



Beyond test accuracy: Benefits of measuring response time in computerised testing

Eugene Gvozdenko and Dianne Chambers
The University of Melbourne

This paper investigates how monitoring the time spent on a question in a test of basic mathematics skills can provide insights into learning processes, the quality of test takers' knowledge, and cognitive demands and performance of test items that otherwise would remain undiscovered if the usual test outcome of accuracy only format ("correct/incorrect") was used. Data from three tests of basic mathematics skills taken by tertiary students in 2004-2006 were analysed. Means and distributions of individual response times on parallel test questions were examined and differences were further investigated.

Analysis of response time data revealed a number of surprising findings in regard to the impact of variables on preferences for written and mental calculation methods and regarding additional cognitive demands of a question. The study examined how simple statistical analysis of response time distribution can be used to investigate abnormalities of the item functioning. These findings may be of value to educators and to test producers by informing them about the potential of utilising response time measurements as a diagnostic facility in computerised tests, for the purposes of improving teaching and learning.

Introduction

The work reported in this paper is situated in the field of computerised testing, an emerging educational technology, which is gaining great societal importance with the current shift in assessment paradigms (Stout, 2002). The summative assessment testing paradigm that has driven test measurement research and testing practice for last fifty years is giving way to a new paradigm that is focusing on formative assessment with an informative function, and whose primary goal is to enable all parties involved in the cognitive process to monitor proactively the progress of skill acquisition. In addition to delivering final test scores, the new approach embraces research into the process of skill formation, aiming to guide teaching and learning whilst it is occurring.

The introduction of computers into assessment is a response to the need to improve assessment processes by maximising evaluation precision and fairness, and minimising the costs. Whereas the cost of teaching is more related to the hours per course and is not greatly affected by how many students attend a lecture, the cost of assessment is a cost per student, which grows proportionally with student numbers. The cost of assessment in higher education is the most rapidly growing component of tuition fees (Ricketts, Filmore, Lowry & Wilks, 2003).

Computerised assessment makes new techniques of unobtrusive data collection available for the purpose of monitoring otherwise hidden characteristics of learning progress. Psychometric analysis of test takers' behaviours was suggested by Masters and Keeves (1999) for the assurance of the quality of assessment measurements. Weiss and Schleisman (1999), Schnipke and Scrams (1997, 1999a, 1999b, 2002), Schnipke and Pashley (1997), Hornke (1997, 2000) and Bergstrom Gershon and Lunz (1994) considered speed an important component of ability and drew attention to the need to develop testing models that would include test takers' response time, that is, the amount of time a test taker spends on a test item in test scoring procedures.

A number of studies (Wainer et al., 2000; Wainer & Eignor, 2000; Schnipke & Scrams, 1997; 1999b; Hornke, 2005) have suggested that further research is required to investigate how response time measurements can improve the precision of cognitive tests by offering additional information about the impact of the question on a test taker.

Schnipke and Scrams (1997), using Thissen's Timed Testing model as a framework for their research, examined the data from computer based tests of verbal, quantitative and reasoning skills involving 7,000 examinees. The study concluded that question response time and accuracy statistics provide separate measures of question functioning.

The results of a previous study (Gvozdenko, 2005) confirmed that response time is a separate variable with a weak relationship with accuracy (Pearson correlation = -0.33), which means that questions that required longer time to complete also tended to result in a slightly higher percentage of incorrect answers.

Descriptive statistics of response times on a test question were suggested as a stable parameter of the question functioning when the number of test takers exceeds 30 people (Gvozdenko, 2005). This paper examines further

the potential for using Mean Question Response Time (MQRT), in combination with distribution analysis of response time, to deliver additional information in the course of monitoring test takers' temporal behaviours and evaluating the equality of test questions.

The current study continues to explore the issue of the usefulness of monitoring response times in the context of testing different cognitive domains in the area of basic mathematics skills. It demonstrates how response time measurements can provide valuable insights for educators, revealing otherwise hidden issues about test takers' current understandings and learning progress, and identifying content areas in which unexpected patterns of time demand prompt further investigation.

Method

The investigation utilises the quasi-experimental type of design justified by the importance of studying cognitive processes in a natural setting (Cook & Campbell, 1979). It reports on three studies conducted in 2004, 2005 and 2006 in an Australian tertiary institution. Question Response Times (RT) were collected from three different cohorts of second year tertiary students: 135 students in 2004, 189 students in 2005, and 203 students in 2006, mostly females between the ages of 18 and 25, who sat a computer based test in basic mathematics skills.

The 2004 and 2005 tests used the same items with random selection of a question from a pool of several parallel questions for each item. The 2006 test had three parallel versions of the test with fixed questions. Items on all three tests (2004, 2005, 2006) were designed to test the same set of cognitive skills.

The tests were a part of a university subject, which ensured that students were motivated to complete all questions. A website for practice tests with unlimited access from campus and from home was made available for two weeks before the real tests. Among other benefits, the practice helped students alleviate any test anxiety. After each practice test the feedback, comprising correct answers and a total score, was automatically generated and presented to students. The level of computer anxiety was assumed to be low as all students had taken a computer skills subject in the previous semester.

The actual tests were administered in online mode under supervised and time restricted conditions in an on campus computer laboratory. Identical hardware and equal speed of broadband connection provided equal technological conditions for all participants. The test duration of 55 minutes allowed all students sufficient time to complete the test. The test can be considered a 'power test', as all items could be answered without the influence of a speed factor, rather than a 'speeded test' which is characterised by speeding behaviour like accelerating or rapid guessing observed for last questions (Schnipke & Scrams, 1999a). The notion that this was a power test is also supported by almost all students completing a practice test within 45 minutes.

Test takers were informed that their response times would be measured but that the measurements would not affect their test scores.

The option of browsing between the questions was allowed to reduce the difference between the computerised and the paper and pen versions of the test, and test takers were able to change their answers using the option of returning to previously answered questions before the final submission of the test.

The online version of the test was based on *Test Pilot* software (McGraw-Hill *Test Pilot Enterprise* (v4), <http://www.clearlearning.com/>). A server side program registered time stamps generated by Java scripts inserted into each web page, to measure the time between downloading different web pages on a client side machine, as test takers browsed between test questions. Each question was presented on a separate web page, which allowed measuring the time spent on each question - a measurement technique employed by Bergstrom, Gershon and Lunz (1994).

In 2004 and 2005 the test takers were presented with a non-adaptive test comprising 26 questions that were randomly drawn out of 24 subgroups containing a total of 72 questions. Questions within each subgroup had been pre-calibrated and evaluated by an expert as being of equal cognitive load. In 2006 the test takers were randomly assigned to one of three parallel versions of the test. Each version had 27 questions. The versions had been prepared by a mathematics content specialist and were expected to be equal. The test takers scheduled themselves to one of ten groups who sat the test in a computer laboratory within one week. The study calculated and interpreted descriptive statistics, such as mean and standard deviation, to quantify typicalities, diversity and relationships among response time and accuracy variables.

Considering the concerns that traditional RT analysis, when restricted to mean RT, could be misleading due to positive skewness of the data (Heathcote, Popiel, & Mewhort, 1991; Brown & Heathcote, 2003) the study also engaged distribution analysis to characterise the spread of response times among test takers and to locate and delete outlier RTs (Ratcliff, 1978).

Results

In this section the study will present further validation of MQRT as a useful measure of a test item performance in regard to identifying parallel questions that have unexpected differences in functioning. Distribution analysis will be undertaken to describe the difference and a hypothesis of the origin of the difference will be generated. Further investigation of the cognitive difference will test the hypothesis and demonstrate how response time measurements can provide useful information.

Reliability of MQRT

Approaching the issues of the usability of RT measurements, the study had first to address the issue of reliability of Mean Question Response Time (MQRT) as a descriptor of test question performance. As one of the possible distortions of MQRT can be caused by outliers, histograms of distributions of individual RTs on each questions were examined to identify outliers in the upper and lower 5% of the distributions. The deletion of outliers reduced MQRT by only 7% on average while the reduction in standard deviation was much greater (34%), which means that removing outliers reduced positive skewness of the distribution curve to a greater extent than it affected MQRT.

Consistency and replicability of MQRT on parallel versions of the 2006 test would support the reliability of the measurements. Figure 1 illustrates the comparison of MQRT between the three versions of the test. It can be seen (Figure 1) that on most test questions, MQRTs of all versions of test questions are close.

One way analysis of variation (ANOVA) returned p-values >0.05 on most items, which indicates that variation between test versions on those items is not statistically significant. However, the analysis of the variation on Q140 and Q240 demonstrated a statistically significant difference between the test versions (p-value <0.001). Q140 will be in the focus of the discussion in this paper.

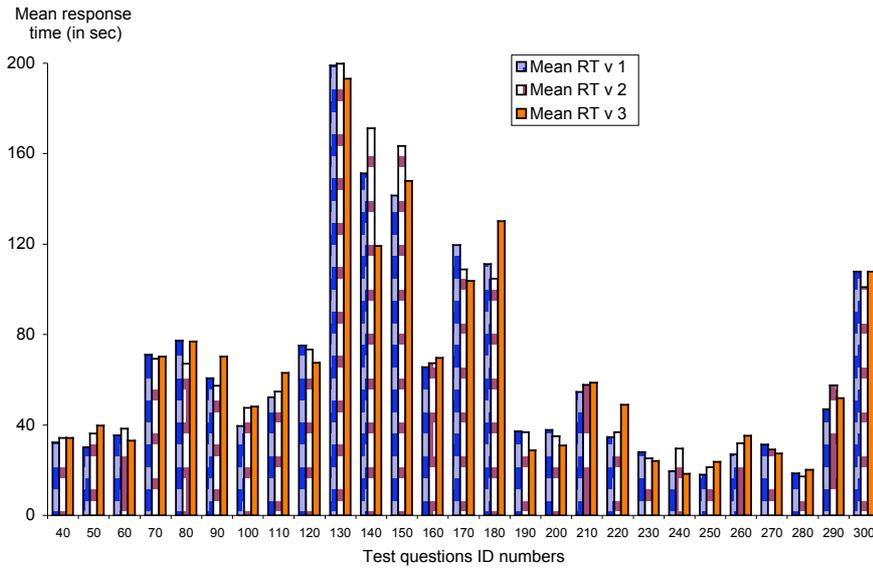


Figure 1: MQRT of three parallel versions of 2006 test (n > 60 for each version)

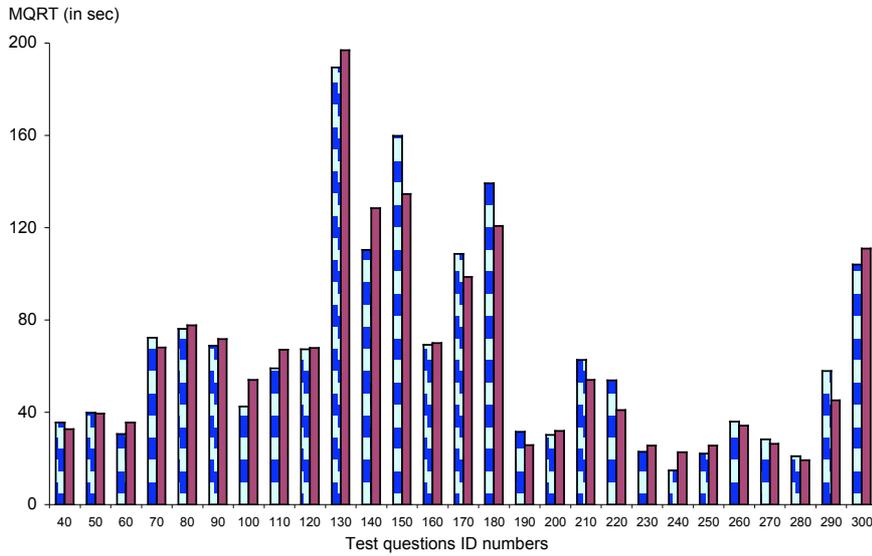


Figure 2: MQRT of halves of the sample of Version 3 of 2006 test (n > 30)

Further analysis of RT within a version of the test was then undertaken to examine whether MQRTs of randomly selected halves of the same version (Figure 2) remain consistent across the test.

It can be seen that after reduction of the sample size from 60 to 30 test takers the variation in MQRT on most test questions was insignificant, with a very strong correlation between the two MQRTs (Pearson correlation = 0.98). Thus, MQRT demonstrated sufficient reliability as an indicator of the impact of a test question on test takers' temporal behaviour.

The monitoring of equality of parallel test questions is suggested as one of the potential uses of MQRT. Although equal MQRT on parallel questions does not necessarily indicate equality of test questions, different MQRTs should be considered a prompt for further investigation in a form of distribution analysis. This notion will be further supported by a few examples.

Applicability of response time (RT) measurements

Identifying differences in solution strategies

Table 1: Targeting Question 140 for further investigation.

Test version	Task	Accuracy (% of correct answers)	MQRT (in sec)
V 1	Izabella is washing windows. It takes her 5 min to wash and polish 1 square metre area of glass. The room she is cleaning has 8 rectangular windows, 50 cm wide and 1.8 m high. How long will it take her (please, round your answer in minutes)?	70%	151.4
V 2	Izabella is washing windows. It takes her 3 min to wash and polish 1 square metre area of glass. The room she is cleaning has 8 rectangular windows, 75 cm wide and 6 m high. How long will it take her (please, round your answer in minutes)?	82%	171.1
V 3	Izabella is washing windows. It takes her 4 min to wash and polish 1 square metre area of glass. The room she is cleaning has 5 rectangular windows, 50 cm wide and 2.4 m high. How long will it take her (please, round your answer in minutes)?	77%	119.1

Analysis of the differences in MQRT between three parallel versions of 2006 test (Figure 1) flagged Question 140 as the question with the biggest

variation (52 sec) between MQRTs of Version 2 and Version 3 (Table 1) (t-test p-value < 0.001); it took test takers 43% longer to complete the question offered by Version 2, which also raises concerns about test fairness.

Histograms presented in Figures 3-5 demonstrate that different versions of Question 140 produced different distributions of test takers' individual RTs. The right side vertical axis represents cumulative percentage of students and helps to illustrate the time range within which most students completed the task. The left side vertical axis represents the number of students for each time bar.

It can be seen (Figure 3) that distribution shape of Version 1 is not greatly skewed. Most RTs are evenly spread between 110 sec and 230 sec. Version 2 (Figure 4), however, is characterised by a shift to the right and negative skewness, which demonstrates that only few test takers were able to complete the question quickly and only 47% of test takers did it within 170 sec.

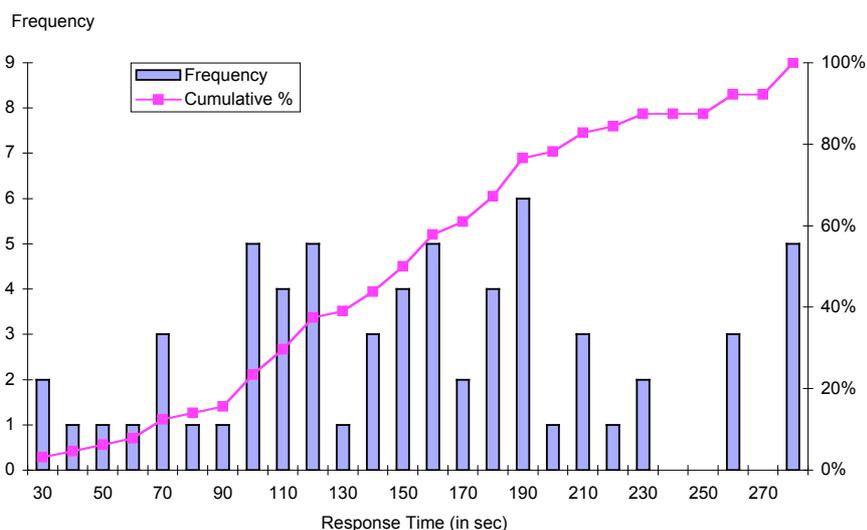


Figure 3: Distribution of test takers' response times on Q 140, Version 1.

Conversely, Version 3 (Figure 5) demonstrates a shift of responses to the left resulting in positive skewness, which means that many test takers answered the question quickly and 80% of test takers completed the question within 170 sec. Thus, the distribution analysis of responses

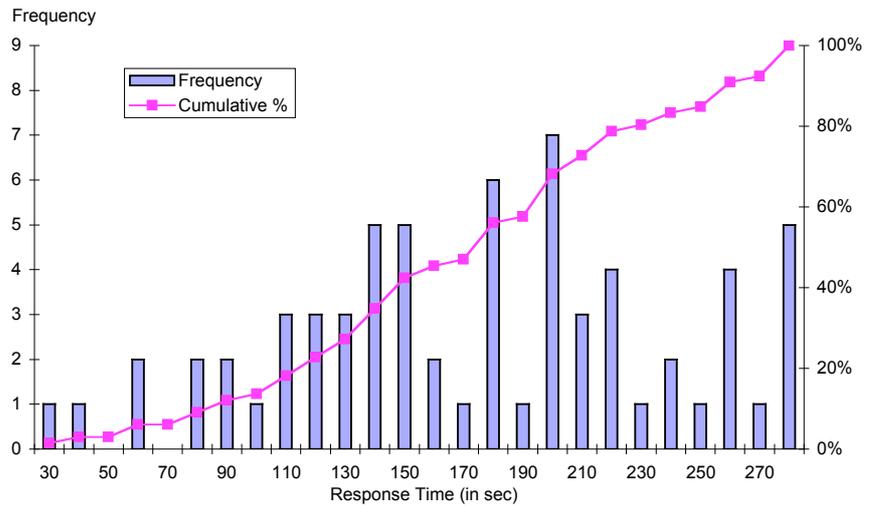


Figure 4: Distribution of test takers' response times on Q 140, Version 2.

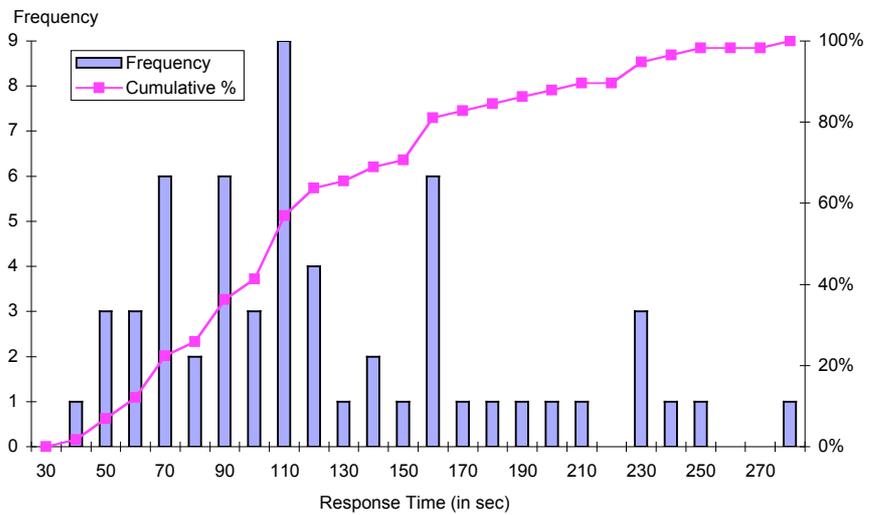


Figure 5: Distribution of test takers' response times on Q 140, Version 3.

supported and expanded information about the differences observed in MQRTs. To confirm visual observations of the difference in distributions, the study used the software *QMPE* (NCL Software Repository, 2007), an open source ANSI Fortran 90 program for response time distribution

estimation. *QMPE* fits the ex-Gaussian distribution, a positively skewed distribution produced by the convolution of a normal and exponential distribution (Heathcote, 1996). The ex-Gaussian has three parameters, the mean and standard deviation of the normal component and the mean of the exponential component. It allows the breakdown of MQRT into a normally distributed component and an exponentially distributed upper tail that determines the skewness of distribution. These two components are affected differently by experimental manipulations. The mean of the normal component reflects a uniform time demand and for simple arithmetics tasks is mainly determined by memory retrieval time. The exponential component is found to reflect the procedural strategies that have been used by test takers (Campbell & Penner-Wilger, 2006).

As can be seen (Table 2) the mean of the normal component and its standard deviation for Version 3 are half that of the corresponding values of Version 2, however the mean of the exponential component is more than twice as large. How the relationship between these parameters can be applied to cognitive diagnostics could be a fertile ground for further research.

Table 2: Results of distribution analysis of RTs (Question 140, test 2006).

Test version	Sample size	Mean of the normal component	Standard deviation of the normal component	Mean of the exponential component
1	58	117.1	59.5	33.3
2	64	146.4	65.0	21.5
3	63	60.2	27.8	56.4

After establishing and quantifying the differences with respect to RT between versions of the test question, the study proceeded to investigate causes of the variation. It was hypothesised that the difference in MQRTs and RT distributions may be associated with test takers preferring different strategies when approaching these questions. A content expert suggested that most test takers would follow three steps of the solution where each of the steps could be done as mental calculation or written calculation and that the difference in time demands between the two methods would generate difference in response time.

Suggested steps of solution for Question 140, Version 1:

- Step 1:* $0.5 \times 1.8 = 0.9$ [calculate area of one window]
Step 2: $0.9 \times 8 = 7.2$ [calculate total window area]
Step 3: $7.2 \times 5 = 36$ [calculate time to clean all windows]

To test this hypothesis another cohort of the same population was given the same questions as part of formative testing. In the following question of the test they were asked to indicate the strategy that they had employed to solve the problem. Their preferences are illustrated in Figure 6 as a distribution between written strategy and mental strategy for each of the steps. It can be seen that at step 1 the percentage of the test takers who chose written strategy is between 60 and 70 per cent for all three versions. At the second step Version 1 produced more answers by written method of calculation than other versions. However, at the third step Version 3 allowed two thirds of test takers to use mental calculation strategy.

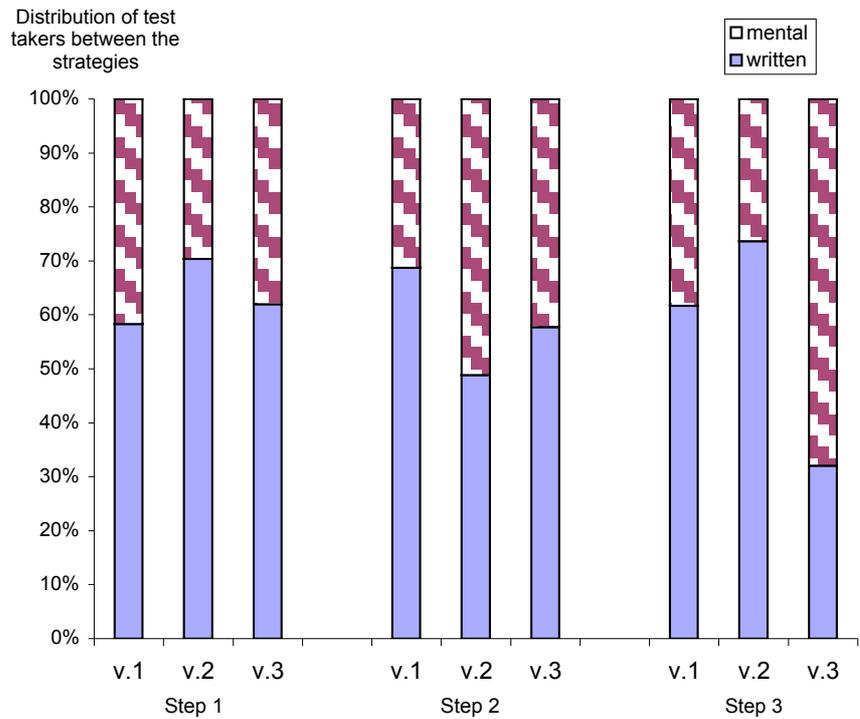


Figure 6: Comparison of written vs. mental methods of calculation for the three steps of the three versions of Q140 of 2006 test.

For Version 2 of the question, test takers demonstrated a preference for the written method in each solution step. Interestingly, it also produced the highest accuracy (82%), perhaps because some of the test takers who

chose mental method on Version 3 (accuracy = 77%) overestimated their ability in mental calculations. However, in this case the difference in accuracy (5%) is only marginal and on its own would hardly trigger further investigation into cognitive process. The analysis of response time data was able to detect the difference and offer substantial grounds for generating a hypothesis about the cognitive nature of the phenomenon.

Identifying differences in cognitive demand

In this section we will present another example of a test where mean question response time (MQRT) data identified questions with different time demands, when a similar cognitive load and time demand were expected. As in the previous example, in this case the difference in cognitive load would go unnoticed if the analysis of test item performance was based on accuracy statistics only. Further investigation of these questions generated insights into the nature of the observed differences.

In the 2005 experiment, 157 test takers were presented with one of two questions randomly drawn from the test pool. The questions were intended to test the skill of conversion from square metres into hectares (Table 3).

Table 3: Conversion from square metres into hectares (2005)

Questions	Q 210	Q 530
Task	27500 m ² = ? (ha)	690 m ² = ? (ha)
Number of test takers	66	91
Index of accuracy (% of right answers)	69%	72%
Mean question response time (in sec)	48	62

These questions would appear to be equal if only the usual measure of index of accuracy was used. However, the values of MQRT indicated that time demands for those two questions differ significantly. The observed difference in MQRT prompted further investigation into the probable cause. The distribution of response times for each task can be seen in the following histograms (Figures 7 and 8).

Analysis of cumulative frequency (Figures 7-8) demonstrated that while half of the test takers completed question 210 within 40 seconds and 90% of test takers did it within 87 seconds, for question 530 it took 60 seconds for the first half of the test takers and up to 105 seconds for 90% of the test takers to complete the question. A significant number of test takers (one way ANOVA p-value <0.05) were thus experiencing additional processing load.

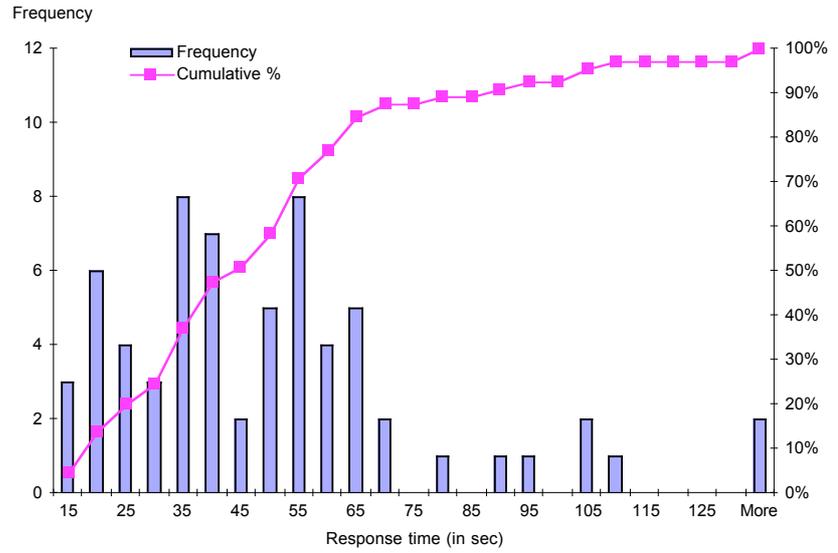


Figure 7: Distribution of response times for question 210 (2005).

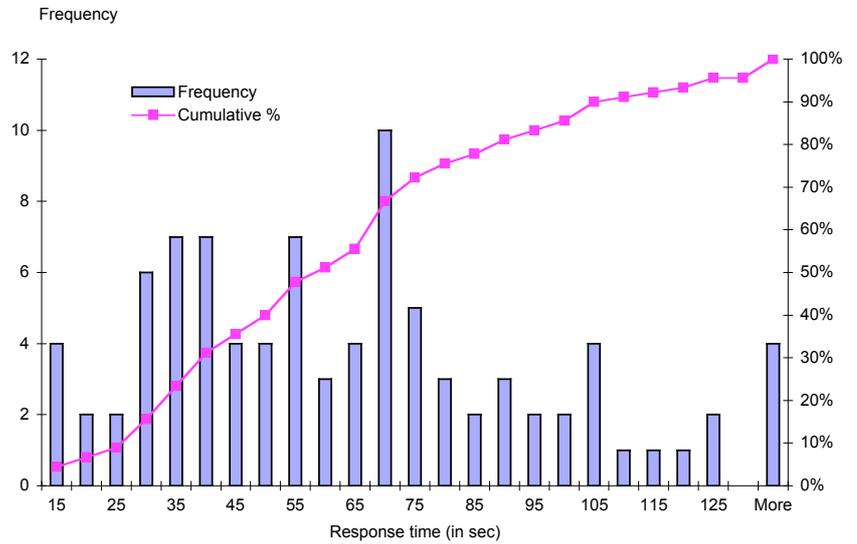


Figure 8: Distribution of response times for question 530 (2005).

The analysis of variation of response times between similar questions in 2004 (Q 210 and Q 530) returned p -value = 0.88, which indicates that the variation between those questions in 2004 was not statistically significant. Comparison between 2004 and 2005 questions (Table 4) led to the proposal that the identified difference may be attributed to the different number of digits in the value being transformed.

Table 4: Difference in MQRT due to additional cognitive problem in the process of conversion

Question	Accuracy (% of correct)	MQRT (in sec)	Value to transform	Expected result
Q 210 (2004)	77	44	12560	1.256
Q 530 (2004)	78	43	9570	0.957
Q 210 (2005)	69	48	27500	2.75
Q 530 (2005)	72	62	690	0.069

It was hypothesised that, as this kind of conversion requires a decimal point to be moved to the left by four digits, the test takers may encounter an additional problem when there are not enough digits just to move the decimal point. They have to recall the procedure of inserting an additional zero after the decimal point in Question 530 (2005), which increases the time spent on the question.

Further analysis of test takers errors confirmed the association of wrong answers involving a misplaced decimal point with the delays in RTs. Consultation with a mathematics educator with specialisation in decimal content (Steinle, 2004) confirmed the findings. This example demonstrates how RT measurements can initiate an enquiry into the test contents and generate a deeper understanding of the cognitive demands of a test question.

Conclusions

The mean question response time (MQRT) has thus been shown to be a valuable source of information about the cognitive load of test questions. It allows identification of otherwise hidden differences in test takers' behaviour. Using only the item's accuracy statistics, as is currently done in most testing models, seems to be insufficient. It was found that a different set of variables in the same task could considerably alter (up to 40%) time demands of the question without a significant impact on accuracy statistics, which raises serious concerns about test fairness in regard to

equal time load, especially in high stakes adaptive tests, where additional time provides opportunities to complete a higher stage.

The study found a number of patterns in test takers' response times in regard to the conversion of measurement units. Analysis of test takers' behaviours on other tasks of basic mathematics is in progress.

It is proposed that question response time and other measures of test takers' temporal behaviours that can be captured only through computerised testing can generate new understandings about test takers' decision making processes and inform improvement of the curriculum. It is an efficient tool that enables educators to focus on specific cognitive skills for the purposes of educational research and practical teaching.

Acknowledgements

The authors acknowledge support of the Faculty of Education, The University of Melbourne, the scholarly input of our colleagues Dr Helen Chick, Dr Steinle and Professor Kaye Stacey, the students who participated in the project, and the Australian Postgraduate Award.

References

- Bergstrom, B., Gershon, R. & Lunz, M. E. (1994). Computerized adaptive testing: Exploring examinee response time using hierarchical linear modelling. Paper presented at the Annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Brown, S. & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behaviour Research Methods, Instruments & Computers*, 35(1), 11-21. [verified 7 Oct 2007] <http://www.newcastle.edu.au/school/psychology/ncl/publications/averaging-curves-web.pdf>
- Campbell, J. & Penner-Wilger, M. (2006). Calculation latency: The $[\mu]$ of memory and the $[\tau]$ of transformation. *Memory and Cognition*. 34(1), 217-226.
- Cook, T. D. & D. T. Campbell (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.
- Gvozdenko, E. (2005). *Question response time in computerized testing: Applicability for test design and prediction of error*. Masters thesis, University of Melbourne, Melbourne.
- Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavioural Research Methods, Instruments & Computers*, 28, 427-445.

- Heathcote, A., Popiel, S. & Mewhort, D. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, 109(2), 340-347.
- Hornke, L. F. (1997). Investigating item response times in computerized adaptive testing. *Diagnostica*, 43(1), 27-39.
- Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicologica*, 21(1), 175-189.
- Hornke, L.F. (2005) Response time in computer-aided testing: A "verbal memory" test for routes and maps. *Psychology Science*, 47(2), 280-293.
- Masters, G. & Keeves, J. (1999). *Advances in measurement in educational research and assessment*. Amsterdam and New York: Pergamon.
- NCL Software Repository (2007). QMPE. Newcastle Cognition Laboratory, School of Psychology, The University of Newcastle, Australia.
http://www.newcastle.edu.au/school/psychology/ncl/software_repository.html
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59-108.
- Ricketts, C., Filmore, P., Lowry, R. & Wilks, S. (2003). How should we measure the costs of computer aided assessment? *Proceedings 7th Computer Assisted Assessment Conference*, Loughborough University, UK. [verified 7 Oct 2007]
<https://dspace.lboro.ac.uk:8443/dspace/bitstream/2134/1924/1/ricketts03.pdf>
- Schnipke, D. L. & Pashley, P. J. (1997). Assessing subgroup differences in item response times. Paper presented at the Annual Meeting of the National Council on Measurement in Education 25-27 March 1997, Chicago, IL.
- Schnipke, D. L. & Scrams, D. J. (1997). Making use of response time in standardized tests: Are accuracy and speed measuring the same thing? Paper presented at the Annual Meeting of the National Council on Measurement in Education 25-27 March 1997, Chicago, IL.
- Schnipke, D. L. & Scrams, D. J. (1999a). Exploring issues of test taker behaviour: Insights gained from response time analyses. Princeton, NJ: Law School Admission Council.
- Schnipke, D. L. & Scrams, D. J. (1999b). Representing response-time information in item banks. Princeton, NJ: Law School Admission Council.
- Schnipke, D. L. & Scrams, D. J. (2002). Exploring issues of examinee behaviour: insights gained from response-time analysis. In Mills, C. N. (Ed.), *Computer-based testing: building the foundation for future assessments* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.
- Steinle, V. (2004). *Changes with age in students' misconceptions of decimal numbers*. PhD thesis, University of Melbourne, Melbourne.
- Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, 67(4), 485-518.

Wainer, H. & Eignor, D. (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing. In H. Wainer, & N. J. Dorans (Eds), *Computerized adaptive testing: A primer* (pp. 271-299) (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L. & Thissen, D. (2000). Future challenges. In H. Wainer & N. J. Dorans (Eds), *Computerized adaptive testing: A primer* (pp. 231-269) (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. & Schleisman, J. L. (1999). Adaptive testing. In G. N. Masters & J. P. Keeves (Eds), *Advances in measurement in educational research and assessment*. Amsterdam and New York: Pergamon.



Eugene Gvozdenko is a PhD student with the Faculty of Education at the University of Melbourne, Australia. Eugene has teaching experience in the area of foreign languages and IT in education in secondary and adult education. Research interests include Information Technology in Education with the focus on online testing, and cognitive neuroscience.

Eugene Gvozdenko (eugeneg@unimelb.edu.au), Faculty of Education, The University of Melbourne, Victoria 3010, Australia



Dianne Chambers is a Senior Lecturer and Assistant Dean (Learning Technologies) with the Faculty of Education at the University of Melbourne, Australia. Dianne teaches in the area of IT in education in early childhood, primary, secondary and adult education. Research areas include IT in teacher education and professional development, online education including online testing, and technology enriched, problem based learning.

Dr Dianne Chambers (d.chambers@unimelb.edu.au), Faculty of Education, The University of Melbourne, Victoria 3010, Australia