AN INVESTIGATION OF THE RESPONSE TIME FOR MATHS ITEMS IN A COMPUTER ADAPTIVE TEST

Chris Wheadon and Qingping He

An Investigation of the Response Time for Maths Items in a Computer Adaptive Test

Chris Wheadon and Qingping He
CEM Centre
Durham University
UK

Chris.Wheadon@cem.dur.ac.uk
Qingping.He@cem.dur.ac.uk

Abstract

An important advantage of computer based testing over conventional paper and pencil based testing is that the response time to items from test takers can be accurately recorded for subsequent analysis. This study investigates the response time for maths items in a computer adaptive test designed as a baseline assessment for pupils aged from 11 to 18 in the UK. The results showed that the response time for all the items in the test generally increases with item difficulty, although significant variability exists. The item difficulty levels and the age and ability of test takers have significant influence on item response time.

Keywords

Item Response Theory, Computer Adaptive Testing, Item Response Time.

Introduction

Information and computing technology (ICT) has been widely used in education at various levels to assist learning in education organisations, and computer based testing (CBT) is becoming increasingly important as an assessment tool (e.g Tymms, 2001; Gardner *et al*, 2002; Ashton *et al*., 2003; Russell *et al*., 2003; Tymms *et al*., 2004; He and Tymms, 2005). CBT can gather more information than conventional paper-and-pencil testing. For example, it is possible to record the time a person takes to answer a specific item in a computer-based test. Of the computerised testing procedures currently in use, computer adaptive testing (CAT) has attracted particular attention in recent years (see Lilley and Barker, 2003; He and Tymms, 2005). Most computer adaptive testing systems are based on the implementation of an Item Response Theory (IRT) model, which generally assumes that, given a test and examinee sample, the overall performance of an examinee is determined by his/her ability and the characteristics of the test items (see, for example, Hambleton and Swaminathan, 1983; Masters and Keeves, 1999; Tymms, 2001; Wang and Kolen, 2001; Tonidandel *et al.*, 2002; Lilley and Barker 2003; He and Tymms, 2005). In a computer adaptive test, for a particular examinee, the items, drawn from an item bank containing items that have been calibrated using an IRT model (i.e. item statistics such as item difficulty and discrimination power have been estimated using an IRT model), are targeted at his/her ability level, and each individual will therefore answer a different set of items.

The study of item response time is important for understanding the physiological behaviour of test takers during the testing process, which is essential for creating effective items and tests that can provide more accurate educational measurements. A number of researchers have conducted work in this area (e.g. Hornke, 2000; Chiu and Bejar, 2001; Bridgeman and Cline, 2004; Moshinsky and Rapp, 2004; Chang *et al*. 2005). In the study undertaken by Chang *et al*. (2005), the authors found that higher ability students showed persistence with test items irrespective of item difficulty and generally spent more time on items than lower ability students, while work by Moshinsky and Rapp (2004) on a high-stake test used for undergraduate admissions in Israel indicated that: more difficult items generally take more time to answer than easier items and that more able students take less time to answer items incorrectly than less able students.

This paper reports results from an investigation of the response time to the items in an adaptive test based on data collected from over 100,000 students, and attention has been focused on studying the effects of item difficulty, and the age and ability of test takers.

The Computer Adaptive Baseline Test

The Curriculum, Evaluation and Management (CEM) Centre at Durham University has been conducting baseline assessments on primary, secondary and post-sixteen students through the administration of paper-and-pencil based tests and questionnaires via a number of performance indicator related research projects, including the Performance Indicators in Primary Schools (PIPS) project, the Middle Years Information System (MidYIS, for Year 7 students aged from 11-12) project, the Year 11 Information System (Yellis, for Year 10 students aged from 15-16) project and the A Level Information System (Alis for Year 12 students aged from 17-18) project (see Fitz-Gibbon, 1997). The baseline data are then linked to students' subsequent academic performance in order to provide value added information for schools to undertake self-evaluation and management. In view of the relatively good IT facilities available today in schools, a two-part computer adaptive test has been developed as an alternative to the conventional paper-and-pencil baseline tests for the three secondary projects (MidYIS, Yellis and Alis). The adaptive test includes an adaptive maths test and an adaptive English vocabulary test. This computer adaptive baseline testing (CABT) system comprises a calibrated English vocabulary item bank, a calibrated maths item bank, and an item display and recording system for displaying items to students and recording responses. The calibrated item banks, in which the item difficulty varies across a wide range, were established by administrating a series of tests to students of various ages and the embedding of common items in the tests, analysis of test results using the Rasch model (see Rasch, 1960; Wright and Stone, 1979), and the equating of the tests using common items. In total there are over 500 vocabulary items in the vocabulary item bank and over 500 maths items in the maths item bank. Effort has been made to make the items content-independent to each other when creating the maths items. Testing is delivered through the Web or from the school's local network. As the items in the item banks cover a wide difficulty range, all three projects use the same adaptive tests with different starting item difficulty to gather baseline information. This has avoided the need to develop separate tests for individual projects. The present study will focus on the items contained in the adaptive maths test.

Results and Discussions

The Effect of Item Difficulty

Theoretical models and empirical evidence suggest that there is a positive correlation between item difficulty and response time (e.g. Moshinsky and Rapp, 2004). This relationship is corroborated to some extent in the present study as shown by Figure 1, which plots the response time against item difficulty for all the items in the adaptive maths test taken by year 12 students. However, significant variability in the mean response times at all levels of item difficulty exists. The information presented in Figure 1 will be useful for constructing more efficient tests by using less time-consuming items across a range of difficulty levels.

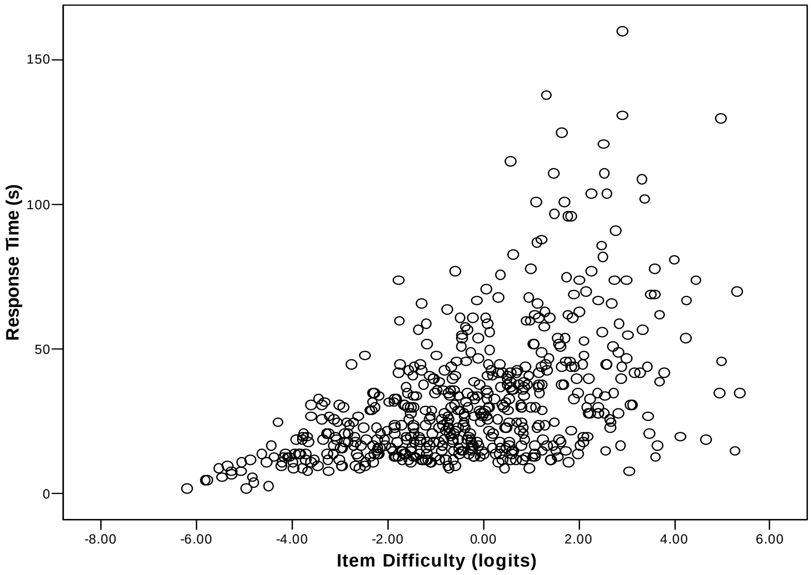


Figure 1 The distribution of response time against item difficulty for Year 12 students

The Effect of Age Groups

As the CABT is undertaken by a large number of students in years 7, 10 and 12, comparisons can be made across age groups. It should be noted, however that the sample size for year 7 decreases as the items become more difficult and the sample size for year 12 increases as the items become more difficult. As an example, a selection of items from the central difficulty range of the maths item bank have been used for comparison, and Figure 2 shows the distribution of response time for different year groups. Figure 2 shows that Year 7 students seem to spend longer than the other year groups on the easier items than on the more difficult items. This may, however, be due to reduced sample size on the more difficult items. It is clear from Figure 2 that the there is generally a positive correlation in the response time between the groups for the selected items, although the response time varies substantially between the items.

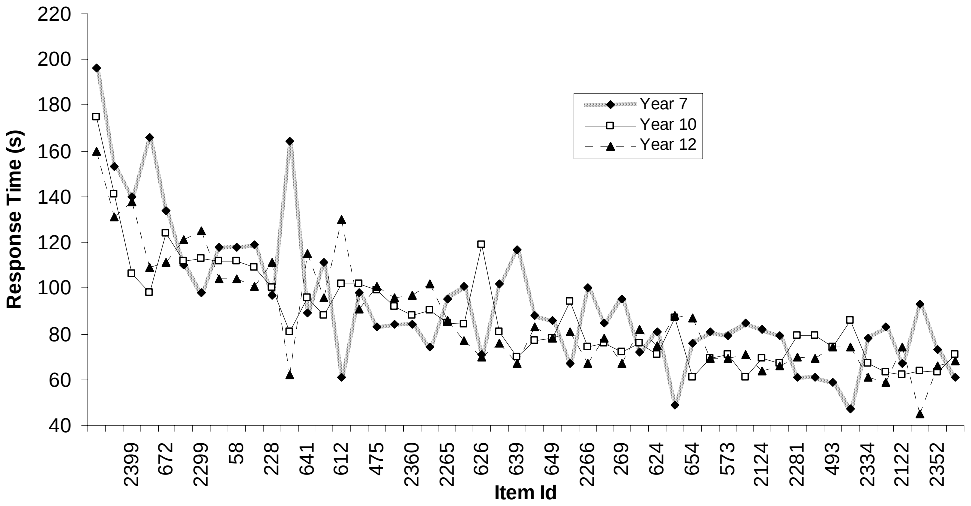


Figure 2 The distribution of response time against year group for a selection of items

Performance Time and Response Accuracy

Figure 3 shows the mean response time by response accuracy across a selection of items. It is clear from Figure 3 that for specific items the mean response time for a correct answer can be greater or less than that for an incorrect answer but there is generally a positive correlation between the two. The overall average response time for correct answers for these items is greater than that for incorrect answers. This is in contrast to the findings from Moshinsky and Rapp (2004). In their study, the authors find that the time reflected in correct responses is less than the time invested in incorrect responses. This contradiction may result from the difference in the nature of the tests being studied: the CABT test is a low-stakes curriculum-free test; the Psychometric Entrance Test is a high-stakes university admissions test. The content domains tested and the age of the test takers may also have contributed to this contradiction.

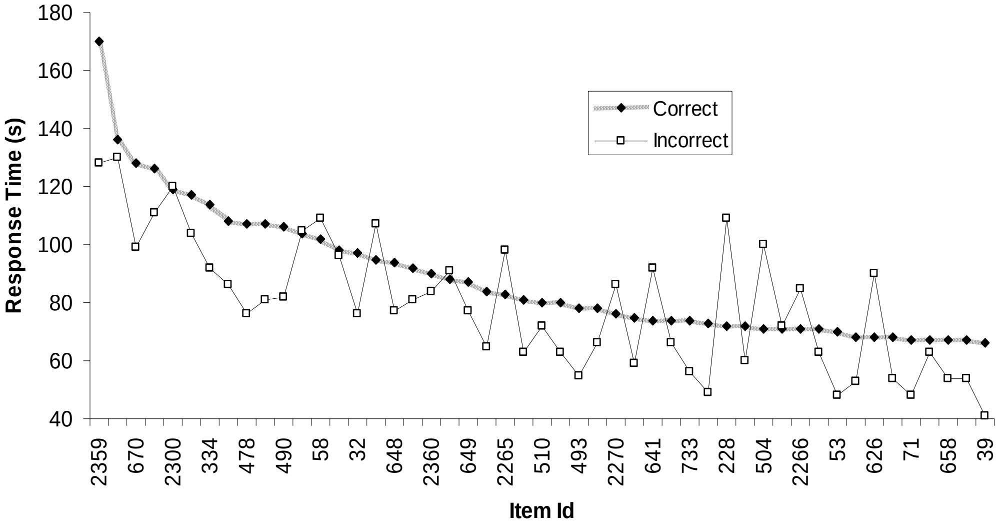


Figure 3 The distribution of the mean response time against response accuracy

The Effect of Ability and Age of Test Takers

Moshinsky and Rapp's (2004) examination of response time found that more able examinees tend to be faster than less able examinees, which is especially true when able examinees know the correct answer. This relationship is diminished when examinees do not know the correct answer, thus the time difference between correct and incorrect answers tends to increase with ability. This was seen to be consistent with the finding that mental ability and mental speed are correlated (see Thissen, 1983). As an adaptive test presents items to candidates that are commensurate with their ability the relationship between time taken on the test and ability is confounded by the difficulty of the items presented to candidates: more able candidates are presented with more difficult items that take longer to solve. Due to the random element of item selection in an adaptive test, and the time it takes for a test to converge on a final estimate of ability, every item is taken by a reasonably wide ability range. The mean ability of 1537 students who took Item 2359 with a difficulty of 2.9 logits, for example, was 2.3 logits, with a minimum of –2.7 logits, a maximum of 8.9 logits and a standard deviation of 1.7 logits. It is therefore possible to analyse the response time of students on individual items which removes any confound with item difficulty. As the CABT is taken by students from the age of 11 to the age of 18 it is furthermore possible to examine the interaction effect of age on performance. As in the study by Moshinsky and Rapp (2004) the correct and incorrect answers are examined separately due to the influence accuracy has on response time.

Three items were chosen for detailed investigation of the effect of age and ability of test takers on response time. These questions, which require a fair amount of time to answer, were chosen from different levels of the difficulty range. The contents of the three items are listed below.

Q.631 Understanding a simple algebraic relationship. Difficulty: -0.6 logits.

The table represents a relationship between x and y. What is the missing number in the table?

X	Y
2	5
3	7
4	?
7	15

a) 9 b) 10 c) 11 d) 12 e) 13

Q.490 Comparing two fractions. Difficulty: 1.1 logits.

Compare the two expressions:

Expression A: $\frac{14+15+16+17+18}{5}$

Expression B: $\frac{17+18+19+20}{4}$

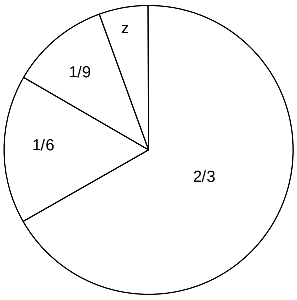
Expression A is greater than expression B

Expression B is greater than expression A

The expressions are equal

Q.2359 Reading a pie chart, working with fractions. Difficulty: 2.9 logits.

The pie chart represents the different colours of cars in Albert Street. If there are 144 cars in total, how many are blue (segment z)?



Free response answer.

Figures 4 to 6 show the relationship between ability and scaled response time (defined as the natural logarithm of the actual response time in seconds) for each year group for the selected items. Care must be taken in interpreting these graphs, however, for a number of reasons. Response time, as noted by Moshinsky and Rapp (2004) tends to be positively skewed. Natural logarithm and cube root transformations make the distribution more symmetrical, but generally the data is unsuitable for parametric tests. Splitting the Year Groups by ability band furthermore results in uneven sample sizes and heterogeneous variance.

Item 631: An easy item

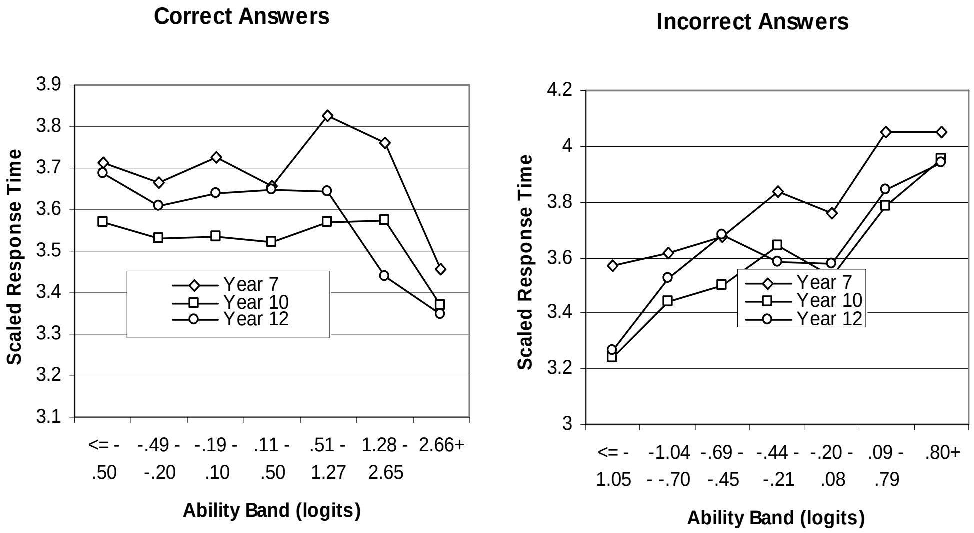


Figure 4 Distribution of scaled response time by ability for correct and incorrect answers on item 631

Figure 4 shows the relationship between response time and ability of test takers (banded) for Q631, which is a relatively easy item. Figure 4 replicates Moshinsky and Rapp's (2004) finding that response time is negatively correlated with ability when the item is answered correctly (r = -.10 p<.001). Moshinsky and Rapp's (2004) finding that incorrect answers are not correlated with ability, however, are contradicted by the positive correlation (r = 0.17 p<.001) between ability and response time when the item is answered incorrectly from our study.

For correct answers, response time was significantly affected by Year Group (H(2)=69.1, p<.001) with Jonckheere's test revealing a significant linear trend in the data, J = 1852769, z = -7.26, r = -.012. Students in year 7 generally took longer to answer this item than students from other year groups. Post hoc Mann-Whitney tests of the difference between the three year groups for correct answers revealed a significant difference between years 7 and 10 (U=608416, r= -.15) and between years 7 and 12 (U=457640, r=-.16), but not between 10 and 12. (the critical value for significance was set at .0167 after application of the Bonferroni correction).

Item 490: A item of medium difficulty

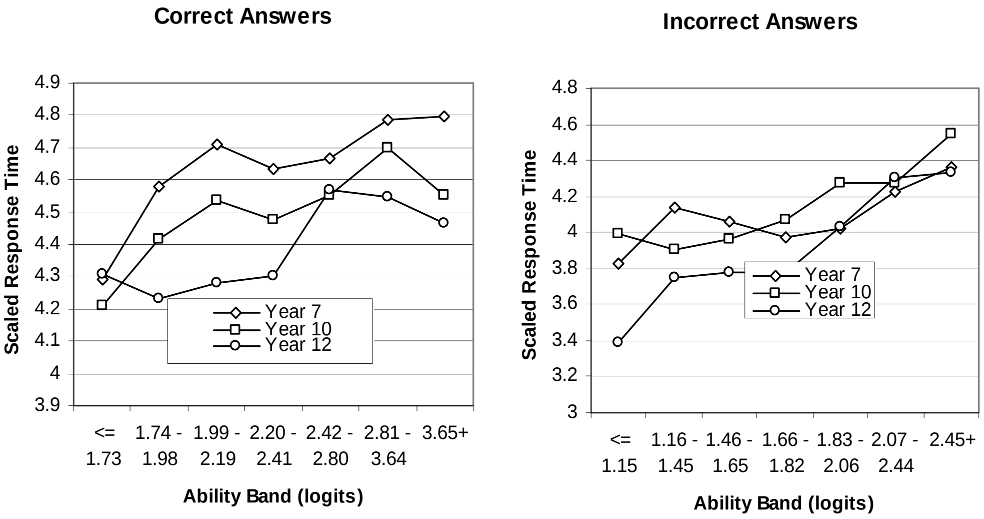


Figure 5 Distribution of scaled response time by ability for correct and incorrect answers on item 490

Figure 5 shows the relationship between response time and ability of test takers (banded) for Q490, which is a medium difficulty item. Figure 5 shows no correlation between time taken and ability for correct answers; contradicting Moshinsky and Rapp's finding that response time is negatively correlated with ability when the item is answered correctly. Moshinsky and Rapp's finding that incorrect answers are not correlated with ability are also contradicted by the positive correlation (r = 0.14 p<.001) between ability and response time when the item is answered incorrectly.

Once again, for correct answers, response time was significantly affected by Year Group (H(2)=18.2, p<.001) and Jonckheere's test revealed a significant trend in the data: as year group increases, the median response time decreases, J=736954, z=-4.3, r=.09. Post hoc Mann-Whitney tests of the difference for correct answers between the three year groups revealed no significant difference between years 7 and 10, but a significant difference between Years 7 and 12 (U=93236, r= -.09) and between years 10 and 12 (U=547821, r= -.07) with the critical value for significance set at .0167 after application of the Bonferroni correction.

Item 2359: An item of high difficulty

Figure 6 shows the relationship between response time and ability of test takers (banded) for Q2359, which is the most difficult item of the three. Figure 6 replicates Moshinsky and Rapp's (2004) finding that response time is negatively correlated with ability when the item is answered correctly (r -.26 p<.001). Moshinsky and Rapp's finding that incorrect answers are not correlated with ability is contradicted by the positive correlation (r = 0.1

p<.001) between ability and response time when the item is answered incorrectly.

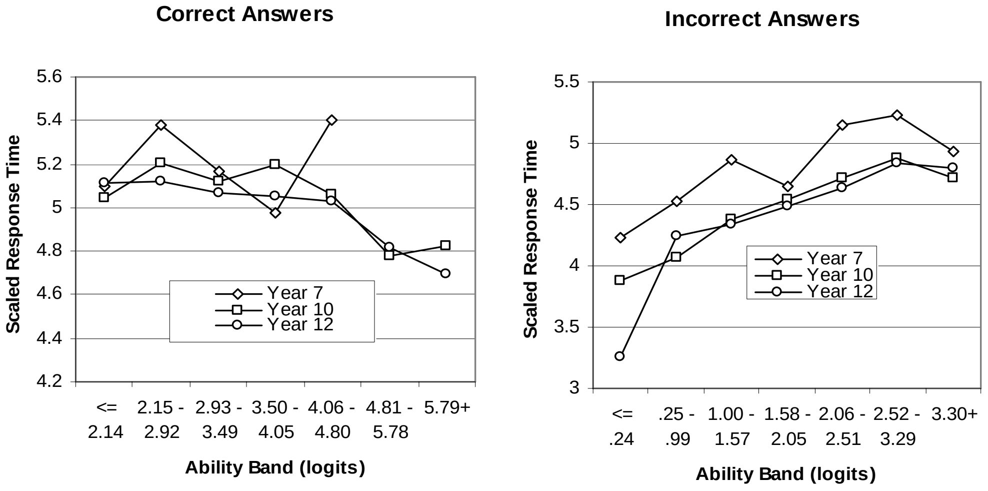


Figure 6 Distribution of scaled response time by ability for correct and incorrect answers on item 2359

Once again, for correct answers, response time was significantly affected by Year Group (H(2)=11.7, p=.003) and Jonckheere's test revealed a significant trend in the data: as year group increases, the median response time decreases, J=20392, z=-3.4, r= -.17. Post hoc Mann-Whitney tests of the difference for correct answers between the three year groups revealed a significant difference only between years 7 and 12 (U=4751, r=0.2) with the critical value for significance set at .0167 after application of the Bonferroni correction.

Perspectives

The use of computer based testing, including computer adaptive testing, represents an important advance in educational assessments. An important advantage of CBT is that it can record the time spent by test takers on specific test items which can be used for studying their behaviour during the testing process. Information gathered on item response time is very important for creating effective items and constructing effective tests to provide more accurate educational measurements. Results from this study indicate that the item response time is influenced by a range of factors, including the content domain and difficulty level of the items, and the age and ability of the test takers. Significant variation of response time exists between items and between students with different age and ability.

As the CABT employed in the current study represents a low-stake non-curriculum baseline test, the results obtained can be viewed as an complement to Moshinsky and Rapp's (2004) findings on a high-stakes adaptive test. Our results contradict their finding that there is no correlation between the time taken on items answered incorrectly with ability. The items in the CABT are not curriculum based and often presented in novel ways to the students. Thus it seems that able students persevere for longer trying to manipulate the item into a form they recognise. The positive correlation between time taken and ability for correct answers found by Moshinsky and Rapp (2004) is not always corroborated. It was most pronounced on the most difficult item where a medium effect size suggested it was an important factor in the response time. This item is most similar to the power items considered in Moshinsky and Rapp's (2004) study, requiring several logical steps, whereas the other items can be answered more quickly in less logical steps. This study furthermore considers the relationship between age and response time. The size of the effect seems to depend on the particular item involved.

While there is a positive correlation between response time and item difficulty, the variation is large. From a technical perspective this offers the opportunity to make the test more efficient, as difficult items that can be answered quickly can be retained in the item bank at the expense of difficult items that take longer to answer. It is not the case that all difficult items are time consuming.

Contrary to Moshinky and Rapp's (2004) findings there does not seem to be a stable relationship between performance time and response accuracy. This may be due to the different levels of familiarity that candidates have on the items.

Further study will involve the use of the response time obtained for individual items contained in the maths item bank of the CABT to design effective tests by selecting items requiring less time to answer at different difficulty levels to investigate the effect of such tests on the accuracy of student ability measurement.

Acknowledgements

The authors would like to thank Michael Cuthbertson, Daniel Bennett, Frank Bell and Peter Clark for their help with data collection.

References

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

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

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

Chiu, C. and I. Bejiar (2001) An empirical evaluation on the quantitative section of the computer-development and delivered GRE: generalizability analysis of response time and test score. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, USA.

Fitz-Gibbon, C. T. (1997) The Value Added National Project: Final Report *Feasibility studies for a national system of Value Added indicator*. School Curriculum and Assessment Authority, UK

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

Hambleton R. and H. Swaminathan (1983) *Item response theory: Principles and applications*. The Netherlands: Kluwer-Nijoff.

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

Hornke, L. (2000) Item response times in computerized adaptive testing. *Psicologica* 21: 175-189.

Lilley, M. and T. Barker (2003) An evaluation of a Computer Adaptive Test in a UK university context. *2003 CAA Conference Proceedings*: 171-182. Loughborough University, UK.

Masters G. and J. Keeves (1999) *Advances in measurement in educational research and assessment*. The Netherlands: Elsevier Science.

Moshinsky, A. and J. Rapp (2004). Performance Time on an Adaptive Power Test. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, USA.

Rasch G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Paedagogiske Institute.

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

Tonidandel, S., M.A. Quiñones and A.A. Adams (2002) Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions. *The Journal of Applied Psychology* 87: 320-332.

Tymms, P.B. (2001) The development of a computer-adaptive assessment in the early years. *Educational and Child Psychology* 18: 20-30.

Tymms, P.B., C. Merrell, and P. Jones (2004) Using baseline assessment data to make international comparisons. *British Educational Research Journal* (in press).

Wang T. and M.J. Kolen (2001) Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement* 38: 19-49.

Wright, B. D. and M.H. Stone (1979) *Best test design.* Chicago, IL: MESA Press.