Response Latencies and Ability Estimation in Computer Adaptive Assessment: A Modified Algorithm

Iasonas Lamprianou
European University Cyprus,
The University of Manchester, UK
lamprianou@yahoo.com

SUMMARY
Computer-based assessment has been used around the world to generate feedback and quantify pupils' learning or knowledge. Pupils' learning may be evaluated by studying the accuracy of their responses but much information may be unearthed by considering their response latencies (response time). This study proposes a new Item Response Theory algorithm for the estimation of the 'ability parameter' of pupils in the context of a computer adaptive testing system. The algorithm weights less the responses that are given too soon after the stimuli are presented, so they may not be considered as legitimate and honest efforts for a correct response. The application of the algorithm on empirical data (N=920) in the context of the mathematics National Curriculum in England gave indications of increased validity and usefulness of test results. The consequences on the classroom teaching and assessment routine are discussed.

KEYWORDS: *Computer-based assessment, Computer adaptive testing, IRT*

INTRODUCTION

Assessment and teaching are today considered to be two sides of the same coin (Heritage and Beile, 2006). In this context, computer-based assessment has been widely employed in classrooms in many countries because it offers flexible assessment opportunities, which are often more well-structured and readily available compared to teacher-made assessments. Wise (2006) has shown that in the context of low-stakes computer-based assessments, such as those used in every-day classroom routine assessment, not all the pupils attempt all questions/items with the same effort. In order to identify those pupils that do not exhibit full effort on all questions, and in order to avoid the 'contaminated' assessment results, Wise and Kong (2005) introduced a measure of examinee effort based on item response times (latencies) in computer-based tests; in fact, their research extended the work of Schnipke (1996) and Schnipke and Scrams (2002), who studied the rapid responses that examinees often give during tests.

Other researchers attempted different approaches/solutions to the problem. For example, Van Der Linden and Guo (2008, in press) proposed a hierarchical model where item response times and the scores on each of the items are used in order to identify memorization or cheating. Other research (e.g. Van Der Linden, 2007) has

also employed a hierarchical model to analyse the responses of people to computer-based tests

The paper aims to propose a modified IRT model that will accommodate for the fact that people taking a computer adaptive test sometimes exhibit rapid response behaviour which my contaminate the assessment information which is reported.

METHODOLOGY

Wise and Kong (2005) assumed that for each item i, there is a threshold, T_i, that represents the response-time boundary between rapid-guessing behavior (when an examinee answers either carelessly, or randomly, or is guessing etc) and solution behavior (the examinee attempts an item in his full potential in order to provide a thoughtful response). Extending this concept, Wise and DeMars (2006) suggested the use of the effort-moderated item response model, which incorporates the concept of item solution behavior in a 'traditional' Item Response Model. On the other hand, Mislevy and Bock (1982) appreciated the need for an Item Response Theory (hence, IRT) parameter estimation method that would be robust, without loosing too much information by dropping too many responses. Capitalizing on the ideas introduced by both Mislevy and Bock (1982) and Wise and Kong (2005) we suggest the computation of a more 'robust' Rasch estimate of examinee ability according to an iterative model. The one proposed by Mislevy and Bock (1982) is presented below:

$$\tilde{\theta}_v^{(k+1)} = \tilde{\theta}_v^{(k)} - \frac{\Sigma W_{vi}^{(k)} (X_{vi} - \tilde{P}_{vi}^{(k)})/S_i}{-\Sigma W_{vi}^{(k)} \tilde{P}_{vi}^{(k)} (1 - \tilde{P}_{vi}^{(k)})/S_i^2}$$

Equation 1

where W_{vi} is the weight of the response, X_{vi} is the observed response, P_{vi} is the expected response, S_i is the dispersion parameter and $\widetilde{\theta}_v$ is the ability of the examinee.

Drawing on the above concepts, this research acknowledges that when an examinee gives an answer too soon (according to a pre-specified criterion), then his/her response may not be accepted as solution behaviour but could be dropped or weighted less. After $T_{\min}$ seconds, the examinee is more likely to give a correct response, as he/she has more time to read the item more carefully, select the information needed, process it and provide an answer. For the purposes of this study, this time is called $T_{\exp}$. An examinee giving a response in t_{vi} seconds ($T_{\min} \leq t_{vi} < T_{\exp}$), has probably 'invested' the necessary time at least to partly deploy his/her full potential for a solution response. In this case, the response of the examinee will be weighted, according to the ratio of t_{vi} to the time $T_{\exp}$ judged by experts as necessary for a person to use his/her full potential for solution response.

We clarify some basic concepts with the following example: Two experienced mathematics teachers studied the item of Figure 1 and agreed that any response

given to this specific item within 3 seconds (i.e. 3sec < T_{exp}= 4sec) will not be considered as solution behavior (i.e. W_{vi} = 0) because 3 seconds is too soon for an 10-year-old pupil even to read the stem of the item (5 lines, around 40 words). However, if a pupil gives his/her response **after** the 3 seconds, but **before** 12 seconds elapsed (i.e. 11sec < T_{exp} = 12sec), his/her response will only be taken partly into account (i.e. 0 < W_{vi} < 1). The two teachers decided that any pupil would normally have adequate time within 11 seconds (a) to read the stem of the item, plus (b) read the table of 16 cells with words and numbers and select the necessary information, **but** (c) **would only partly** have time to process the information and key in a thoughtful response. On the item of Figure 1 the judges decided that pupils are reasonably expected to give their answers after 12 seconds, even if they are really fast.

However, using the RT alone, responses to this item given **at or after** 12 seconds (T_{exp} = 12sec), may be considered as full solution behavior (i.e. W_{vi} = 1).

$$W_i = \begin{cases} (1 - U_i)^2 \text{ where } U_i = \log(T_{i,\exp}) - \log(t_{vi}) \text{ when } T_{i,\min} \le t_{vi} < T_{i,\exp} \\ 0 \text{ when } t_{vi} < T_{i,\min} \\ 1 \text{ when } t_{vi} \ge T_{i,\exp} \end{cases}$$

Equation 2

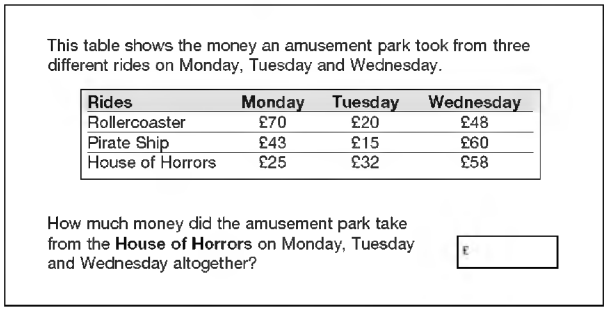
where t_{vi} is the time that took examinee *v* to give his/her response to item *i*. The estimation of the robust ability estimates is given by Equation 1 where W has the meaning explained above[1].

Although answering too soon is an indication of rapid response, answering too slowly is not necessarily an indication of solution response: we do not mean that $t_{vi} \ge T_{i,epx}$ is an indication of a thoughtful response. We acknowledge that an examinee might waste this time just by looking around the classroom (as we have personally witnessed this happening in practice in real life repeatedly). We also acknowledge that different pupils have different reading and reaction times, as well as different skills with computers.

The practical application of the method obviously demanded the operationalization of T_{min} and T_{exp}, that is, to assign a value for each item. In order to avoid the contamination of the results with subjective judgments, two experienced mathematics teachers were asked to study each of the items and to agree on T_{min} and T_{exp}values for each item. The teachers were asked to provide a more lenient and a more severe weighting set. In the first case, the teachers were instructed to be more

1. Hence, 'weighted ability' is estimated using equation 1 where W is defined by equation 2. When we simply refer to ability estimates, we refer to the use of equation 1 where W=1 (the 'usual' Rasch model). All estimations employ the UCON method (Wright & Stone, 1979) using the 'Analysis' package (Lamprianou, 2008). The standard error of the estimate refers to the asymptotic standard error.

lenient, therefore setting larger values of $T_{\min}$ and $T_{\exp}$ (ability estimates would not be too trimmed). In the second case, the teachers were instructed to be more severe, therefore setting smaller values of $T_{\min}$ and $T_{\exp}$ (ability estimates would be more trimmed).


This table shows the money an amusement park took from three different rides on Monday, Tuesday and Wednesday.

Rides	Monday	Tuesday	Wednesday
Rollercoaster	£70	£20	£48
Pirate Ship	£43	£15	£60
House of Horrors	£25	£32	£58

How much money did the amusement park take from the **House of Horrors** on Monday, Tuesday and Wednesday altogether?

£

Figure 1: *Item 11Q8a ($T_{\min}$ = 4sec, $T_{\exp}$ = 12sec)*

For IRT estimation purposes, the difficulty of the items was treated as fixed (we used the values provided by the test developer). Each ability was computed three times:

a) Under the null hypothesis using their full response pattern and the simple Rasch model, hence, producing a b_0 for each pupil v
b) Under the alternative hypothesis that some of the pupils exhibited rapid response behavior, estimating their weighted ability, hence b_w. For each pupil v
   a. we estimated a leniently weighted ability estimate, hence $b_{w_{lenient}}$;
   b. we estimated a severely weighted ability estimate, hence $b_{w_{severe}}$ by increasing the values of $T_{\min}$ and $T_{\exp}$ by 3 seconds.

RESEARCH AIM AND QUESTIONS

This study uses empirical data to investigate the effect of a 'response-time weighted' IRT algorithm on the ability estimates of pupils in the context of a computer adaptive testing system for the mathematics National Curriculum in England. The study has four research questions:

1. What will the effect of the suggested method be on the ability estimates?
2. What will the effect of the suggested method be on the ability standard errors?
3. Regarding the face validity of the method: Will the effects of the suggested method be defensible and reasonable under the light of specific case studies?
4. What are the possible implications for the every-day classroom practice?

THE TEST AND THE DATA

The computerized adaptive testing system used in this study is commercial

software built around the mathematics National Curriculum in England, for ages 7 to 14. It is used by schools as a classroom computer-based assessment instrument to provide diagnostic feedback to teachers regarding the mathematical performance of their pupils. The database of the version of the software of this study consists of around 400 items, either multiple choice or open-ended in format. The items consist of a stem and usually of a figure, table or picture (see Figure 1 as an example). The dataset used in this study consists of the responses of 920 pupils (8 to 11 years old) to 220 items (items covering the secondary education curriculum are not included in the sample). The items of the tests were calibrated by the test developer and their difficulty estimates (all items being dichotomously scored) ranged from -3.16 to 3.73 Rasch logits[2], with the mean difficulty set to 0 logits (standard deviation = 1.45 logits).

RESULTS

The pupils spent, on average, 31 seconds on each item, with response times ranging from 0 seconds to several minutes. The item with the shortest average response time (average RT=15 seconds) was answered by N=93 pupils. On the other hand, the item with the longest average response time (average RT=89 seconds) was answered by only 55 pupils. Most of the items were having, on average, a RT of 20 to 50 seconds. On average, the pupils responded to 37 items, however, they could, if they wanted (provided their teacher agreed), to stop the test at any time. As a result, a small number of pupils stopped the test after they completed only 5 items (minimum allowed by the test).

The parameters b_0, $b_{w_{lenient}}$ and $b_{w_{severe}}$ as well as e_0, $e_{w_{lenient}}$ and $e_{w_{severe}}$ were estimated (Table 1). From the 920 pupils, 31 had full (all responses correct) or zero score (all responses incorrect), and were excluded[3] from this part of the analysis.

Table 1: Comparison of ability estimates and standard errors

(N=889)	Ability Estimates			Standard Errors		
	Rasch	'leniently' Weighted	'severely' Weighted	Rasch	'leniently' Weighted	'severely' Weighted
Mean	-0.85	-0.82	-0.79	0.53	0.54	0.55
Median	-0.85	-0.80	-0.75	0.43	0.43	0.44
Stand Deviation	1.46	1.47	1.46	0.25	0.29	0.35
Minimum	-5.03	-7.98	-7.96	0.31	0.31	0.32
Maximum	3.84	3.84	3.84	2.41	3.9	4.91

2. A logit (log-odds unit) is a unit of interval measurement which is well-defined within the context of a homogeneous test. The definition of the odds of a correct answer is the ratio of the probability of it occurring to the probability of it not occurring and is governed by the ability of the pupil and the difficulty of item.

3. The UCON estimation method used in this study does not converge for extreme scores.

The change between the Rasch and the leniently weighted ability estimates is statistically significant according to a paired t-test (t=-4.291, p<0.001) but negligible. Also, the change between the Rasch and the severely weighted ability estimates is statistically significant according to a paired t-test (t=-5.768, p<0.001) but negligible.

The standard errors of the weighted ability estimates are always larger than the standard errors of the Rasch ability estimates. This is expected since the weighted ability estimates are computed using only part of the information. Overall, under the leniently weighted model, 56.5% of the pupils had no responses discarded or weighted. In total, 74.8% of the pupils had no discarded responses or lost less than 3% of their response information[4]. Finally, 5.4% of the pupils lost more than 10% of their response information. Still, the loss of information was not so important, since the standard errors of the ability estimates were not affected noticeably.

Overall, under the severely weighted model, 39.6% of the pupils had no responses discarded or weighted. In total, 55.8% of the pupils lost no responses at all or lost less than 3% of their response information. Finally, 6.6% of the pupils lost more than 10% of their response information. Still, the loss of information was not so important, since the standard errors of the ability estimates were not affected noticeably.

A very small number of pupils (2.5%, N=23) 'suffered' sizable losses of information. We will next inspect two of those cases more closely.

Case Study 1: "Peter's" rapid responses at the end of the test

Peter is a primary school pupil who took the test in May 2007 during a normal class session. He took the usual 3 practice items before the beginning of the test, and then responded to 35 mathematics items. He received a raw score of 9 (out of 35) and an ability estimate of -0.73 logits, almost identical to the average of the whole sample.

This profile is not making any justice to Peter. He spent 4 minutes and 27 seconds on the test. He responded to the last 20 items (average difficulty 0.33 logits) within 53 seconds and got none correct. He spent 214 seconds for the first 15 items (average difficulty 0.75 logits) and got 9 correct.

Using the lenient weighting method, Peter lost 48% of his response information, because most of his responses to the last 20 items were discarded or weighted less. His ability estimate rose to 0.93 logits and the error of estimate rose only to 0.51 (from 0.43). Using the severely weighted method, the ability estimate of Peter rose to 1.54 logits by loosing, 66% of his response information. Figure 2 shows that, for the right response, the time Peter spent is proportional to the average time the rest of the pupils spent on each of the items. However, even in the first

4. When we say that a pupil lost x% of the information of his/her response pattern, we mean that the standard error of his/her ability estimate increased by x% as a result of the time-weighting of his/her responses.

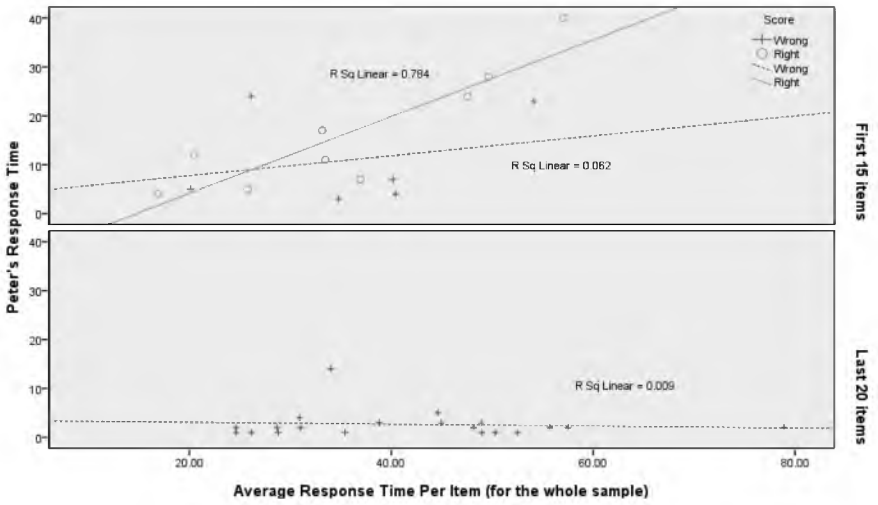


Figure 2: *Response time and score (right/wrong) for items answered by Peter*

15 items, he spent less time on the items he got wrong than the average time the rest of the sample spent on those items. Therefore, the weighted algorithm has arguably managed to create a more robust picture of Peter's performance on the test.

Case Study 2: "Kate's" rapid responses across the test

Kate is a primary school pupil who completed 55 items on the test. She received a score of 10, and an estimated ability of -2.13 logits. In total, she spent 933 seconds (around 15 ó minutes); an average of 17 seconds per item. She has given a series of rapid responses (RTs of 1-3 seconds) which are very difficult to explain, in the sense that they are scattered across the test (Kate gave rapid responses to items on which the rest of the sample 'invested' much time). Using the leniently weighting method, Kate lost 13.8% of her response information. Her ability estimate rose slightly to -2.04 logits and the error of estimate rose only to 0.40 (compared to the unweighted 0.39). Using the severely weighted method, Kate's ability estimate rose to -1.83 logits but lost 28% of her response information.

Kate was presented, immediately after the end of the test, (on the screen of the computer) the question "How do you rate your own computer skills"? Her response was "Very Good" and her teacher later agreed. She was also presented with another question "How more difficult was it to take the test on the computer rather than take a paper and pencil test"? and she responded "The same". She also said that she were about the same as nervous as she would be when taking a paper and pencil test.

Kate invested more time on the more difficult items and less time on the more difficult items. Kate spent proportionally a lot of time on items she finally got correct, but less time on items she got incorrect. She also invested more time while the difficulty of the item increased, but this was true only for the items she finally got

right. She invested less time while the difficulty of the item increased (for the items she finally got wrong).

CONCLUSION

The study showed there were pupils who systematically generated rapid responses and others that would only produce rapid responses occasionally: the information conveyed by those responses is likely to mislead the teachers about the mathematical ability of the pupils. Incorporating response time weighted routines in computer adaptive tests should technically be a straightforward issue. Provided the teachers find the facility useful in their every-day work, this method could save a lot of time and improve the quality of the feedback the teachers and the pupils get from computer-based assessment.

This research has not taken into account important background information of the pupils such as their typing speed, motoric dexterity, their computer literacy skills and the like. Future research should take these factors into account. The teachers should always take into account these background factors before deciding to assess competencies or abilities using computer-based assessment techniques.

REFERENCES

Heritage, M., & Baile, A. L. (2006). Assessing to teach: an introduction. Educational Assessment, 11 (3 & 4), 145-148.

Lamprianou, I. (2008). Analysis: Rasch analysis [Computer software]. Cyprus: ReLabs Research Laboratories. Retrieved from the Web (free), December 20, 2006. www.ReLabs.org/software.html

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. Educational and Psychological Measurement, 42, 725-737.

Mosteller, F., & Tukey, J. (1977). Exploratory data analysis and regression. Reading, Mass: Addison-Wesley.

Schnipke, D. L. (1996, April). How contaminated by guessing are item-parameter estimates and what can be done about it? Paper presented at the annual meeting of the National Council on Measurement in Education, New York. (ERIC Document Reproduction Service No. ED400276)

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.). Computer-based testing: Building the foundation for future assessments (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Searle, J., & Hallam, N. (16-21 September, 2007). Response times for mathematics items in computer adaptive testing in secondary schools. Paper presented to the 33rd IAEA Annual Conference, Baku, Azerbaijan.

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

Van Der Linden (2007). A hierarchical framework for modelling speed and accuracy on test items. Psychometrika, 72 (3), 287-308.

Van Der Linden, W. J., & Guo, F. (2008, in Press). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. Psychometrika.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. Applied Measurement in Education, 19 (2), 95-114.

Wise, S. L., & DeMars, C. E. (2006). An application of Item Response Time: the effort-moderated IRT model. Journal of Educational Measurement, 43 (1), 19-38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. Applied Measurement in Education, 18, 163-183

Wright, B.D. & Stone, M.H. (1979). Best test design. MESA: Chicago.