that girls consistently took a few seconds longer on average to respond than boys and this difference tended to increase with difficulty of the questions. The mean difference is 2.7 seconds. .

Figure 5 shows the variation in mean response time with the question difficulty categories for years 7, 8, 9 and 10.

Figure 5 Variation in mean response time with age

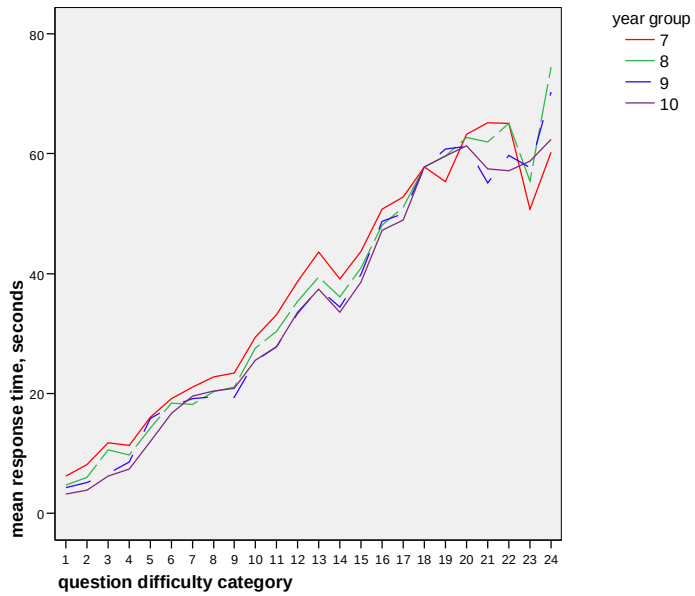
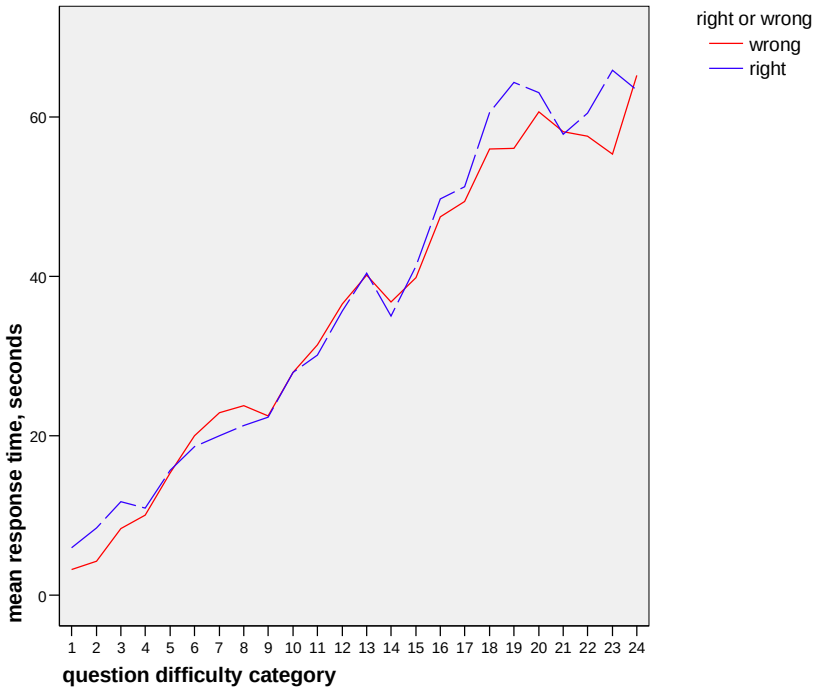


Figure 5 indicates that the students are taking a little longer on average to respond the younger they are, except for the most difficult questions. The proportion of students from a year group who answer the more difficult questions increases with age, as may be seen by reference to Figure 2.

It is interesting to investigate whether these trends are different if students get an item correct or incorrect. During the CABT the students receive no feedback as to whether they are right or wrong in their response. They are simply presented with the next question.

Figure 6 shows the variation in mean response time with question difficulty comparing right and wrong answers.

Figure 6 Variation in mean response time with correct or incorrect response



On average, we see whether a student gets a question right wrong makes very little difference to the mean response time.

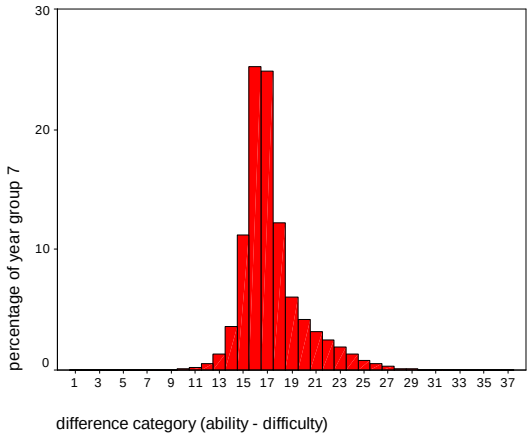
4. Looking at the data from an alternative perspective.

The analysis of data from the CABT has as an objective the selection of items suitable to construct tests for particular groups of students. The analysis above in Section 3, although it indicates some trends also indicate considerable variation. In a computer based test we need to feel assured that the students are engaging with the questions and trying to answer them as best they can, and are not just making guesses to move the software onto the next question or finding the items to be very easy.

The Rasch model allows ability and difficulty to be measured on the same scale so we now introduce a third categorisation. This is the difference between the ability of a student and the difficulty of the question he or she is answering. There are 37 categories in the data; category 1 is a difference of –8.0 to –7.5 logits; category 37 is a difference of 10.0 to 10.5 logits. However, there is very little data outside of the categories 8 and 32, and this is illustrated in Figure 7.

Figure 7 shows the distribution of these categories across the year groups.

Figure 7 Year 7



Year 8

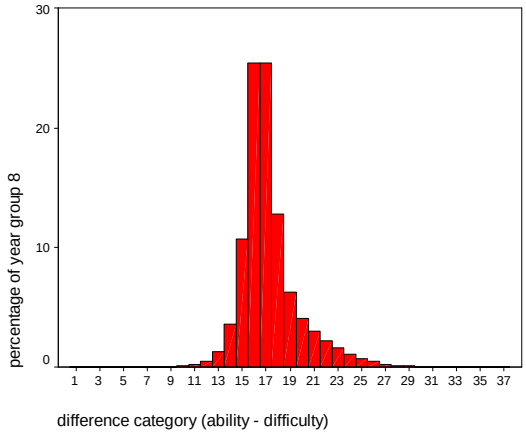
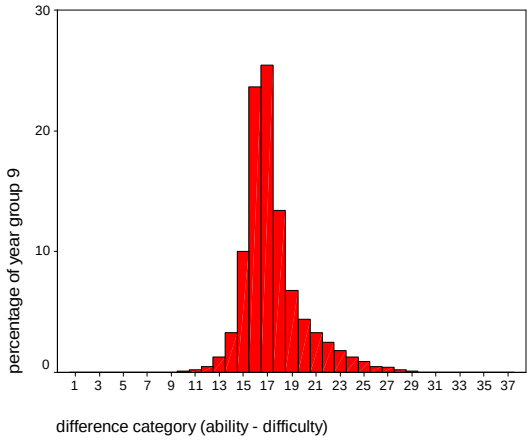
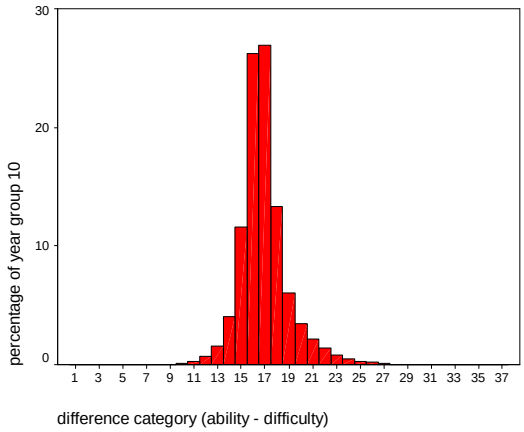


Figure 7 Year 9



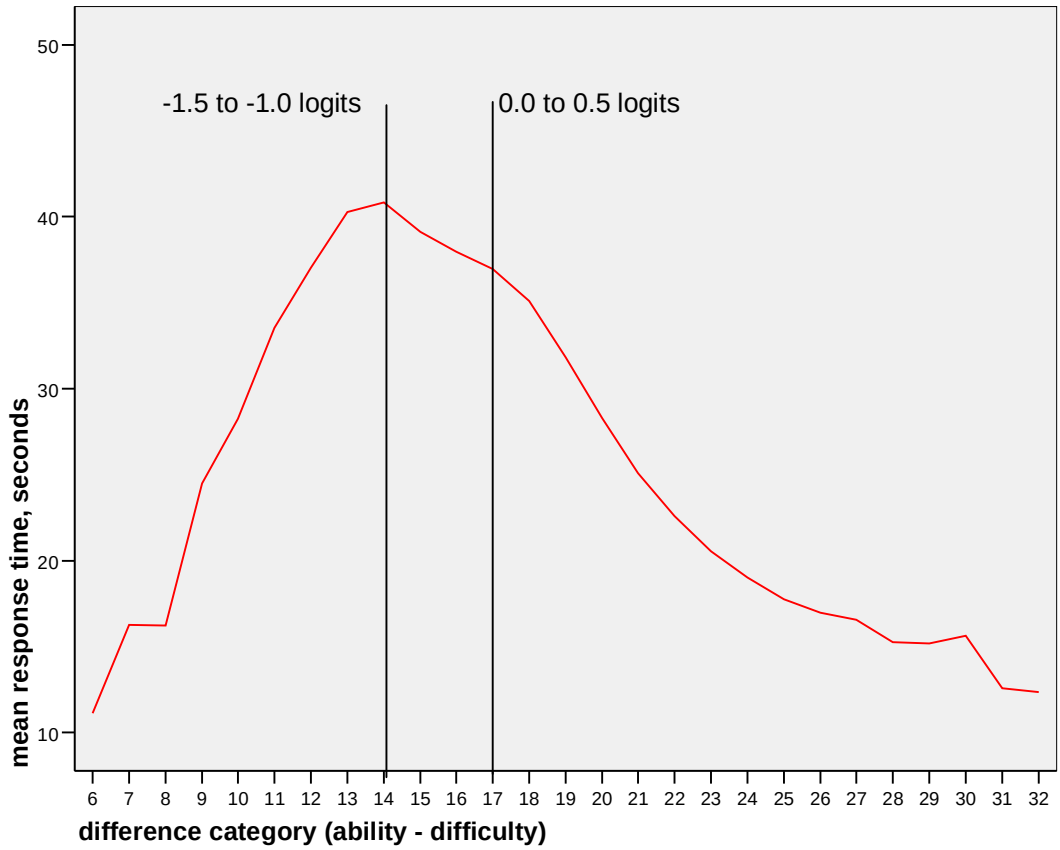
Year 10



It is notable that these distributions are very similar and centre around category 17, which is 0.0 to 0.5 logits. However, it is seen that at the extremes of the distribution where ability greatly exceeds difficulty or vice-versa, the amount of data in each category is small. This resulted in some anomalous behaviour at the extremes of the distribution when this data was further analysed graphically. In the graph shown in Figure 8 the extreme categories have been omitted.

Figure 8 shows the variation in mean response time with the difference category.

Figure 8: Variation in mean response time with (ability – difficulty); all students.

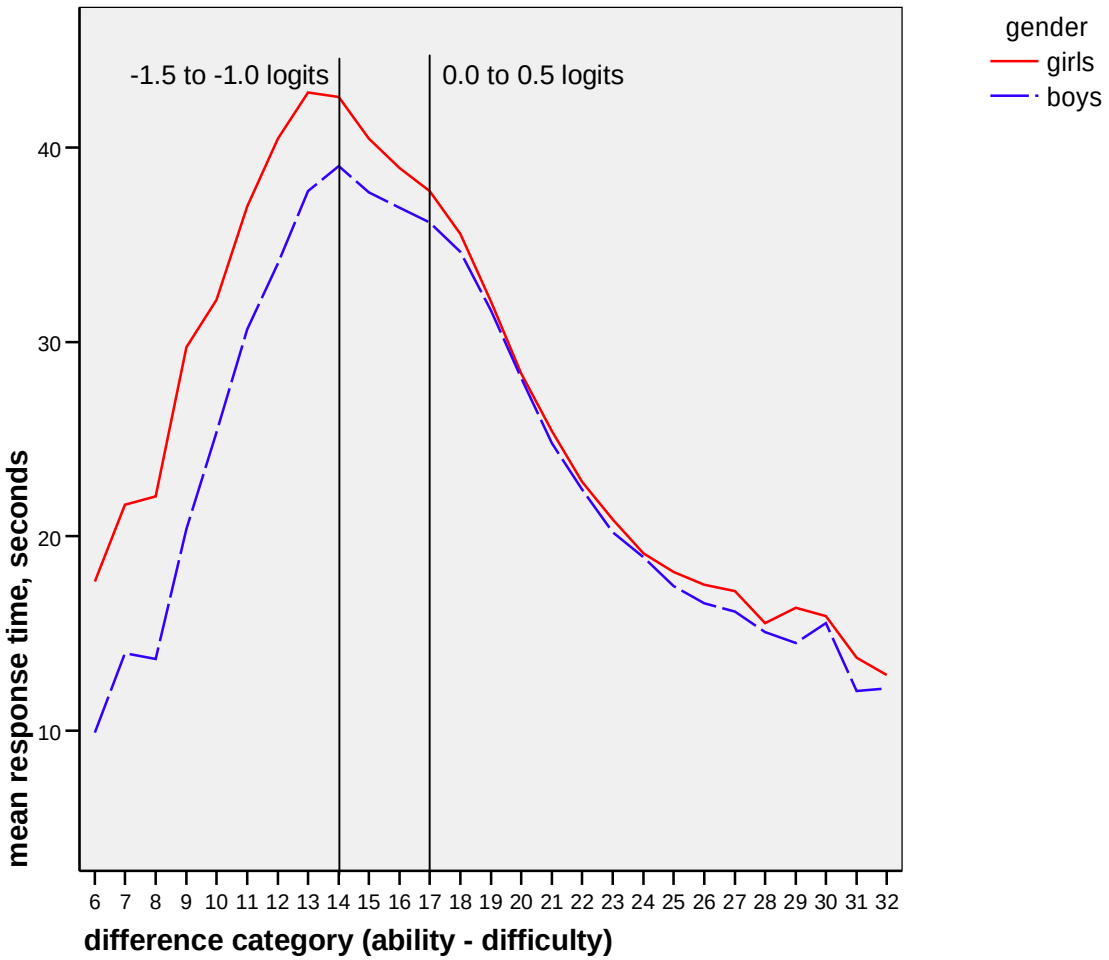


It is immediately notable that the data here is following a trend of a curve that peaks around difference category 14; that is –1.5 to –1.0 logits and a mean response time of about 40 seconds. Other features of this curve are notable; it is steeper to the left of the peak than to the right. To the right of the peak, ability exceeds difficulty after category 17 (0.0 to 0.5 logits) and the student responses are being made more rapidly the greater the difference. As ability exceeds difficulty we would expect most of these responses to be correct. To the left of the peak, ability becomes increasingly less than the difficulty of the questions and we can surmise that the shorter time a student is engaging with a question that they might be guessing the answer rather than trying to work it out or not submitting an answer. The cut off point at which students stop guessing and engage meaningfully with a question is uncertain, but we note students in general are engaging with questions for 20 seconds or more between categories 8 and 24 (or a difference of about –4.5 to 3.5 logits), and for 30 seconds or more between categories 10 and 20. (or a difference of about -3.5 to 1.5 logits).

We now investigate further how the response time and the difference between ability and difficulty varies with gender and age of the students and also with whether they got a question right or wrong.

Figure 9 shows this variation for boys and girls.

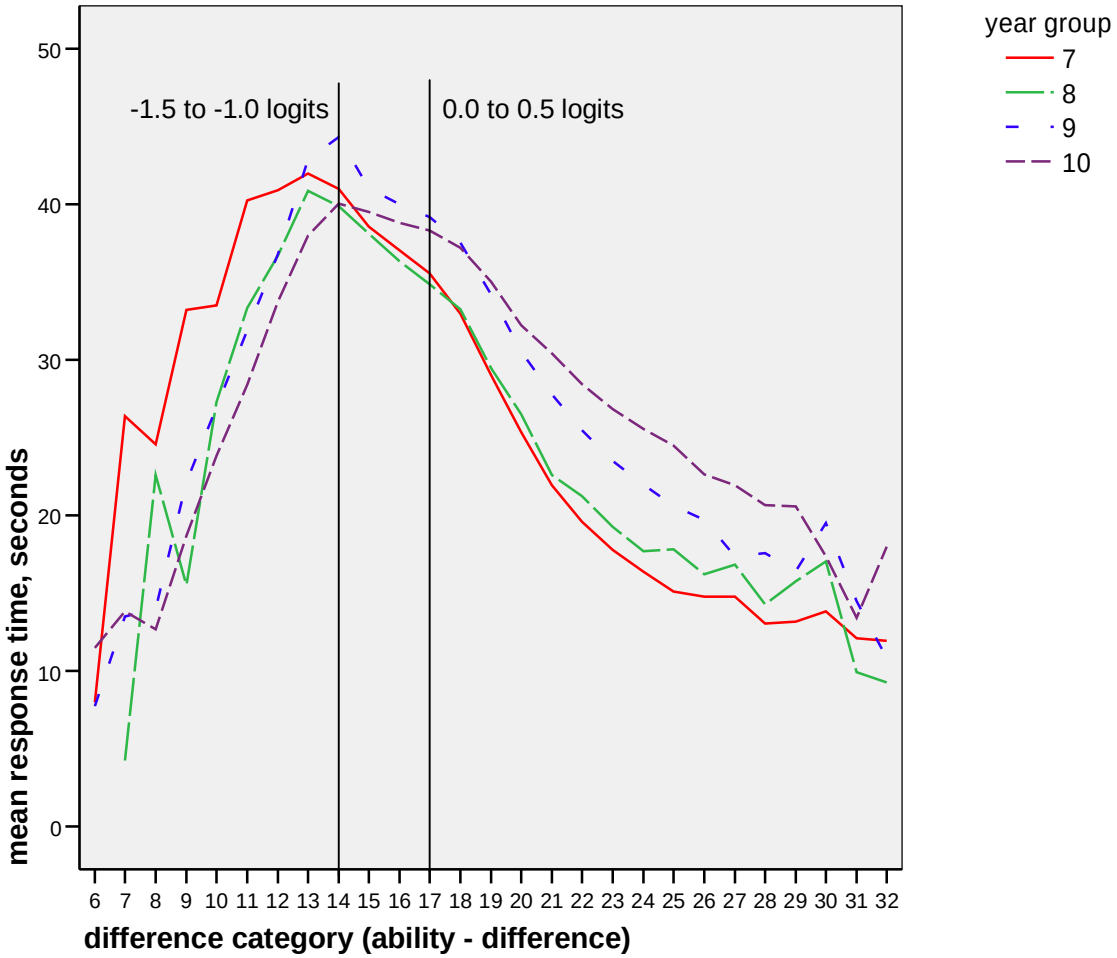
Figure 9: Variation in mean response time with (ability – difficulty) boys and girls



The notable difference between the response time of boys and girls is to the left of the peak, which whilst still at about category 14, is a little more to the right for the boys compared to the girls by about half a logit. To the left of the peak where difficulty exceeds ability, we see in general girls taking longer to respond by about 6 to 7 seconds. To the right of the peak this difference diminishes to zero around difference category 19, or a difference of 1.0 to 1.5 logits. This suggests girls are more prepared than boys to engage with more difficult questions but as questions become easier less and less thinking time is required by both sexes.

Figure 10 shows the variation for the different year groups.

Figure 10 : Variation in mean response time with (ability – difficulty) age of students

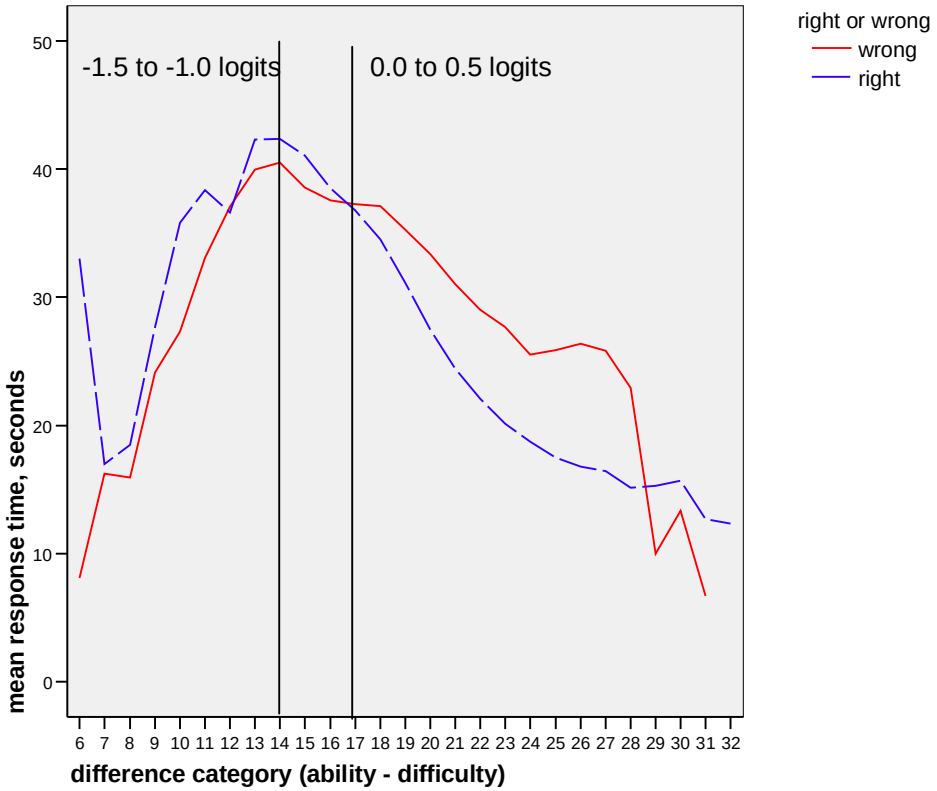


The trend is the same for all four year groups but the peak for years 7 and 8 has moved by about one difference category to the left. This suggests younger students will engage with a question for the greatest time if the difference between ability and difficulty is about –2.0 to –1.5 logits compared to –1.5 to –1.0 logits for older students, or a greater difference of about half a logit.

It is notable that either side of the peaks the graphs for year 10 and year 7 swap over their relative positions. A possible explanation is that younger students are prepared to spend more time than older ones engaging with questions where the difficulty is greater than their ability and vice versa when ability exceeds difficulty; older students given questions they should be finding relatively easy spend a little more time considering them than younger students. Alternatively it may be that older students tend to get more difficult questions than the younger ones, so they may give up more easily (left of the peak) and spend a longer time working on questions (right of the peak).

Figure 11 shows the variation for questions answered correctly or incorrectly.

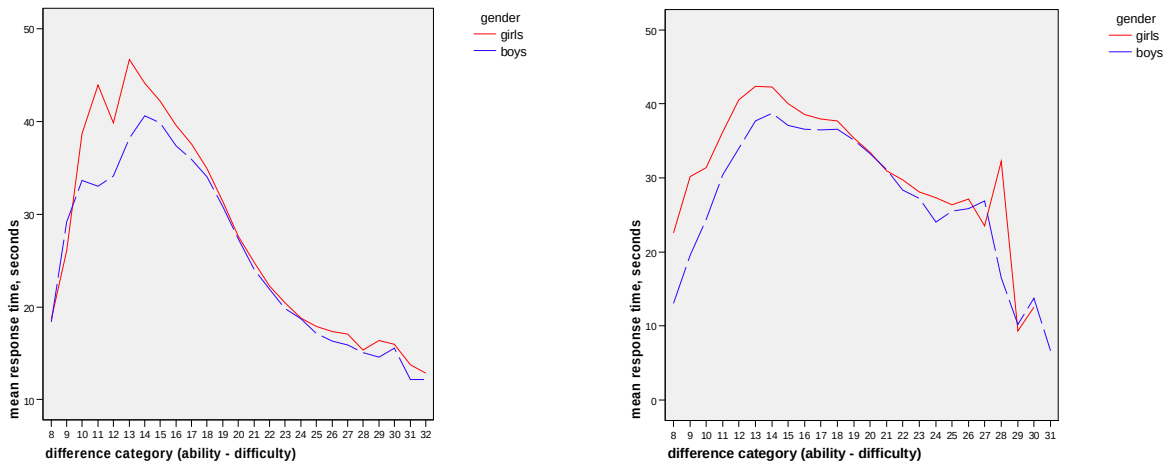
Figure 11: Variation in mean response time with (ability – difficulty)
with correct or incorrect response



Here again we see the relative position of the two graphs change either side of the peak. There is some anomalous behaviour at the extremes of the difference categories but the interesting observation is towards the right where it appears that students invest more time in getting a question they should be finding easy, wrong.

Figure 12 shows the right or wrong responses graphs broken down by gender

Figure 12

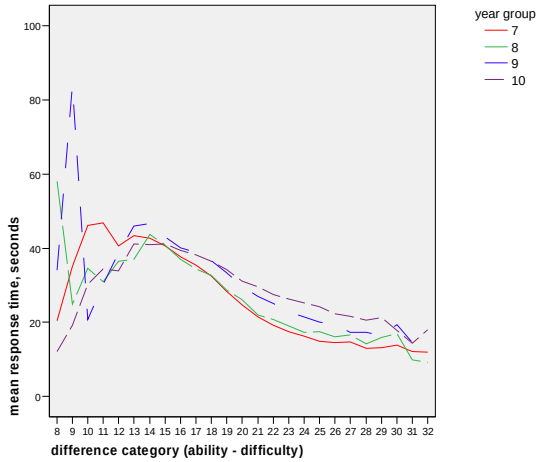


right responses wrong responses

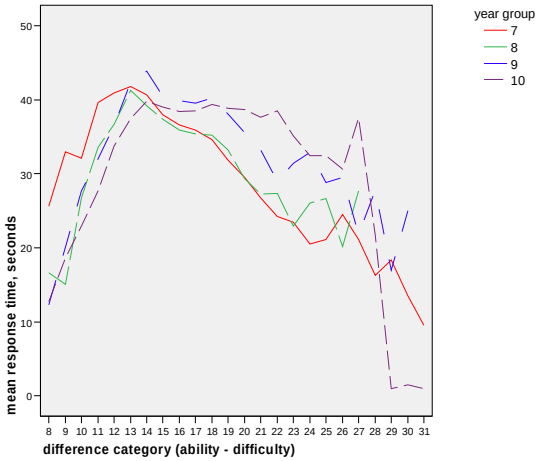
There is some anomalous behaviour at the extremes of the difference categories, and Figure 12 is shown between categories 8 and 32 as a result, but these graphs are similar to that of Figure 9. To the left of the peak around category 14, the girls spend more time answering questions that are difficult for them, whereas the difference between the sexes is hardly noticeable as ability increasingly exceeds difficulty.

Figure 13 shows the right and wrong responses graphs broken down by age of student

Figure 13



right responses



wrong responses

Again anomalous behaviour is seen towards the extremes of the difference categories but otherwise these graphs show similar behaviour to those of Figure 10 with the relative positions of the graphs for years 7 and 10 changing place either side of the peak around category 14.

5. Implications

The principal conclusion from this analysis is that there is an optimum difference between a student's ability and the questions he or she is asked to respond to, if we want the students to engage with the questions and give a considered response. If students are given questions that are very easy for them (i.e. their ability is much greater than the difficulty of the questions) or are given questions that are too difficult for them (i.e. their ability is much less than the difficulty of the questions) then it can be argued that the responses are of little use to teachers in terms of information about the progress of their students in mathematics. When constructing a test for students, the conclusion from this analysis is that ideally the difficulty of the questions should exceed the ability of the pupils by about 2 logits for year 10, reducing to about 1.5 logits for year 7 Also if time limits are to be set for such a test it may be noted that girls in general will require a little longer than the boys to answer the same questions. It is interesting to note the difference in response times broken down into right and wrong responses, but students, and possibly their teachers, cannot anticipate in advance of taking the test what they will get right and wrong.

Further analysis
It would be interesting to see to what extent these results are affected by the type of a question a student is presented with. These could be categorised by multiple choice and free response questions, and also the area of mathematics that the question comes from; numeracy; algebra; shape and space and handling data. It would also be interesting to investigate to what the extent the results are replicated in other curriculum areas, particularly some aspects of English, which are included in the reading and vocabulary parts of the CABT.

The present conclusion re mathematics is that students are most interested and engaged with what they are doing when the test is suitable for the ability of the students and that will be when the difficulty of the questions is a little greater than the ability of the students. This might have some relationship to Vygotsky's zone of proximal development, and it would be interesting to pursue this further. This conclusion is appropriate for "low stakes" tests taken at school level; there are wider implications for "high stakes" national tests, where the question can be raised as to whether the tiering system currently in use in England in the Key stage tests is adequate to assess all students, or whether there should be a move towards more customised testing.

References

Ainley J., Banks D. and Fleming M. (2002)
The influence of IT: perspectives from five Australian schools.
Journal of Computer Assisted Learning 18, 395-404

Ashton, H.S., Scholfiled, D.K. and Woodger, S.C.(2003)
Piloting summative web assessment in secondary education.
2003 CAA Conference Proceddings:19-29. Loughborough University, UK.

Bond. G, and Fox, C. (2007)
Applying the Rasch Model; fundamental measurement in the human sciences.
Pub. Laurence Eribaum Associates, Inc. US

Bridgeman, B. and Cline, F. (2004)
Effects of differentially time consuming tests on computer-adaptive test scores
Journal of Educational measurement 41: 137-148

Chang, S., Plake, B. Ferdous, A. (2005)
Response times for correct and incorrect item responses on computerised adaptive tests
Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

He, Q. and Tymms, P.B. (2005)
A computer assisted test design and diagnosis system for use by classroom teachers.
Journal of Computer Assisted Learning 21: 419-429

Hornke, L. (2002)
Item response times in computerised adaptive testing
Psicologica 21175:189

Gardner, L., Sheridan, D. and White, D. (2002)
A web-based learning and assessment system to support flexible education.
Journal of computer assisted Learning 18:125-136

Lilley, M. and Barker, T. (2003)
An evaluation of a computer adaptive test in a UK university context
2003 CAA Conference Proceedings: 171-182 Loughborough University, UK.

Moshinsky, A. and Rapp, J. (2004)
Performance time on an adaptive power test.
Paper presented at the Annual Meeting of the American Educational Research Association, San Diago, USA

Russell, M., Goldberg, A. and O'Connor, K. (2003)
Computer-based testing and validity: a look into back into the future.
Assessment in Education: Principles, Policy and Practice 10:279-293

Ware, M. and Woodger, S. (2004)
The role of technology in assessment-online in Scotland; the Pass-IT project (phase 1)
Paper presented at the 30th IAEA Conference Annual Conference, June 13-18, 2004, Philadelphia, USA.

Tymms, P.B., Merrell, C. and Jones, P. (2004)
Using baseline assessment data to make international comparisons.
British Educational Research Journal 30, 673-689.