Question Response Time in Computerized Testing:

Applicability for Test Design and Prediction of Error

Evgueni Gvozdenko

B.Ed.(Hons)
(Taganrog Pedagogical Institute)

Submitted in partial fulfilment of the requirements for the degree of
Masters of Information Technology in Education
in the Faculty of Education at The University of Melbourne

2005
# Table of Contents

Table of Contents ........................................................................................................ i  
List of Figures .................................................................................................................. iii  
List of Tables .................................................................................................................. iii  
Declaration of originality ................................................................................................. iv  
Acknowledgement ........................................................................................................... v  
Abstract .......................................................................................................................... vii  
Introduction ..................................................................................................................... 1  
  Nature of knowledge and the role of educational evaluation ........................................... 1  
  The research problem and the aim of the study ............................................................... 3  
  Research approach .......................................................................................................... 4  
  Thesis overview ............................................................................................................... 5  
Chapter 1. Review of the Literature: Computerized Assessment .................................... 7  
  External factors of computerized testing environment affecting performance speed .......................................................................................................................... 7  
    Computerized mode of test delivery .............................................................................. 7  
    Power vs. speeded tests ................................................................................................. 10  
    Test time limitation ...................................................................................................... 11  
    Adaptive vs. non adaptive tests .................................................................................. 12  
    Difficulty of a test question .......................................................................................... 14  
    Test question design: formats and styles .................................................................. 17  
    Item pool enlargement ............................................................................................... 19  
    Question review ........................................................................................................... 19  
  Internal psychological and cognitive perception factors affecting performance speed ......................................................................................................................... 20  
    Individual differences in performance speed ............................................................... 20  
    Mental speed, load, fatigue and accuracy ..................................................................... 23  
    Motivation, arousal and emotional state ....................................................................... 26  
    Age factor .................................................................................................................... 27  
    Gender factor .............................................................................................................. 28  
    Verbal intelligence ....................................................................................................... 29  
    Visuo-spatial ability ...................................................................................................... 29  
    Mathematical ability ................................................................................................... 30  
    Concluding remarks ..................................................................................................... 30  
  Response time in computerized testing ......................................................................... 31  
  Conclusion ....................................................................................................................... 32  
Chapter 2. Methodology .................................................................................................. 35  
  Introduction ..................................................................................................................... 35  
  Sample ............................................................................................................................ 35  
  Ethical considerations .................................................................................................... 35  
  Development of the experiment ..................................................................................... 36  
  Materials used in data collection .................................................................................... 37  
  Data analysis .................................................................................................................. 38  
    Mean Question Response Time .................................................................................. 39  
    Correlation between the first presentation and total question time .......................... 40  
    Variability of Response Time between test takers .................................................... 41  
    Test takers’ total test accuracy and accumulated Response Times for the first presentation of the questions ................................................................. 42
List of Figures

Figure 1. Questions characterised by different levels of difficulty: IRT approach.....16
Figure 2. Speed-accuracy relationship. .................................................................25
Figure 3. Correlation between the means of the first presentation and total time. .....40
Figure 4. Correlation between the index of return and number of correct answers....49
Figure 5. Correlation between index of return and MQRT. .................................50
Figure 6. Comparing mean time for conversion by multiplication and division.......55
Figure 7. Conversion of measurement units with a prefix milli- : litres vs. metres....56
Figure 8. Accumulated time and accuracy. ..........................................................60
Figure 9. Correlation between mean response time and standard deviation on a question: between-test takers variability. ..........................................................61
Figure 10. Histogram of distribution of Standard Deviation of test takers’ Standardised and Normalised Response Time. .........................................................64
Figure 11. Pacing patterns. ..................................................................................65
Figure 12. Variation of SNRT vs. test accuracy. .....................................................67
Figure 13. Variation of SNRT vs. mean SNRT within a test taker..........................68
Figure 14. Summary of distribution of SNRT across all entries.........................72
Figure 15. Quasi-equal questions (Q.1190) about reflections. ............................86
Figure 16. Quasi-equal questions (Q.1200) about reflections. ............................86
Figure 17. Time used by test takers to complete a practice test. ..........................87

List of Tables

Table 1. Test question formats (from Haladyna, 1999).......................................17
Table 2. Test question styles (from Haladyna, 1999).........................................18
Table 3. Conversion from square metres into hectares. .......................................51
Table 4. Quasi-equal questions: difference in RT and index of return. .............53
Table 5. Quasi-equal questions: difference in accuracy. .....................................54
Table 6. Pacing pattern and accuracy.................................................................66
Table 7. Difference between MQRT for right and wrong answers for all questions..70
Table 8. Difference between mean SNRT for right and for wrong answers within a test taker.................................................................71
Table 9. Glossary and abbreviations. .................................................................81
Table 10. Equal questions with a smaller number of test takers..........................85
Table 11. Equal questions with longer RT. .........................................................85
Declaration of originality

This thesis does not contain material that has been accepted for any other degree in any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by any other person, except where due reference is given in the test.

Signature: 

Evgueni Gvozdenko
Acknowledgement

I would like to thank my supervisor Dianne Chambers for great support and guidance through all stages of the study. Dianne’s encouragement gave me a feeling of involvement in academic environment; her sharp logic and challenging questions gave my thesis the shape and content of an academic manuscript.

I greatly appreciate the help of other staff members of the Department of Mathematics and Science in Education: the Head of the Department, Ann McDougall, for assistance with discussing and focusing the topic and methodology of the study; the subject co-ordinator, Helen Chick, for enthusiastic support of the experimental part of the study that involved introduction of computerised testing in DSME; Anthony Jones, for patience and support in ethics clearance issues; Vicki Steinle, for valuable comments on the Results chapter; John Warner, for sharing his clear vision of mental processes; Nick Reynolds, for computer group discussion leadership; Stephen Goldstraw, for providing excellent conditions for studies in the postgraduate room. Advice on statistical analysis for this study was received from Dr Sue Finch of the Statistical Consulting Centre, the University of Melbourne.

This study would not be possible without financial assistance of the Australian Postgraduate Award.

It is my family that will celebrate the completion of the study the most. Encouragement and support from my dear wife Inna allowed me focusing on the research issues; web technology support by my son Yaroslav accelerated the preparation phase of the experiment, my home office network was maintained by my eldest son Ilya. Special thanks are to my father Nickolay who crossed an ocean to give me everyday encouragement and support to complete the study.
Abstract

This empirical study collected data from a computerized non-adaptive test in basic mathematics skills taken by 135 tertiary students. The study focused on the number of times questions were reviewed and the time it took test takers to attempt a question. The number of test takers’ returns to the same question (index of return) was found to be a construct that differs from the construct of question difficulty which is measured by the percentage of wrong answers. The index of return is suggested to be a measure of a question’s difficulty as perceived by the test taker and thus may be used as a new and additional tool in question calibration. It may also be useful for analysis of the learning process.

The analysis of test takers’ response time (RT) suggests that the means of RT can be used to verify the equality of questions as has been determined by an expert. The study introduced a new notion of quasi-equality of questions in regard to the questions that look similar, and has been judged as equal by an expert, but exhibit different time demands or elicit different behaviour of test takers. Additional investigation into these questions prompted by time measurements may reveal difference in cognitive demands and alter expert’s opinion of the equality of these questions.

The study observed a high variability of RT between test takers on a question and within a test taker across the test questions. In order to avoid a danger for slow performers to be discriminated against, the study suggests that the question selection algorithm should include a response time parameter to balance a set of test questions by expected time required for a question completion. Accelerating/slowing down patterns of performance speed were identified with the results supporting previous research. The study examined correlation of the patterns with test takers’ final scores and found statistically significant difference between two of the patterns. The finding that wrong answers take longer to complete that correct answers conform with prior research in this area. The study observed regularity regarding the difference in distribution of RT for right and wrong answers, and probable association of some ranges of standardized RT with a higher probability of a wrong answer.
Introduction

Nature of knowledge and the role of educational evaluation

Starting from Plato and Aristotle the philosophers were approaching the basic question: what distinguishes true (adequate) knowledge from false (inadequate) knowledge? It took hundreds of years to agree that knowledge is context-bound and therefore, inherently, has some degree of uncertainty ("relativism"). Following the Renaissance, two main epistemological positions dominated philosophy: empiricism, which sees knowledge as the product of sensory perception, and rationalism, which sees it as the product of rational reflection.

The constructivism proposed by Jean Piaget (1896-1980) introduced an idea of dynamic knowledge acquisition based on existing personal experience, rational thinking and ongoing adaptation. Constructivist Theory of Learning (Bruner, 1960, 1996) argued that the ability to solve problems through analytic thinking is best formed in the process of active interaction with the environment, that is, that sensory perception is a required element of building individual knowledge.

Constructivist Theory in general and Piaget’s works in particular are especially interesting to mathematics educators because mathematics is a study of structure, with learners developing a holistic picture of knowledge through building it with conceptual models of (i) elements; (ii) relations; (iii) operations; (iv) patterns (Lesh & Carmona, 2003). It provides a tool for understanding a complex process of learning, supporting an idea that learning is not an instantaneous event, it is not just flowing of information from one head to another, but a gradual process of active construction of a new knowledge on the foundation of existing knowledge and beliefs.
Determining the location of a learner on the stairs of the knowledge construction process is one of the main functions of educational assessment.

Educational evaluation has a dual purpose of assessing learners’ progress and supplying a feedback for education system modification. Such dualism is reflected in some confusion, which might arise in application of the common terms used by researchers and practitioners in this area.

The terms test, measurement, evaluation and assessment, though used in the literature sometimes interchangeably, convey distinct ideas and should be differentiated for the purpose of this research. Test, the narrowest of these four terms, connotes the presentation of a set of questions to be answered by a test taker. As a result of a test some numeric value is obtained to indicate the person’s knowledge. Measurement is a broader concept of obtaining information in a quantitative form using also additional options, for instance: observations, rating scales, questioning peers. Evaluation is a professional judgment about a set of quantitative and qualitative parameters (Mehrens & Lehmann, 1978). Assessment is a currently popular term, which can be defined as a process and a result of a comprehensive evaluation of a learner’s progress on the basis of standardized and computerized procedures. A new approach to assessment is based on an informal synthesis of a wide variety of evidence and is designed to evaluate student’s progress on a scale of achievements (Weiss & Schleisman, 1999), which shifts assessment to being more criterion-referenced than norm-referenced, that is, evaluating student’s progress relatively to determined subject mastery scale without considering the student’s achievement relatively to other students.

The introduction of technology into teaching and learning is a trend of recent decades and a result of “value for money” concerns that are increasingly driving
higher education. However, not everyone realizes that whereas the cost of teaching is more related to the hours per course and is not greatly affected by how many students actually attend a lecture, the cost of assessment is a cost per student which grows proportionally to student numbers. If to consider all costs of assessment including setting time, invigilation time, marking time, administration process, required facilities and resources at both faculty and university levels, the cost of assessment in higher education may now surpass the cost of teaching (Ricketts, Filmore, Lowry, & Wilks, 2003). It can be argued that teaching should also be understood in a broad terms as a process involving administration and facilities maintenance costs, and not only direct costs of tutoring hours.

The research problem and the aim of the study

Use of computer-based tests (CBT) has dramatically expanded during the last decade. It is now a common form of test delivery for licensure, certification and admission tests. The moving forces of CBT implementation are reduction of test administration costs, benefits of individualized testing and instant evaluation. Currently available computing power provides achievability for the test designs combining traditional CBT advantages with new features of diagnostic feedback and adaptive item selection employed in computerized adaptive tests (CATs).

Most of currently employed testing models base their scoring algorithm on a bi-modal result of a test taker’s response: right or wrong. The extent of knowledge quality or the strength of the response remains unaccounted for. It is suggested by Linn et al. (2002) that test takers’ behaviour in computerized tests could be a valuable source of information about estimated ability and cognitive strategies employed. Further research is needed to explore how learners’ behaviour and how Response Time (RT) in particular can inform the interactive process of item selection. The
recommendation for future research included such questions as whether or not RT as a
dependant measure of speed and pacing behaviour can increase the validity of an individual score. Schnipke and Scrams (2002) suggest that further research is required on the relationship between speed and accuracy and on RT as a predictor of finishing time. The investigation of these issues could contribute to the discussion about generalizable relationships between proficiency and RT, which could lead to introducing RT parameter into adaptive testing models (Wainer et al., 2000) where RT could confirm or alter the value of the pass/fail result of an item response.

As was stated by Rabbit (1996), “The findings that individual variance in speed of performance across qualitatively different tasks is systematic, rather then random, offers us new ways in which to analyse decision times in order to explore cognitive differences associated with general ability and age” (p. 85).

The study aims to scrutinize the performance speed of online test takers with a focus on the potential value of item RT measurements for test questions calibration.

**Research approach**

To investigate some of the many factors influencing test takers’ behaviour the study considered these questions:

1. Does the number of return visits to a question measure the same construct as the number of right answers for that question? Can the number of return visits be used to evaluate the equality of test questions?
2. Can equality of a group of questions be verified using average question response time?
3. What is the variability of response time between test takers? Are slow-speed performers in danger of being discriminated against?
4. What is the variability of response time within test takers? How does a test takers’ response time change across a test?
Is there any significant relationship between within-test taker’s variation of response time and his/her average performance speed or accuracy?

5. Can test taker’s response time be associated with probability of error? If delayed answers can predict an increase in probability of error, what would be the range of association?

**Thesis overview**

In the review of the literature (Chapter 1) historical progress and the current state of research in the field of computerized testing with regard to time aspect are reviewed to identify the key issues. Chapter 1 examines the literature about the impact of external and internal factors on test takers time patterns. External features of test administration include test delivery mode, testing time constraints, test design options, question style and content, and question selection process. Further, the chapter reviews research on internal, psycho-physiological factors that may impact the time spent by test takers on a question. The influence of cognitive factors and individual differences on temporal behaviour is examined to complete a comprehensive outlook of potential variables determining test takers’ response time.

The Methodology chapter (Chapter 2) describes the research design, methods of data collection and analysis, and a setting of the experimental part of the research.

Chapters 3 and 4 present the results, analysis and discussion of the data collected at the experimental phase. Chapter 3 presents the findings about applicability of measurement of question response time and “index of return” for validating test questions equality. Chapter 4 focuses on the implications of response time variability for test design and predictive power of response time measurement in regard to probability of a wrong answer. The findings of the study are compared to the previous research covered by Chapter 1 and 2.
Chapter 5 conveys the summary of the findings with concluding remarks, description of limitations and recommendations for further research.
Chapter 1. Review of the Literature: Computerized Assessment

Delivering computerized testing online is an emerging technology of educational assessment. This literature review attempts to identify potential external and internal factors that can effect the time it takes a test taker to answer a test question, the response time (RT). It will discuss the findings of prior research concerning the impact of online delivery mode; computer hardware; test time limitations; test scoring and composing procedures; and test question design on RT. The issues of extending test question pools will be analysed with respect to different scoring models, highlighting the importance of question equality for maintaining test fairness.

The second part of the chapter will focus on the research of internal psychological factors and individual differences that may influence the time patterns of test takers. The impact of state of anxiety, speed-accuracy trade-off, difference in cognitive abilities, gender and age are reviewed to build a list of potential factors to account for the variation in the performance speed of test takers.

External factors of computerized testing environment affecting performance speed

Computerized mode of test delivery

Computer-based delivery of educational assessment began in the late 1980s soon after the large-scale supply of personal computers in education began. This format of testing has expanded dramatically over the last two decades offering a number of advantages unavailable in paper-and-pen (P&P) versions of tests.
Substantial research about Computer-Based Testing (CBT), also called Computer Assisted Assessment (CAA), investigated the issues of test transition (translation) into an electronic format. A considerable amount of study has been done to establish if P&P and CBT assess the same constructs, and whether the test results obtained with CBT represent the same level of proficiency in the same cognitive task.

The results of these studies are rather controversial and warrant a cautious approach. Neuman and Baydoun (1998) found no difference across modes between P&P and computer-based clerical tests. Puhan and Boughton (2004) reported the absence of any substantial difference between CBT and P&P versions of teacher certification tests. In the field of testing in mathematics, Straetmans and Eggen (1998) examined and compared three different test administration procedures for making placement decisions in adult education: a P&P test, a computer-based non-adaptive test (CBT), and a computerized adaptive test (CAT). The results suggest that test performance was not differentially affected by the mode of administration.

On the contrary, McDonald (2002) suggested that individual differences should be recognised as having the potential to significantly affect the equivalence of computer- and paper-based assessment, and argued that CBT and P&P tests provide test takers with qualitatively different experiences, which may affect the statistical equivalence of scores.

The impact of using electronic media for perceiving information was investigated by Garland and Noyes (2004) who found that screen reading is slower and possibly less accurate than reading from paper. The differences in the way in which information was retrieved varied between the presentational formats and were time related. Garland and Noyes suggest that repeated exposure to accessing
computer-based information is necessary to equate knowledge perception and presentation with that achievable from hard copy alternatives.

In general, most studies agreed that the underlying constructs measured by both CBT and P&P versions are more comparable for non-adaptive tests than for adaptive tests (Neuman & Baydoun, 1998). Adaptive tests will be discussed later in this chapter.

Another common concern is that results of a computer-administered test may be confounded with other factors, such as levels of computer proficiency/literacy and computer anxiety. Computer literacy in this context is understood as a demonstrated ability to efficiently interact with testing software and hardware. Kim and Mclean (1994) reported that test anxiety and computer literacy were significantly related to test performance, while computer anxiety was not related to test performance, on an adaptive test in algebra administered to 208 South Korean college students maths. Lee and Weerakoon (2001) compared student performance in computer-based and P&P multiple-choice tests and found that neither computer experience nor anxiety correlated significantly with performance in the computer test. However, it can be argued that as this study had only 51 participants, it could hardly provide a basis for statistical significance.

Addressing this concern, Guidelines for the Developers of Computer-Based Test Services (American Psychological Association, 1986) suggested that the user interface should be designed to minimize reliance on computer skills that are not part of the construct being measured by the assessment. Test administrators should provide adequate assistance with regard to the computer skills required to complete a test. Test taker’s anxiety should also be reduced by providing tutorials and CBT sample tests to candidates in advance of the testing session to minimize examinee frustration. These
suggestions have been included into the recommendations of International Test Commission (2005). This topic is more fully addressed further later in this chapter while discussing the implication of psychological factors on a test taker’s temporal pattern.

Due to the technical nature of CBT its implementation is constrained by the reliability of hardware. Accidental computer failure can result in the complete loss of test information on a test taker’s or a test administrator’s machine. Recent advancements in computer hardware have made most concerns such as slow speed of a graphical display or insufficient hard drive capacity irrelevant. However, some innovations in test item design, like short segments of television-like episodes or involving speech recognition can still pose a challenge to hardware. Green (2000) suggests that uniformity of test hardware should include computer station, monitor and peripheral device equivalence. It can be added that operating system and the type of web browser can be crucial to test takers as unfamiliar controls can easily cause frustration and distraction. Garland and Noyes (2004) found that difference in cathode-ray tube (CRT) monitor characteristics of refresh rates, fluctuating luminance, and contrast levels were capable of causing cognitive interference and effecting the temporal pattern of a learner.

A natural limitation of web-based testing is that it relies on Internet and network connections. Any interruption in the services of an Internet provider can undermine the test results as test integrity and time restriction would be altered.

**Power vs. speeded tests**

Tests can be classified as ‘power’ tests and ‘speeded’ tests according to the time limits imposed on test takers. A power test aims to measure whether a test taker is able to answer the items correctly. It is assumed that for a power test all test takers
are given sufficient time to complete their answers without being influenced by time limits. In reality most tests are speeded to some extent, which can be measured by the percentage of unreached questions for non-adaptive tests. This extent is called speededness by Schnipke and Scrams (1999) who suggest that traditional speededness indices underestimate the amount of speededness on most multiple-choice tests because some test takers rapidly mark answers as time expires in the hopes of getting some of the items right by chance. This issue could be addressed by measuring question RT to eliminate “lucky guess” strategies (Schnipke, 1999).

**Test time limitation**

Time limits on tests may have at least three major functions. First, they are needed for organizational purposes to regulate the costs of test administration. Second, time limits might be applied in order to measure a necessary construct, for example, RT, when the tasks are so easy that they would be answered correctly by all participants if there was sufficient time. The third reason would be to provide standard conditions to all test takers.

Establishing fair time limits for Computerized Adaptive Tests (CAT) is not an easy issue as high-achievers are presented with progressively more difficult items with higher time demands. Sometimes it may be the case that an actual ceiling of test taker’s score is not achieved because of time constraints (Wise, 1997).

A good indication of the degree of speededness of a test can be the performance of test-takers on last 10% of test items. If a rapid guessing behaviour is clearly observed it may indicate that the time allocated to the test is not sufficient (Schnipke & Scrams, 1997).

How additional time would affect the score of a test taker has been investigated by Bridgeman (2004) on the materials of *Scholastic Assessment Test*
(SAT) and Graduate Record Examinations (GRE) tests, which are university admission and graduation tests and are designed to minimize the time effect on the score. The results of the tests where additional time was given to test takers suggest that allowing more time per question had a minimal impact on verbal scores, producing gains of less then 10 points on the 200-800 SAT scale, and that for the Maths sections high scoring students tended to benefit more than lower-scoring students with a difference of up to 30 points, which is still not very significant (Bridgeman & Cline, 2000). The study did not find any racial/ethnic and gender differences with extra time.

Adaptive vs. non adaptive tests

The advance of testing procedures after the transition of tests from paper-and-pen to computerized administration was the introduction of individualized test item selection and scoring. Computer-based tests are non-adaptive if they use a randomised draw of test items from a pool of test items and are adaptive if the test taker’s answer affects the selection of the item for the next question.

The objective of Computerized Adaptive Tests (CATs) is to select, for each examinee, a set of test questions from a pre-calibrated item bank that most effectively and efficiently measures the knowledge, skills or fluid ability of the individual. CATs offer a number of benefits compared to non-adaptive CBTs.

The most important of the advantages is that the test time can be significantly reduced as the selection process excludes from the test those items which exceed the test taker’s abilities. Another benefit is a reduced frustration experienced by test takers, as they do not have to struggle with questions that they cannot resolve (Lilley, Barker, & Britton, 2004).
CAT models can be differentiated by the underlying theories – Item Response Theory (IRT) or non-IRT based. IRT is a statistical framework in which mathematical models are used to process the results of actual performance on test items, pre-determined item statistics, and estimated examinee ability to predict the relative level of proficiency of the test taker. IRT suggests that all test items in the item pool should be assigned a value along a unidimensional scale of difficulty. During the test an examinee is first presented with a randomly chosen item. The selection of each following item is based on the response to the previous item. All items on the scale are equally important and the test stops when the position of examinee’s abilities on the scale is determined with sufficient probability. As was suggested by Wainer et al. (2000), so long as a single dominant dimension underlies the items, and so long as the items can be ordered in difficulty on that dimension, CAT is fair.

The idea of unidimensional testing is not generally accepted. Flaugher (2000) points out that in order for the adaptive test to be efficient, as many item pools need to be created as there are dimensions, which would be equal to creating separate subtests with separate scores derived from them and used as a weighted combination to compute a final test score. It can be argued that a general idea of knowledge in most subjects would be better characterized by multidimensional parameters. The task of assigning a linear value of complexity to each of several hundred items can be easily affected by individual bias unless some objective parameters are involved.

A problematic issue with IRT-based CATS is the algorithm offering a less difficult item if the person has failed the previous one. Neither omitting nor reviewing items are permitted in most CAT implementations as omitting items interrupts further item-selection process and reviewing items could confuse the item-selection path.
Another limitation of IRT is that item response is analysed at the pass/fail level. The information, carried by the exact choice in favour of one of false options, is disregarded despite the fact that it may convey a very important piece of data about the examinee’s misconceptions. Even if the same misconception leads to errors in other questions, it will not be detected by current models of CAT.

Wainer (2000) suggests that performance-assessment test items (that is, essay prompts or other higher level construction response items) are less likely to be used in a CAT, as they take a longer time and they are more expensive to pre-test on a large scale than multiple-choice items.

A non-IRT model for CAT was developed as a sequential testlet-based procedure which randomly selects each testlet (the subset of items) from a pool of parallel testlets and determines whether an examinee passes, fails or continues the test (Lewis & Sheehan, 1990). The testlets are composed of the same number of items and are supposed to be equivalent with respect to content coverage.

Adaptive testing is now at the initial stage of common practice. Further research needs to be done on issues of automated item calibration and incorporating supplementary factor of test taker’s behaviour analysis into item selection and scoring procedures.

**Difficulty of a test question**

Mehrens and Lehmann (1978) offered the following definition:

Average test difficulty is defined as the ratio between the average score and the total test score. For example, if on a test where the maximum score was 100, the average score was 70, we would say that the average test difficulty is $\frac{70}{100} = .70$. If all pupils who took the test received a perfect score, the average test difficulty would be $\frac{100}{100} =$
1.00; if no pupils answered any item correctly, the average test
difficulty is 0/100 = 0.00. (p. 326)

This definition seems to be counter-intuitive as defining the difficulty of a test or of a
question as the percentage of test takers who get the item right elicits some
confusion, as an increased value, for example from 60% to 80%, means that more test
takers have answered the test question correctly, which actually means that the test
was less difficult in a common sense meaning of the word. It is hard to see the reason
why the term “test item difficulty” should not convey the meaning of the percentage
of wrong answers, or should not be replaced with the term “test/item ease” as was
suggested by Gronlund (1981) and others.

In computerized adaptive testing the notion of “item difficulty” is central for
developing a test item pool. Item Response Theory (IRT) (Baker, 2001) postulates
that each item can be described by the relationship between probability of a right
answer, which is derived from “item difficulty”, and the ability of test takers. The test
item characteristic curve (Figure 1) is the basic block of the theory as all other
constructs depend upon this curve. Baker, 2001 defines difficulty as a location index
of the curve.

Figure 1 illustrates three item characteristic curves with the same
discrimination but different levels of difficulty. However, it can be argued that if the
probability of a correct answer (P) of a person with maximum ability level (3) is not
reaching its maximum it should be treated as a warning sign of a faulty item. A
perfectly designed test item should provide the highest probability of success to a test
taker with the highest ability.
Figure 1. Questions characterised by different levels of difficulty: IRT approach.

The need for modification of IRT-based procedures to include more objective parameters like question Response Time (Wainer, 2000), or information-time index (Lau & Wang, 2000) has been expressed by a number of researchers. Lau and Wang (2000) suggested that using information index could reduce by 12% item consumption and using Response Time for determining information-time index can enhance by 14% time consumption on an adaptive test. Item consumption is a critical issue for adaptive testing as the procedure of replacing an over-exposed item is the most labour-consuming part of the test administration. Over-exposure may be caused by an extensive use of the test item when the item has been presented to too many test takers or published for a wide audience, which resulted in compromising the novelty of the item. Decreased time consumption, that is, the time spent by a test taker to complete a test, is one of the advantages of adaptive testing allowing significant reduction in the total test time.

To summarise the issue of difficulty in IRT, it seems appropriate to quote Wainer et al. (2000): “That a single continuous variable in an IRT model accounts for all systematicity in examinees’ item responses is not proposed as a serious
representation of cognition, but as a caricature that may help solve some practical testing problems” (p. 236).

Test question design: formats and styles

Test questions are designed to fit the test purpose and the objectives of assessment. Test publishers offer a wide range of question templates or styles defining a form and content of the test presented to test takers. A set of question styles can be subdivided by question formats. Bridgeman and Cline (2000) found substantial variability of response time between different test question designs, which is a concern for test fairness for CATs. Haladyna (1999) suggested that there are three types of test items formats: high-inference constructed response (CR), multiple-choice (MC) and low-inference constructed response (CR). These formats differ in structure, objectivity of scoring results and limitations (Table 1).

Table 1. Test question formats (from Haladyna, 1999).

<table>
<thead>
<tr>
<th>High-inference constructed response (CR)</th>
<th>Multiple-choice (MC)</th>
<th>Low-inference constructed response (CR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus: Command or question</td>
<td>Stimulus: Command or question</td>
<td>Stimulus: Command or question</td>
</tr>
<tr>
<td>Conditions for performance</td>
<td>Student selection of choices</td>
<td>Conditions for performance</td>
</tr>
<tr>
<td>Student performance</td>
<td>Objective scoring</td>
<td>Student performance</td>
</tr>
<tr>
<td>Judge-mediated scoring using a descriptive rating scale</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Haladyna (1999) discusses the issue of preferable format in the framework of five validity arguments: prediction, content equivalence, proximity to criterion, gender bias and cognition. The essence of the prediction argument is that current test results are expected to have a high correlation with the previous and future performance of individuals. The content equivalence concerns interpretableness of test scores based on the definition of the construct, whether it is knowledge-based,
skill-based or a fluid ability. The proximity to criterion argument measures how strong the relationship between criterion measure and the criterion is. The gender bias argument cautions about possible effect of gender on test item perception and elicited cognitive process. Haladyna (1999) suggests that the most efficient and reliable way to measure knowledge is with multiple-choice formats and to measure skills or a fluid ability is with constructed response format. Table 2 presents question styles most commonly used in computerized testing.

Table 2. Test question styles (from Haladyna, 1999).

<table>
<thead>
<tr>
<th>Multiple-choice (MC)</th>
<th>Constructed response (CR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yes or No, True or False</td>
<td>Fill in the blank</td>
</tr>
<tr>
<td>2. Choose one correct</td>
<td>Short essay (key words)</td>
</tr>
<tr>
<td>3. Choose all that apply</td>
<td>Graph plotting</td>
</tr>
<tr>
<td>4. Match by a criterion</td>
<td>Insert section breaks in a text</td>
</tr>
<tr>
<td>5. Rank</td>
<td>Point and click on a map</td>
</tr>
</tbody>
</table>

Gadalla (1999) examined the equivalence of multiple-choice and constructed response (discrete) formats on the Canadian Achievement Test in mathematics administered to 1028 secondary school students. He found that response format had a significant effect on the performance of students in Grades 2 and 3 but was not significant in Grades 4, 5 and 6.

As far as one of the main aims of CBT is to achieve automated objective scoring, the usage of high-inference CR that deals with more abstract notions of creativity, style and expressiveness is rather limited (Christie, 2003). Even low-inference CR, such as a short answer essay, is very demanding on test makers due to the potential existence of right answers that have not been registered in a list of possible correct options and, therefore, they may be automatically marked as wrong.
For the purpose of this study, it is important to keep in mind that the individual behaviour of test takers may depend upon a particular question format and style as they may elicit different cognitive strategies and patterns.

**Item pool enlargement**

Content exposure increases with each test presentation and control over item over-exposure involves the issues of the efficiency of item administration and overall question pool enlargement. Creation of a new question in a non-adaptive test is rather straightforward and is well-defined in the classic testing approach (Mehrens & Lehmann, 1978). However, automatic generation of a new question and verification of question equality is still under discussion (Davey & Nering, 2002).

Randomisation of item selection and the ability of the item selection system to (i) automatically generate test variables from the defined range; (ii) line up groups of items in a pre-defined order and automatically draw a number of questions from a designated item pool can also reduce item exposure. Chapter 3 explores question equality disturbance that may be caused by automated item generation and suggests the procedure for validation of question equality.

**Question review**

Allowing test takers to review questions has been found to be highly desired by test takers and to be something very controversial for test makers. Advantages of allowing question reviewing are multiple: students feel more comfortable when they have some control over the test (Revuealta, Ximenez, & Olea, 2003), it significantly decreases anxiety level and test takers considerably improve their scores (Olea, Revuelta, Ximenez, & Abad, 2000), (Wise, 1997).

Opposition to reviewing option is mostly caused by concerns of adaptive test makers about the effect that reviewing items and changing responses might have on
the scoring algorithm (Weiss & Schleisman, 1999). In addition, the question selection process in adaptive testing is based on analysis of the previous response. If that response gets changed later along the test it confuses the question selection process and has a potential to unfairly inflate the test score (Stocking, 1997).

Stone and Lunz (1994) explored the effect of reviewing test questions on the efficiency of CAT administered to 712 medical students. They found that 32% of the students improved their score slightly after review, but did not change their pass/fail status. The average efficiency of the test was decreased only by 1% after review had been allowed. This led to the conclusion against prohibiting review of questions on CATs. However, the way in which reviews were allowed leaves many questions unanswered. First of all, the students were given a review option at the end of the test, which means that the item selection process was not affected by any response alteration. Second, it raises the question of whether a student would have got a different set of questions if he/she had reviewed previous answers in the middle of the test; and whether a student would have been able to respond more correctly to the other set of questions. It seems there is a need for further research in this area.

Internal psychological and cognitive perception factors affecting performance speed

Individual differences in performance speed

Starting with Galton (1961) psychology has accumulated a great deal of data about individual differences in cognition. Factor analysis introduced by Spearman (1863-1945) allowed mapping subgroups of abilities and areas of human cognition that were characterised by systematic variances of human performance across
qualitatively different tasks. The main criteria of performance – accuracy and speed – are found to be in an interdependent relationship and vary according to the nature of the task.

In The Theory of Multiple Intelligences Gardner (1983) suggests that the traditional notion of intelligence, based on I.Q. testing, is far too limited. Instead, he proposes eight different intelligences to account for a broader range of human potential in children and adults:

- Linguistic intelligence ("word smart")
- Logical-mathematical intelligence ("number/reasoning smart")
- Spatial intelligence ("picture smart")
- Bodily-Kinesthetic intelligence ("body smart")
- Musical intelligence ("music smart")
- Interpersonal intelligence ("people smart")
- Intrapersonal intelligence ("self smart")
- Naturalist intelligence ("nature smart")

A different approach to individual abilities within a study of human intelligence was suggested by Cattell (1998) who empirically identified cognitive areas as factors predicted in cognitive abilities theory, known as $G_f-G_c$ theory (Horn, 1994; Cattell, 1998). Later the list of factors was extended by Woodcock (1998) up to 10 broad abilities as follows:

- Short term memory – the ability to hold information in immediate awareness and then use it within a few seconds
- Verbal conceptual knowledge – breadth and depth of knowledge, including verbal communication, information and reasoning when using previously learned procedures
- Quantitative knowledge – the ability to manipulate numerical concepts and symbols
- Reading-writing – basic reading and writing skills
- Visiospatial thinking – operating visual images and spatial orientation
- Auditory thinking – operating audio signals
- Long-term storage-retrieval – efficiency of long term memory
- Automatic processing speed – the ability to rapidly perform automatic and very simple cognitive tasks
- Novel reasoning – the ability to reason, form concepts, and solve problems that often include unfamiliar information or procedures. Manifested in the re-organization, transformation, and extrapolation of information
- Correct decision speed – speediness in finding correct solution to problems of moderate difficulty. (p. 138)

In the situation when a test taker is presented with questions engaging the same set of abilities, his/her performance speed would be expected to keep a constant rate, which is suggested to be a dependent variable of decision speed and level of competence in that area (Rabbit, 1996). Any alteration to decision speed on another question would be accounted for by a change in the corresponding ability. In this case any variation of test question completion time between-test takers could be explained by individual differences in abilities and relative competence, provided that all other external and internal factors are matched.

Individual cognitive differences are found, in general, to be both inherited and cultivated by the environment (Kimura, 1999; Horn, 1998). The proportion of the influence of genetic and social factors is still under debate as it was in Galton’s times when he wrote (Galton, 1961): “I acknowledge freely the great power of education and social influences in developing the active powers of the mind, just as I acknowledge the effect of use in developing the muscles of a blacksmith’s arm, and no further” (p. 1).
Arguing with Galton’s understanding of abilities as being mostly inherited, Blewett (1961) concluded his studies of relative genetic determination of the various primary mental abilities with suggestion that very little, if any, of the individual differences in speed, as measured by the test, are due to inheritance.

**Mental speed, load, fatigue and accuracy**

Differences in test performance speed can be determined by some constructs of mental parameters of an individual, like basic mental speed or accuracy; and by a temporary affective state of a test taker, that is, by his/her current emotional state, motivation, fatigue experienced and overall mental load (Cohen, Sparling-Cohen, & O'Donnell, 1993). The interest in mental speed has been kindled by a search for theoretical explanations of intelligence. Jensen and Vernon (1986) came to the conclusion that individual differences in performance speed on relatively simple tasks are highly correlated to differences in basic speed of mental information processing.

The idea of global neurophysiological efficiency produced a number of different theories of adult cognitive maturation and aging.

Recent study by Roberts and Stankov (1999) focused upon the relationship between general intelligence and process measures derived from elementary cognitive tasks. The relationship between fluid intelligence as defined in Gf-Gc theory and processing speed was found dependant on task cognitive complexity. The researchers argued that the search for a basic process of intelligence by means of mental speed frameworks alone is most likely misguided. Other additional factors must be considered while investigating performance speed.

The impact of work load on information processing has a long history of research (Broadbent, 1958; Kahneman, 1973). The term “work load” means the amount of information perceived per time unit. The ability to process the
informational flow is characterized by the number of processing resources available or “processing capacity” of an individual. It is well established (Cohen et al., 1993) that the total individual capacity has an upper limit, which can vary within some range being dependent on other factors like motivation, fatigue and kind of material processed. The performance level on a task can also be increased by increasing the level of effort but only until the upper ceiling of the individual capacity is reached (Kahneman, 1973). Different approaches attribute such limitation to different factors: the channel model argues that this limitation is caused by a narrow entry channel, while the capacity resource model is based on performance-resource function (Cohen et al., 1993). Another broad discussion is focused on the question whether a human brain engages a parallel distributed or sequential information processing (Rumelhart, McClelland, & Group, 1986; Clark, 1989). Though the current study acknowledges the importance of these issues for investigating temporal behaviour of an individual, they are beyond the scope of the study.

Psychometric data support the idea of a direct relationship between time restrictions and rate of errors (Jensen, 1987). In many types of perceptual-motor tasks there is a trade-off between how fast a task can be performed and how many mistakes are made in performing the task. When asked to perform a task as quickly as possible, people will try to apply various strategies that may optimize speed, optimize accuracy, or combine the two. A general picture of speed-accuracy relationship can be seen in Figure 2, suggested by Jensen (1987) on the basis of Hick’s Paradigm.
Figure 2. Speed-accuracy relationship.

Adapted from Jensen (1987, p 171).

Jensen (1987) suggests the differentiation between the trade-off events within the subject and across the subjects. The diagram shows that though the subjects A, B and C have the same response time on a simple task (marked with a white circle), with the introduction of a more complex task subjects A, B and C will need different time frames to accomplish the task. Subject C needs the most time to cope with the task maintaining the minimum level of errors. If subject C is asked to do the more complicated task within the same time as a simple task, he will have the biggest number of errors. Subject A who is the fastest with a more complicated task will also make the least number of mistakes when asked to accelerate.

In regards to the testing conditions the data imply the following:

- the performance speed of test takers can be expected to vary subject to the task complexity;
- the factor of test speededness will affect test takers to different degrees;
- the error rate of the fastest test takers will be the least affected.

For this reason, comparing the performance of test takers cannot be done on the basis of speed or accuracy alone, but both values need to be known.
Motivation, arousal and emotional state

The relationship between cognitive process and emotional state has been researched for the last two hundred years. The James-Langer theory of emotion proposed that emotional feeling was a by-product of a person’s mental interpretation activity as a behavioural response to an external stimulus. It is interesting to note that emotion shared the label of a by-product with some other constructs of mental activity like attention and mental effort with a major difference being that, though the functional role of emotions in cognition remains unclear, the relationship with physiological responses like brain electric activity, heart rate, blood pressure and others has been long established (Cohen, 1993). On the contrary, attention and effort are broadly used by current psychological models of mental processes, though the problem of substantiation by neurological research remains.

A new model of the emotion-cognition relationship was introduced by Lewis, Sullivan and Michalson (1988) who suggested that emotion and cognition are neither separate nor independent processes. Instead, they are interwoven into each other as elements of continuous behaviour. Different emotions produce different patterns of emotional arousal, which in its turn affects general mental arousal.

Arousal theory, referred to as Yerkes-Dodson law, postulates an inverted-U relationship between the level of arousal and efficiency of performance of any task. The law states that a medium level of arousal is optimal for performance of a task. As arousal increases or decreases from that point, it results in the decline of the efficiency. A high level of arousal or “hyperarousal” is a key characteristic of anxiety and its negative impact on productivity is well established (Martindale, 1981). There are several typical kinds of anxiety that may be observed during computerized testing, the major of those are: computer anxiety, caused by the fear of the inability to operate
the device correctly and test anxiety, caused by the fear of failure. The opposite extreme to anxiety is under-arousal or “hypoarousal”, which can be caused by different reasons starting with the fatigue after a sleepless night and up to paralysing and inhibiting feeling of despair. Any of these factors can come into play unexpectedly and may dramatically change the pattern of the performance speed of an individual.

Another factor sought to account for the internal influences mediating stimuli and responses could be represented by the concept of “motivation”. Motivational states were considered by the first behavioural theorists to be a critical determinant of behaviour (Cohen, 1993). The level of motivation underlies arousal level and determines the persistency of a test taker thus providing stability of performance speed. The different degrees of motivation can be a serious compounding factor while making time comparison of a test taker’s behaviour on formative and summative tests.

**Age factor**

Other factors that may play an important role in producing individual differences are age and gender. Age-cognition relations can be interpreted in terms of a relation between age and speed (Salthouse, 1992). The negative correlation of age and performance speed has been firmly established (Salthouse, 1992), however, there is a broad discussion about the cause of the change and whether the correlation can be accounted for by one general factor or by multiple factors corresponding to different cognitive abilities. The major neurophysiologic hypotheses are that increased age causes a slower speed of neural transmission, a lower attenuation of signal on inter-neuron level and a delayed neuron activation. Significant change is observed when the age difference exceeds 10 years (Craik & Salthouse, 1992).
Horn (1998), using factor analysis of the relationship between cognitive abilities and age, suggests that different intelligences stem from different combinations of genetic and environmental determinants and that cognitive speed factors are not general in the way they relate to other variables.

The impact of age on performance speed would differ for different tasks, depending whether it is a domain of fluid intelligence ($G_f$) or crystallized intelligence ($G_c$) suggested by the Triadic Theory (Cattell, 1998). Fluid intelligence, termed ($G_f$) showed its highest value in performing complex analysis where there was no advantage from previous knowledge. After reaching its maximum between the age of 16 – 21, it steadily declines throughout adulthood. On the contrary, crystallized intelligence ($G_c$), which is found on cultural performances and the use of what one has learned and stored up, slowly improves throughout active life (Cattell, 1998).

Based on the research literature it can be inferred that the age parameter should be accounted for in data processing if participants have significant differences in age.

**Gender factor**

Gender factor research has always been rather vulnerable to social prejudice and theoretical limitations. Any empirical finding of gender difference in favour of females could be hardly accepted by a mostly male political elite, the data in favour of males could be interpreted as a claim for male superiority and thus be publicly condemned by the feminist movement. Such an environment had its own advantage of creating the highest standards of scientific responsibility and accuracy in this area of research, which produced a rather objective picture of a complex relationship between gender and abilities. Starting with a comprehensive review by Maccoby and Jacklin (1974) a rapid growth in this field allowed building a solid knowledge of gender
differences in human cognitive performance. The most important conclusion of the latest research (Kimura, 1999; Cornoldi, 2003) is that the general categories of abilities, such as visual-spatial and mathematical, have not provided a clear-cut divide between genders and so need a more detailed and refined approach.

**Verbal intelligence**

The traditional impression is that women are more talkative than men. The research found that young girls articulate and enlarge their vocabulary, and start using a more complex grammar earlier than boys (Maccoby, 1966).

Though tests of verbal intelligence at mature age have not produced any significant difference for women, a more detailed investigation confirmed a constant female superiority in some particular tasks involving verbal memory recall. On average, women give a more accurate account of the content of memorised material, whether it is a list of unrelated words or a logically meaningful text. Women tend to show even better results on a visual memory tasks if it involves names, colours and faces.

In general, female advantage on verbal memory is one of the strongest sex differences favouring women (Kimura, 1999).

**Visuo-spatial ability**

Gender difference in visuo-spatial tasks seems to be concentrated around a selective group of functions of mental rotation and manipulation. The tasks that display the most distinct difference require active working memory processes and involve a specific visuo-spatial modality (Cornoldi, 2003). If the tasks entail less active operations like simple image recall from immediate working memory, no gender difference is reported.
Mathematical ability

Mathematical ability remains one of the most controversial issue in the area of sex differences. On one hand females get higher grades throughout their school years, on the other hand males constantly score higher on high school and college entrance mathematics aptitude tests with a difference of about one-half a standard deviation. The maths ability, however, is not homogeneous: girls tend to get higher scores on calculation /computation parts of the tests and boys perform better on the parts requiring mathematical problem solving; males have an advantage in spatial 3-dimensional rotational ability, which directly affects such maths areas as reflection, rotation and symmetry (Kimura, 1999).

In a comparative study of 13-year-old schoolchildren in the USA and Thailand Engelhard (1990) suggests that in both cultures boys perform better as the level of cognitive complexity increases and the content changes from arithmetic through algebra to geometry.

Concluding remarks

The analysis of the prior research in this area suggests that the particular characteristics of a task can benefit one gender or another. Males are found to have an advantage in mental rotation, reflection or other manipulations in visuo-spatial aspects; and problem solving in reasoning and logical manipulations, and females outperform males in all tasks involving verbal actions. However the origin of this difference (nature or nurture) is widely debated. The difference was reported to be dependant on genetic factors and individual hormonal balance on one side, and education and social environment on the other. Kimura (1999) demonstrated that both biological and experimental factors should account for the observed difference. Furthermore, this issue has a dynamic nature as some researchers reported that gender
difference is gradually diminishing over the last 40 years. We cannot exclude the chance that in ten years the picture of gender differences will differ dramatically from what we know now.

**Response time in computerized testing**

Computerized delivery of tests allowed new unobtrusive data gathering techniques. A number of studies (Wainer et al., 2000; Schnipke & Scrams, 1999b) have suggested using question response time to collect additional information about the impact of a question on a test taker on one hand and about the approach of a test taker to the question on the other hand.

A number of studies have investigated the issues of applicability of response time for improving the efficiency of cognitive tests, for example, Schnipke and Pashley (1997); Schnipke and Scrams (1997, 1999a, 1999b, 2002), Hornke (1997, 2000) and Bergstrom, Gershon, and Lunz (1994).

Bergstrom *et al.* (1994) collected data from a certification examination in 1991 using a computerized adaptive algorithm. They found that examinees spent more time on items that produced wrong results than on items where they reached correct answers. Test score was not correlated with response time; and the age, sex, first language and ethnicity of examinees did not predict between-person variance. This study raised a concern that most of response time variation had not yet been explained.

Hornke (1997) presented results from 100 test takers who had sat a computerized adaptive test on general mental ability. He examined the variance in test takers’ response time and found a bigger variance of response time for correct answers than for incorrect answers and some correlation between achievement and speed of response. There was no substantial correlation between response time
variance and achievement found. Hornke (1997) reported that wrong answers took examinees more than 20 seconds longer on average than right answers.

Rammsayer (1999) reported that his results on 120 subjects, who were confronted with perceptual and cognitive discrimination tasks, confirmed that response times were significantly longer for incorrect answers on perceptual and cognitive discrimination tasks. He suggested that the findings support the assumption that timing behaviour may represent an independent personality trait.

**Conclusion**

The reviewed studies outline factors that may affect test takers’ response time and potentially alter the test score. Some of the factors are external to a test taker, such as test delivery components (computer equipment, internet access), or test administration features (test design, question format, scoring model). Other factors are internal, such as psychological parameters (mental speed and accuracy, emotional state, difference in cognitive abilities) or age and gender.

Though it seems fair to believe that most of the factors affecting individual test performance speed have been accounted for, it is not clear how great the total variance in performance speed will be and which factors will be responsible for the biggest part of variation.

Reports on correlation between performance speed and test scores are controversial. The potential impact of response time variance on fairness of a computerized test in mathematics has not been examined in prior studies. Further research needed to establish if wrong answers take longer than right answers on a mathematics test.

The adoption of computerized testing by large-scale admission tests has greatly increased the need for item development to compensate item exposure.
(Wainer & Eignor, 2000). Creating parallel (equal) items to extend a test item pool is a major solution to this problem. The current study investigates the issue of validating questions equality using measurements of response times.
Chapter 2. Methodology

Introduction

This study has a quantitative nature and utilizes the quasi-experiment type of design, justified by the importance of studying cognitive processes in a natural setting (Cook & Campbell, 1979). The study investigates the potential of using easily-obtainable and unobtrusively measured data about the time aspect of the test taker-computer interaction to gain new perspectives on test question design and on test takers’ behaviour.

Sample

Quantitative data of test takers’ Response Time (RT) was collected from second year tertiary students, mostly females between the ages of 18 and 25, in Australia. The population consisted of 228 students who selected themselves into one of 13 groups to sit an examination in mathematics. Logistic demands of test administration determined the maximum size of a group to be 30 students and scheduling the groups across five days. Due to some technical problems the first four groups (out of 13) could not be included in data collection. Therefore, the data for the study was gathered from the other nine groups that consisted of 139 students.

Ethical considerations

The test was a part of an existing university subject and thus it was ensured that students were motivated to complete all questions. It was the first year when this test was administered electronically. As a step in transition from paper-and-pen to computerized mode of delivery a test training website was made available two weeks before the test to help students alleviate test and computer anxiety. Developing practice tests that provide test-takers the opportunity to familiarise themselves with
testing procedure and design is a recommendation of International Test Commission (2005). Students were allowed unlimited access, both on campus and from home, to practice tests. Feedback comprising correct answers and a total score was automatically generated immediately following the completion of a practice test.

The practice test website offered two options: anonymous testing and non-anonymous, time measured testing. In private correspondence some students expressed opinion that this site structure allowed a gentle introduction to computerized test format: they started with the anonymous self-assessment mode as they felt more secure when there was no identification used. Then, when they had gained more confidence, they moved to a time measured non-anonymous version, which gave them the feeling of doing a real test. The factor of computer anxiety was assumed to be low as all students had taken a computer skills subject in the previous year.

The study did not gather data from anonymous tests and it did not focus on the impact of practising on final test scores. Pre-test conditioning as a result of practising would be important in comparing performance between test takers. The present study focused on analysis of response time variation within an individual, which was assumed to be independent of the amount of pre-test practise-

### Development of the experiment

The test was administered under supervised and time-restricted conditions in a computer laboratory. The test duration of 55 minutes allowed all students sufficient time for completion. The test can be considered a power test, rather than a speeded test (please, refer to the definition on p.10), as all items could be answered without the influence of a speed factor. This is also supported by the fact that over 97% of
students were able to complete a practice test within 47 minutes (please, refer to the histogram in Appendix 5).

To further eliminate the difference between the computerized and paper-and-pen version of the test, the option of browsing between the questions was allowed. Test takers were able to change their answers using the option of returning to previously answered question before final submission of the test.

The test was administered in an on-campus computer laboratory in online mode. Identical hardware and equal speed of broadband connection provided equal technological conditions to all participants.

**Materials used in data collection**

The online version of the test was based on TestPilot software (McGraw-Hill Test Pilot Enterprise (v4), http://www.clearlearning.com). Additional server-side program registered timestamps 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 a particular question – a measurement technique employed by Bergstrom et al. (1994).

The test takers were presented with a non-adaptive test that used random drawing of the question out of sub-groups containing homogeneous (by the expert opinion) question pool. The test comprised 26 questions that were drawn out of 24 subgroups containing 72 questions, which had been pre-calibrated and assigned to one of the subgroups manually by an expert to minimize possible difference in difficulty.
Data analysis

The accumulated data contained the following information:

- The test taker’s name and login (at a very early stage of information processing the data were de-identified).
- The time of starting the test (as per the server time).
- The duration of the test (as per the server registration).
- The time on a test taker’s computer when a browser started downloading each web page with a test question. These timestamps allowed computing the time spent by a test taker on a question.
- The test question pool.
- The list of the questions presented to each test taker and his/her answers (marked as correct/incorrect).

For the purpose of this study the time between downloads of consecutive web pages by a test taker is referred to as Response Time (RT). This measure includes the time spent by a test taker on reading the task, finding the solution and filling in the answer into the form on the web page and the web page loading time. The latter had a value of less than 0.5 seconds for questions without graphics and less than one second for questions with graphics. The web page loading time was equal for all computers of the network of the computer laboratory where the test was administered, and thus did not affect individual differences in response time.

The set of RT was examined to adjust the data against unreasonably quick answers. All RTs of less then three seconds were considered as indication that a test taker was passing this question in order to reach another one. This information is not relevant to this research and these entries were removed.

As question reviewing was permitted, the data had to be separated into subgroups according to the number of approaches of the test taker to each question.
For the purpose of this study the time of the first approach and the accumulated time for all approaches to a question were calculated for each test taker.

The first presentation of the question was considered valid if a test taker spent at least 7 seconds on the question. This number is justified by the minimum time required for comprehending and answering the question. The entries of five participants were identified as clearly a “fly-through” strategy, that is, when test takers used the first item presentation to just look through and to move to another item. The data from these participants were excluded from the study.

The research focused on two parameters: (i) time spent on a question; and (ii) correctness of the response (accuracy). The information about accuracy was represented by a data set of final entries by a test taker. Correct answers were placed in each spreadsheet cell to contrast with the student’s answer. The summary table of the data held the time taken by a test taker on a question plus a colour code for accuracy. This allowed analysing the time data separately for correct and incorrect answers.

**Mean Question Response Time**

For the purpose of analysing the average time allocated by test takers to a test question (MQRT) the study used raw time data that were not logarithmically transformed or standardized.

The time spent by a test taker when attempting a question the first time was used as indication of a test taker’s time allocation pattern. To justify the approach the study needed to establish if the first presentation time was a valid representation of the total time spent by test takers on a question.
The means for the first presentation of each question and for all presentations of each question (total time) were calculated separately to examine any correlation between the means of the first presentation time and the means of total question time.

**Correlation between the first presentation and total question time**

The scatter plot (Figure 3) demonstrates the correlation between the mean total time that test takers spent on a question during the test and the mean time that test takers spent on a question when it was presented the first time. It can be seen that the first presentation time is highly correlated (Pearson correlation = 0.99) with the total time. For simple questions, characterised by short RT, the total MQRT does not differ much from MQRT for the first presentation. Test takers spent little time reviewing these simple questions. For more complex questions the reviewing time grew proportionally in a linear manner (Fig. 3). The data confirm that MQRT of the first presentation is a valid representation of the total time on a question and it will be used for the rest of the study in data processing and analysis as a mean of question response time.

*Figure 3. Correlation between the means of the first presentation and total time.*
Prior to the test all test questions had been rated by an expert and grouped into subgroups according to similarity. Questions within the same subgroup were evaluated by the expert to be equally difficult for test takers. The research analysed the collected data to find if expert opinion about equality of test questions was supported by this data on test takers’ behaviour.

Analysis of the number of returns to the same questions (index of return) has been introduced by this study as a potential source of information about test takers’ behaviour and test question functioning. In order to establish if index of return represents a new construct the study examined its correlation with the number of correct answers (index of accuracy) and Mean Question Response Time (MQRT). If more difficult questions are found to attract significantly more students returning to them later to re-check the solution then the index of return would be a dependant variable of index of accuracy. If more time-consuming questions are found to attract significantly more students to return to them later to re-check the solution then the index of return would be a dependant variable of MQRT.

If there is no or only weak correlation between these variables then the index of return should be considered an independent construct and it may provide an insight into unexpected qualities of a test question or point out at new regularities in test takers’ decision making processes. Chapter 3 presents Pearson correlation coefficients and scatter plots demonstrating the relationship between the index of return and these two variables.

**Variability of Response Time between test takers**

The variability of RT between different test takers can be expressed with a standard deviation (SD) value, which represents a measure of the range of values in a set of numbers. This statistic was used as a measure of the dispersion in a distribution
to examine the extent of difference between slow performers and fast performers to establish if slow performers may be in danger of being discriminated against under test time limits.

*Test takers’ total test accuracy and accumulated Response Times for the first presentation of the questions*

In this instance the accumulated time of the first presentation is the time it took a test taker to complete the first attempt at all questions. This time was then correlated with the per cent of correct answers of the test taker to determine if fast test takers tended to produce better results. The Pearson correlation was calculated to show the relationship. The relationship between standard deviation and MQRT is demonstrated on a scatter plot.

*Within-test takers variability of Response Time*

The normal distribution is one of the key distributions providing the basis for probability statistics. When a particular distribution is not normal some transformation of the data may be attempted so as to achieve a normal distribution, for example charting the logarithms of values instead of the values themselves. This is known as normalizing the distribution.

As initial analysis of the data distribution histogram showed that the data were considerably positively skewed the advice from Statistical Consulting Centre of the University of Melbourne was sought; and the decision was made to normalize the data with natural logarithmic transformation to create a more normal distribution required by most statistical procedures. This procedure of normalising and standardising the data collected from measuring response time is generally accepted in prior research (Schnipke, 1999a).
To compare the performance speed of a test taker across different questions, his/her question RT for each question can be expressed with reference to the average time of all test takers for that question. This reference can be established with “a standard score” – the Standardized and Normalised Response Time (SNRT), which is calculated using the following formula:

\[(T_{\text{ind}} - T_{\text{avg}}) / SD\]

Where \(T_{\text{ind}}\) is the time spent by a test taker on the question, \(T_{\text{avg}}\) is the average time spent by all test takers on the same question, and \(SD\) is a standard deviation for individual times on the question.

If a test taker’s time was equal to the average time (for all test takers) on that question, the test taker’s Standardised and Normalized Response Time (SNRT) would be 0.00. If the test taker is faster than average, the SNRT would be positive. If the test taker is slower then average, the SNRT would be negative. A SNRT equal to -1.00 would mean that the test taker’s time is longer than the average time for a value of one standard deviation.

**Analysing pacing patterns**

For each test taker the means of his/her SNRTs were calculated separately for the first (9 questions), middle (8 questions) and end (9 questions) groups of questions to determine if the performance speed was changing during the test. Then the calculated values were compared to determine if they fitted into one of four possible models:

1. The performance speed is accelerating towards the end of the test.
2. The performance speed is slowing down towards the end of the test.
3. The performance speed is slowing down in the middle and then re-accelerating towards the end of the test.
4. The performance speed is accelerating in the middle and then slows down towards the end of the test.

Correlations between variation and accuracy, and between variation and average standard score of a test taker

The variability of response time within a test taker can also be examined using a standard deviation (SD) value, which in this case will indicate the range of variation of the SNRT of a test taker across different questions.

The variation of SNRT, expressed by SD value, was correlated with total test accuracy, expressed by a per cent of correct answers, to establish if the test takers who achieve a high score also have a more stable performance speed. Pearson correlation was calculated to show the relationship.

A relation between the variation of SNRT and average SNRT of a test taker was examined to establish if there is any correlation between performance speed and performance stability. Do generally faster test takers have less variation in their speed? Pearson correlation parameter and a scatter plot demonstrate the finding in Chapter 4.

Relationship between test takers’ Response Time and probability of error

To answer the question: “Can individual RT predict the probability of a wrong answer?” the data were analysed separately for right and wrong answers to determine if there are any regularities in distribution of MQRT and SNRT for right and wrong answers.

After the data were normalized with natural logarithmic transformation the means for right and wrong answers across the questions were compared. To test the hypothesis that wrong answers take longer then right answers it is required to determine the weight of evidence for rejecting the null hypothesis. This can be done
by calculating the level of significance (P-value) of the statistical test ‘*t-test paired two samples for means*’. It will show how likely the difference in means can be the result of a sampling error. For the purpose of the research the probability level used in calculations was 0.05 (5%), which allows us to say that the results are significant at the .05 level. If the absolute value of $t_{\text{Stat}}$ is larger than $t_{\text{Critical two-tail}}$ we will have to reject the null hypothesis of no difference in favour of the alternative hypothesis that there is a statistically significant difference between RTs for right and wrong answers.

To compare within-test takers’ results across different questions each logarithmically transformed value was converted into a standard score - SNRT. It reflected the relationship of an individual time to the average time for the question in terms of standard deviation and thus it allowed compensating the difference in question complexity across test questions, as it was suggested by Schnipke (1995).

After separate means for individual right and wrong answers were computed, a null hypothesis stating that there is no significant difference between individual average standard score for right and wrong answers was tested. Again, this was done by calculating the level of significance (P-value) of the statistical test ‘*t-test paired two samples for means*’. The probability level used in calculations was 0.05 (5%), which allows us to say that the results are significant at the .05 level. If the absolute value of $t_{\text{Stat}}$ is larger then $t_{\text{Critical two-tail}}$ the null hypothesis has to be rejected.

The *t-test* statistic and the distribution chart of all standard scores of all test takers distinguished by the parameter of correctness will be demonstrated in Chapter 4. The results may provide a general picture whether the delay in answering the question can be a predictor of error probability.
Chapter 3. Results and Discussion: Validating Questions Equality

This chapter presents the results of data analysis that investigate the following research questions:

- **Index of return**
  Does the number of return visits to a question (the index of return) measure the same construct as the number of right answers for that question?
  Do more time consuming questions have a higher index of return?

- **Equality of questions**
  Can an expert’s opinion about the equality of a group of questions be validated using the average question response time?

Norm-referenced testing interprets a score of an individual by comparing that score with those of other individuals called a norm group. The validity of the comparison is based on the assumption that the norm group is appropriate and that the questions administered to the individual and the norm group are equal (Mehrens & Lehmann, 1978). Criterion-referenced testing determines how well a test taker has mastered a cognitive skill by comparing a score of the individual with the maximum score that could be achieved. Its validity is based on the assumption that the question presented to a test taker is measuring the same constructs as the reference question.

Growing usage of randomly selected or computer-generated question variables raises concerns about the equality of tests. Classic testing theory and most computerized testing programs use the percentage of correct answers as a question parameter called “question difficulty” to monitor question functioning, that is the response the question generates in test takers. Therefore, if two questions testing the same cognitive skill generate the same value of question difficulty these questions are
considered to be equal and thus can be used interchangeably by the question selection algorithm.

A number of studies (Wainer et al., 2000; Schnipke & Scrams, 1999b) have suggested using question response time to collect additional information about the impact of a question on a test taker. Computerized delivery of tests allows the introduction of new unobtrusive data gathering techniques to provide additional tools for validating question equality. This study focuses on measuring response time and collecting statistics on the number of visits to a question that may assist test designers in the process of item calibration, which is one of the most labour-consuming parts of adaptive test design involving human expert opinion and pilot trials.

**Index of return**

The total time spent on a question may differ from the time of first presentation (examined in Chapter 2: Analysis) for three probable reasons: (1) test takers may go through all questions a second time to verify that they did not miss an answer; (2) test takers may re-do a task to check if the answer is correct; (3) some test takers may attempt questions they have not been able to solve at the first presentation.

For the purpose of this study the data about the first option were excluded from the number of visits as they were too short to effectively review the answer. Only meaningful reviews (longer then three sec) were used to calculate the average number of return visits to a question (the index of return).

It was hypothesized that the index of return can measure a construct different from that measured by RT or the index of difficulty (accuracy). Schnipke and Scrams (1997), using Thissen’s Timed-Testing model as a framework for their research, examined the data from CBTs of verbal, quantitative and reasoning skills involving 7,000 examinees. The results of his study suggest that response time and accuracy
provide separate measures of performance. If the index of return displays sufficient independence from these factors and consistency for equal questions, then it can be offered as an additional source of information about test takers’ behaviour that can be used by test designers to monitor test question functioning.

It may be hypothesized that while the percentage of correct answers of a question reflects general difficulty of the question, the index of return may indicate the question difficulty as it is perceived by test takers, that is a recognized difficulty. The study investigated a correlation between the number of correct responses and the index of return for a question. Figure 4 demonstrates the relationship between these two variables.

Figure 4. Correlation between the index of return and number of correct answers.

As can be seen, there is a slight tendency for test takers to return more to the questions that produce a lower proportion of right answers. This relationship is expressed by a weak negative correlation (Pearson correlation = -0.48), meaning that a bigger average number of returns to a question is moderately associated with a higher per cent of incorrect responses.
Additionally, the index of return also has a weak positive correlation (Pearson correlation = 0.61) with MQRT as demonstrated in Figure 5. This means that students only slightly tend to return more to the questions that need more time for solution.

*Figure 5. Correlation between index of return and MQRT.*

The observed weak correlation of the index of return with both accuracy and MQRT and the absence of correlation between accuracy and MQRT (Pearson correlation = -0.33) suggest that the index of return is a separate construct measuring a different trait of test taker’s behaviour that has only partial relationship with average time spent on a question and percentage of correct answers.

The relationship between index of return and test taker’s certainty in the correctness of the answer can be a subject of further investigation. This hypothetical relationship may have the potential to indicate the quality of test taker’s knowledge and thus may be able to inform the CBT scoring algorithm to adjust the weighting of test takers’ answers.
Index of return is also suggested to be used as a test designing tool as it may supply additional information about test question functioning. Application of index of return in the investigation of question equality will be demonstrated further in this chapter.

Equal and quasi-equal questions

The study suggests that the MQRT can verify an expert opinion of question equality and can provide a reliable measure of differences in complexity of test questions. To approach this issue the study needed to establish whether possible variance of MQRT does not compromise its applicability for test design process.

Several groups of questions were rated by an expert as “absolutely equal” and they were used as probes of MQRT variation. Questions 210 and 530 test the skills of conversion from square metres into hectares, a procedure that involves multiplying or dividing by powers of 10 (Table 3).

Table 3. Conversion from square metres into hectares.

<table>
<thead>
<tr>
<th>Groups of question</th>
<th>Hectares</th>
<th>Hectares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions 210</td>
<td>210</td>
<td>530</td>
</tr>
<tr>
<td>Task</td>
<td>12560 m² = ? (ha)</td>
<td>9570 m² = ? (ha)</td>
</tr>
<tr>
<td>number of test takers</td>
<td>75</td>
<td>60</td>
</tr>
<tr>
<td>% of right answers</td>
<td>77%</td>
<td>78%</td>
</tr>
<tr>
<td>Mean of 1st presentation (in sec)</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>Mean of the total time (in sec)</td>
<td>63</td>
<td>58</td>
</tr>
<tr>
<td>index of return</td>
<td>1.61</td>
<td>1.67</td>
</tr>
</tbody>
</table>

Each of 135 test takers was presented one of these questions, randomly drawn from the test pool. As can be seen from the table, these questions yielded similar means of the first presentation time (44 sec and 43 sec), with the difference less than 1.3%. Both questions produced a similar rate of accuracy, means for the total time, and indexes of return. These data fully support the equality of these questions.
The validity of the measurement was not affected by a smaller number of participants (22-30 test takers for questions 70 and 180, please refer to Appendix, Table 10) or a longer average time (162-172 sec for questions 270, 630 and 640, please refer to Appendix, Table 11). The data fits with suggestions by (Bridgeman & Cline, 2000) about requirement to a mean of response time to be based on the performance of at least 20 students.

In the cases mentioned above, RT proved to be a good measurement tool for validating the equality of questions and was at least as good, or even better, than the percentage of correct answers (accuracy), which is usually used as a tool for testing question difficulty, and which offered less precision than RT with up to 10% tolerance for equal items.

While investigating whether question equality could be validated using RT, it was found that there were the questions, that had been categorised by experts as “equal”, that were found not to be equal when the RT measurement was employed. These questions were found to greatly resemble each other but to differ in respect of question time demand, which may indicate difference in some cognitive demands. The expert has classified these questions as equal or “nearly equal” and only after additional investigation prompted by RT and the index of return measurements the difference was identified.

For the purpose of this research, questions that greatly resemble each other but turn out to be measuring different cognitive skills or eliciting different temporal behaviour will be called quasi-equal questions.

A good example of this issue is the pair of questions 950 and 960 (Table 4), which were questions in the same question pool in the computerized test. These questions have identical wording except for the numbers. In question 950 test takers
manipulated numbers such as 75 cm and 2.4 m, while in question 960 they had numbers 80 cm and 5 m, which are easier to multiply (electronic calculators were not permitted on the test).

Table 4. Quasi-equal questions: difference in RT and index of return.

<table>
<thead>
<tr>
<th>Groups of question</th>
<th>window sq</th>
<th>window sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions</td>
<td>950</td>
<td>960</td>
</tr>
<tr>
<td>Task</td>
<td>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 5 rectangular windows, 75 cm wide and 2.4 m high. How long will it take her (please, round your answer in minutes)?</td>
<td>Izabella is washing windows. It takes her 2 min to wash and polish 1 square metre area of glass. The room she is cleaning has 6 rectangular windows, 80 cm wide and 5 m high. How long will it take her (please, round your answer in minutes)?</td>
</tr>
<tr>
<td>Number of test takers</td>
<td>54</td>
<td>78</td>
</tr>
<tr>
<td>% of right answers (accuracy)</td>
<td>69%</td>
<td>79%</td>
</tr>
<tr>
<td>Mean for 1st presentation (in sec)</td>
<td>193</td>
<td>155</td>
</tr>
<tr>
<td>Mean for the total time (in sec)</td>
<td>319</td>
<td>230</td>
</tr>
<tr>
<td>index of return</td>
<td>2.32</td>
<td>2.00</td>
</tr>
</tbody>
</table>

As can be seen from the table, question 960 resulted in a higher percentage of correct responses and smaller RT and index of return than question 950. Despite all apparent similarity of these two questions they are characterised by different time demands, which is indicated by RT, and difficulty level, which is indicated by accuracy and the index of return, and so they should not be regarded as equal questions, but rather as quasi-equal. Additional considerations by experts would be required to decide whether these questions should be placed in the same test item pool. Identifying questions such as these, using RT and the index of return is likely to be beneficial when developing item pools of equal questions for use in computerized testing.
There are some cases when accuracy (percentage of right answers) for two questions does not match as well as the MQRT, for example in the questions 1190 and 1200 involving reflections (Table 5).

<table>
<thead>
<tr>
<th>Groups of question</th>
<th>Reflections</th>
<th>Reflections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions</td>
<td>1190</td>
<td>1200</td>
</tr>
<tr>
<td>Task</td>
<td>(see Appendix, Fig. 15)</td>
<td>(see Appendix, Fig. 16)</td>
</tr>
<tr>
<td>Number of test takers</td>
<td>135</td>
<td>135</td>
</tr>
<tr>
<td>% of right answers</td>
<td>67%</td>
<td>92%</td>
</tr>
<tr>
<td>(accuracy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean for 1st presentation (MQRT)</td>
<td>54</td>
<td>58</td>
</tr>
<tr>
<td>(in sec)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean for the total time (in sec)</td>
<td>80</td>
<td>77</td>
</tr>
<tr>
<td>index of return</td>
<td>2.40</td>
<td>2.40</td>
</tr>
</tbody>
</table>

A more detailed analysis suggests that question 1190 happened to have more powerful detractors than question 1200 (see Appendix, Figures 15 and 16). In this example accuracy and RT probably measure different constructs as was suggested by Schnipke and Scrams (1997). RT may indicate mainly the complexity of the questions, in other words, how many and what kind of steps are required to reach the solution, and it may have only a minor component of difficulty, that is the ability to find the right path to the solution. The number of correct answers represents mainly how many test takers failed to recall the right path and it correlates to some extent with the number of steps required to reach the solution.

**Question Equality and Response Time**

**Measurement units: conversion by multiplication and by division**

Three questions (nos. 20, 30 and 90) involve multiplication for conversion of measurement units while three other questions (nos. 70, 40 and 230) involve division of the same units. For example, question 20: “Complete the following statement: 8.65
L = ? (mL)” and question 70: “Complete the following statement: 2580 mL = ? (L)” were rated as “nearly equal” by an expert. The study compared the MQRT for these questions (Figure 6).

Figure 6. Comparing mean time for conversion by multiplication and division.

As can be seen (Figure 6), for all three pairs of conversion tasks the questions involving division (conversion from smaller units to bigger ones) take less time (20-30%) than those involving multiplication. Consequently, these pairs of questions should be considered quasi-equal, that is that they appear to be equal but they are not. This finding is prompting further investigation into the steps that test takers undertake to complete these tasks and is an example of how data available from computerized testing, such as Response Time, can allow educators to be alerted to ways students think.
**Conversion of measurement units with a prefix “milli”**

The Metric prefix “milli-” indicates difference between numbers of 1000 times. In order to convert litres into millilitres or metres into millimetres a test taker needs to multiply the first unit by 1000. Conversion of litres into millilitres or metres into millimetres therefore appears to be the tasks requiring equal competency (that is, multiplying by 1000). However the MQRT suggests that these questions are not equal, as conversion from metres into millimetres takes longer than conversion from litres into millilitres and the same is the case for the reverse conversions which require division by 1000 (Figure 7).

*Figure 7. Conversion of measurement units with a prefix milli- : litres vs. metres.*

The source of the MQRT difference for these tasks is an interesting subject for further investigation. One of possibilities is that school curriculum introduces conversions between litres and millilitres as a single step procedure. However, it presents conversion from millimetres to centimetres as a separate skill first and then moves to conversion from centimetres to meters sequentially, thus making the conversion from millimetres to metres a two-step process. Again, RT data offers
researchers insight into student’s knowledge acquisition that may benefit both test development and general teaching and learning practice.

Summary

The results presented above form substantial grounds to answer the first two research questions:

**Index of return** was found to be an independent construct having only weak correlations with percentage of correct answers and mean question response time. Index of return demonstrated a sufficiently small variation for equal questions and thus can be suggested for use in test question calibration. It is suggested that monitoring the index of return may provide a new insight into the discussion of question difficulty and test takers’ confidence.

**The mean of response time of a question** has been found to have a very low variation (2-8%) for equal test questions. This variability proved to be stable in the full range between 20-30 seconds for a shorter RT and 200-250 seconds for a longer RT for the questions engaged in the test. MQRT variation is shown to be a sufficiently predictable variable for groups of 25 and more participants, which makes MQRT a reliable tool for verification of expert opinion of question equality or measuring the extent of difference for quasi-equal questions. Monitoring MQRT has demonstrated a high potential in revealing hidden quasi-equality of test questions, which can be missed by experts, and which may be worthy of further investigation.

**Index of Return and Response Time** are suggested both for using in validating the equality of questions in a test pool and for gaining insights into student learning. The data about temporal behaviour offers a new approach to developing models for automatic computer generation of test questions (parallel items). It seems that using only item parameters statistics as it is done in IRT-based adaptive testing
for verification of items equality may be insufficient. Pre-testing calibration of test questions should be based on a computerized mode of delivery and the RT data generated can offer considerable benefits in comparing the items. Application of RT measurement can be done on small groups of test takers which can greatly reduce the risk of item over-exposure at the pre-testing stage, a concern expressed by Hambleton (2002).

As was found (Table 4), a different set of variables in the same task can considerably alter (up to 25%) time demands of the question. In cases when a random generation of the numbers in a mathematical task is used for question modification during a test administration, it may be useful to change random selection from a range of numbers to the selection from an array of pre-determined numbers that have been tested for equivalence. This will help to avoid presentation of too easy or too difficult to calculate variables. The issue of implementation of these data in adaptive testing is a rich area for further research.
Chapter 4. Results and Discussion: Patterns of Response Time

This chapter presents the results of data analysis that investigate the following research questions:

- **Variability of Response Time (RT) between test takers**
  - Does variability of Response Time between test takers raise concerns about test fairness to slow performers?
  - Do slower performers get lower scores?

- **Variability of RT within a test taker**
  - Does variability of RT within a test taker correlate with the test taker’s average speed or accuracy?

- **RT distribution for right and wrong answers**
  - Do wrong answers take longer than right ones?
  - Can a delayed response be associated with a higher probability of a wrong answer?

**Accumulated time spent on all questions and accuracy**

Though the test administered by this study had a time limit (55 min), 86% of all test takers completed it before the time expired. 5% of test takers used the total test time and still did not get a pass mark. It seems unlikely that additional time would change the outcome for those test takers. This suggests that the time pressure of the test can be excluded from the number of factors affecting performance. This makes it possible to investigate if the finding that accuracy and speed are not correlated on complex tasks under unlimited time conditions (Jensen, 1987) is true for this mathematical testing environment.

The relationship between time taken to attempt all questions for the first time and the percentage of correct answers (accuracy) is illustrated in Figure 8. It was found that a test taker’s test score is not correlated (Pearson correlation coefficient = -0.05) with the time it took to attempt all questions. Test takers who completed last
could have as high a final score as those who completed first. This fits with the finding of Bergstrom et al. (1994) that low scoring examinees did not take longer to respond to questions than high scoring examinees. The absence of correlation extends to this mathematical testing environment the finding of Jensen (1987) that accuracy and speed are not correlated on complex tasks under unlimited time conditions.

*Figure 8. Accumulated time and accuracy.*

The variability of Response Time between test takers

Variation of test takers’ RT was analysed within a question and then within a test taker across all questions. Consistent with previous studies (Bergstrom et al., 1994; Schnipke & Scrams, 1999b), the between-test takers difference in RT was characterised by relatively high standard deviation for each question, that is, questions that take longer to complete, have larger standard deviation, than questions that take less time to complete.

It can be seen from Figure 9 that there is a linear relationship (Pearson correlation coefficient = 0.96) between standard deviation and the MQRT of a
question, which reflects that the impact of the difference in performance speed grows in direct proportion with increasing question complexity.

Figure 9. Correlation between mean response time and standard deviation on a question: between-test takers variability.

Significant differences across test takers in their normal performance speed should be considered to provide fairness of final scoring on a test with high time demands as time restrictions can have different impact on fast and slow performers, as it was found by Jensen (1987). Jensen (1987) established that time pressure can compromise accuracy of a slow test taker dramatically before it had any impact on a fast test taker.

Time pressure on the test taker on a non-adaptive test can be examined by the number of test takers unable to complete all questions before the end of the test. Time pressure of an adaptive test is harder to establish because for lower achievers the test is over when their upper level of ability has been determined by test scoring algorithm. Besides, more able test takers receive harder items, which may involve more time-consuming computations. Bridgeman and Cline (2000) did not find any
indication that students who were administered items with long RT were disadvantaged. However, their study did not specify the time pressure of the test nor did it address the question of what score would have been achieved by a test taker whose progress was stopped by a time limit. Bridgeman and Cline (2000) conclude that difference in MQRT has a potential to create unfairness on a timed test.

When a computerized test employs an adaptive model or randomised question selection in non-adaptive model, each examinee receives a unique set of test items (Wise, 1997). If the question selection process does not recognize quasi-equality of questions there is the possibility that it may offer more complex questions with higher time demands and from domains that are a low speed area for the test taker thus discriminating against the test taker.

The data (Figure 9) suggests that individual differences in RT should be considered at the stage of test design. The question selection process could be designed to incorporate a capability to use the statistic of RT from previous test administrations to compute an expected time for each question to equalize the sets of presented questions according to time demands.

The variability of Response Time within test takers

At the time when this study began it was not clear if ranking test takers by their response time on a question would exhibit any consistency across the test. That is, would the fastest test taker be the fastest on all questions? If a test taker’s RT were not consistent, how much would his/her performance speed vary between questions? Would that variation be bigger or smaller for the test takers who get the highest score or who are the fastest?

The idea of using RT ranking for the purpose of knowledge estimation is attractive. If RT ranking were found to be consistent through a test, it could make
additional information about a test taker’s internal state and strategies he/she applies during the test available to a scoring procedure. If a correlation between change in RT rank on a question and the degree of certainty in that cognitive area was established, RT measurement could deliver an insight into the quality of a test taker’s answer and may help to differentiate the answers by degree of certainty. That is, a delayed response could be treated as indication of partial knowledge, decreasing the weight of the correct answer for scoring purposes, and thus making it possible to test the strength of the knowledge on a more precise basis than pass/fail.

Additionally, RT could potentially reveal the extent of time pressure on a test taker by monitoring the change in his/her performance speed across the test. As was suggested by Schnipke and Scrams (1999a), RT can help to detect the “lucky guess” strategy employed by a test taker on a question and could inform the scoring process to compensate use of this strategy.

Analysing the performance speed of a test taker across different questions requires finding a reference point. It could be a set of pre-calibrated questions to determine a test taker’s speed to compute expected RT for other questions or it could be a relative speed parameter defined by the relation of the RT of an individual on a question to the arithmetic mean of the question response time (MQRT). For the purpose of further data processing this study used the latter option. Each response time (RT) value was replaced with standardised and normalised response time (SNRT), that is a standard score calculated by comparing the logarithmically transformed RT of a test taker on a question and the average RT of the same question, in terms of the standard deviation on the question (for the procedure, please refer to Chapter 2, Data Analysis).
As the first step in investigating this issue the study measured the variation of RT within a test taker and then investigated if the variation correlates with other variables of a test taker’s performance, such as total test accuracy and average performance speed.

A distribution histogram of standard deviations computed for each test taker’s set of SNRT (Figure 10) demonstrates that the distribution of within-test takers’ variation of SNRT is characterised by a wide range of distribution. The value “0.00” on the (x) axis would mean that the RT of an individual was consistent across all questions relatively to the average time. As it can be seen from this graph there is no such person. The test taker whose performance is most consistently close to the average time on all questions still has a standard deviation value of 0.54. Most test takers had their SNRT varying within 0.72 and 1.33 standard deviations.

*Figure 10. Histogram of distribution of Standard Deviation of test takers’ Standardised and Normalised Response Time.*

This graph demonstrates that a test taker’s rank is highly inconsistent across the test that makes it impossible to rank test takers without factoring out the major variables affecting response time. Bergstrom *et al.* (1994) found that such examinee
variables as test anxiety, gender, ethnic background, age and language accounted for 2% of the variance in RT and question variables, such as question difficulty and design, accounted for 19% of difference. Most variation in RT remains unexplained up to date. Investigating if the degree of certainty could be one of the factors at least partially explaining this variation is suggested as a focus of further research.

**Changing speed across the test: four pacing patterns**

The SNRT was computed separately for the beginning, middle and end groups of questions for each test taker to determine if the alteration of performance speed across the test possesses some regularity in regard to a phase of the test. Each test taker’s pattern was grouped according to one of four models (Figure 10), (for details, please refer to Chapter 2: Data Analysis).

The distribution of test takers between accelerating (22%) and decelerating (22%) patterns found by this study match the findings of Schnipke and Scrams (1999a) who found 23% of Graduate Record Examinations (GRE) examinees accelerated as they proceeded through a quantitative test and 22% of examinees decelerated.

*Figure 11. Pacing patterns.*
It should be noted that though GRE administration time (45min) and number of questions (28) are similar to that of the test used by the current study, GRE is based on an adaptive item selection algorithm. Bridgeman and Cline (2000) in their study of GRE results of 5,957 examinees suggested that examinees who are responding rapidly at the end are to some extent balanced by examinees who are responding more slowly at the end. It seems that such accelerating and decelerating patterns are true both for non-adaptive and adaptive testing.

The study observed that the test takers who accelerated during the test, also tended to get a slightly higher score (Table 6). Analysis of variance indicated that there was a statistically significant difference between pacing groups in the mean percentage of correct answers ($F(3,134) = 3.10, P = 0.03$). Tukey's pairwise comparisons indicated that the difference between the group of the test takers accelerating during the test and the group of test takers going slower in the middle of the test is the only one that is statistically significant (at the 0.05% level).

Table 6. Pacing pattern and accuracy.

<table>
<thead>
<tr>
<th>Pacing pattern</th>
<th>Number of test takers</th>
<th>Average per cent of right answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerating during the test</td>
<td>30</td>
<td>83%</td>
</tr>
<tr>
<td>Decelerating during the test</td>
<td>30</td>
<td>76%</td>
</tr>
<tr>
<td>Going slower in the middle of the test</td>
<td>34</td>
<td>74%</td>
</tr>
<tr>
<td>Going faster in the middle of the test</td>
<td>40</td>
<td>79%</td>
</tr>
</tbody>
</table>

This tendency was not reported by the prior research of Schnipke and Scrams (1999a) or Bridgeman and Cline (2000) and warrants further investigation.

**Test score and variation in Response Time**

In order to explore possible relationships underlying the variability of RT the study investigated the degree to which a variation in SNRT correlates with other test performance variables, such as total test score (accuracy) and mean SNRT of a test
taker. The correlation between SNRT variation and test accuracy can be examined on a scatter plot (Figure 12), which demonstrates that the variation of SNRT is not correlated with total test accuracy. The SNRT of the top 20% of achievers varied as much as the middle or lower 20% of achievers. That is the performance speed of a high scoring test taker varies as much as the performance speed of a low scoring test taker.

*Figure 12. Variation of SNRT vs. test accuracy.*

Response Time variation vs. mean Response Time of a test taker

The data were further analysed for possible relationship between RT variation and average performance speed of a test taker. Figure 13 demonstrates that SNRT variability is not dependant on the mean SNRT of a test taker. This means that generally quicker performing test takers have as much variation in their performance speed as the test takers with medium or low performance speed.
The analysis of within-test taker’s variability of response times indicates that individual response time greatly varies across different questions. This variation does not correlate with average performance speed or the accuracy of a test taker.

The study observed a high variation in RT for different tasks within the same cognitive domain where the variation cannot be accounted for by the seven different kinds of mental speed that could influence cognitive abilities in different cognitive domains as was proposed by Cattell (1998). The results of the current study support the conclusion of Roberts and Stankov (1999) who argue that other additional factors must be considered while investigating the variation in performance speed.

**Difference in Response Time for right and wrong answers**

The next research question is whether the probability of wrong answers is correlated with test takers’ RT. Establishing a correlation could lead to the possibility of using RT as a predictor of correctness of an answer.

First, the study needed to find if there was a significant difference between response time for right and wrong answers. To determine if wrong answers take longer than right answers, as was found by Bergstrom et al. (1994), Hornke (1997)
and then supported by Rammsayer (1999), the data were analysed by two parameters: (1) difference in mean question RT for right and for wrong answers for each question and (2) difference in mean SNRT for right and wrong answers for each test taker.

The application of the second parameter is required as there is controversial evidence about the role played by the factors of general ability (g\textsubscript{f} and g\textsubscript{c}) as defined by Cattel (1998) in the difference between mean time for right and for wrong answers for a question. Thissen (1983), as cited by Schnipke and Scrams (1999a), suggested that test takers with a higher level of general ability tend to answer questions more correctly and quicker. On the contrary, Schnipke and Pashley (1997) reported relatively high correlation between a test taker’s ability and slowness for the verbal- and quantitative-reasoning tests. In general, Schnipke (1997) found that high-scoring test takers tended to take more time then low-scoring test takers. However the difference was found to be non-existent for the analytical-reasoning test.

In order to deal with these contradictory findings the data had to be analysed within test takers to determine if a difference between the performance speed on right and wrong questions exists both within a question and within an individual. To contrast right and wrong answers within each question the study calculated MQRT separately for right and wrong answers for all questions and compared them using a t-test: Paired Two Sample for means (for more details, please, refer to Chapter 2, Data Analysis).

The t-test indicated that the null hypothesis that there is no difference in RT for right and wrong answers has to be rejected in favour of the alternative hypothesis, which states that there is a statistically significant (at the .05 level) difference between the mean time for right and for wrong answers.
Table 7. Difference between MQRT for right and wrong answers for all questions.

<table>
<thead>
<tr>
<th></th>
<th>Right</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (in sec.)</td>
<td>65.18</td>
<td>71.76</td>
</tr>
<tr>
<td>Variance</td>
<td>3351.99</td>
<td>3727.81</td>
</tr>
<tr>
<td>Observations</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Hypothesized Mean Difference</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>t Stat</td>
<td>-2.66</td>
<td></td>
</tr>
<tr>
<td>t Critical one-tail</td>
<td>1.67</td>
<td></td>
</tr>
<tr>
<td>P(T&lt;=t) two-tail</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>t Critical two-tail</td>
<td>2.00</td>
<td></td>
</tr>
</tbody>
</table>

For 60% of test questions the mean time to complete wrong answers was longer than to complete right answers. Further investigation is needed to analyse why wrong answers do not take longer for some questions. One of the findings of previous research (Swanson, Featherman, Case, Luecht, & Nungester, 1997) is that while high achievers allocate more time to more difficult questions, low achievers do not recognise difficult questions and do not increase time allocation proportionally. This pattern may account for some wrong answers that were not well thought out, as the test takers did not recognize the complexity of the question. This may have affected the MQRT by reducing the mean time for wrong answers on some questions.

To determine the difference between RT for right and wrong answers across different questions within a test taker, separate mean SNRT for right and wrong answers were compared using a t-test: Paired Two Sample for means (Table 8). The test resulted in the null hypothesis of no difference being rejected in favour of the alternative hypothesis, which states that there is a statistically significant (at the .05 level) difference between right and wrong mean SNRT within individual results. The mean SNRT for wrong answers was larger for 60% of test takers, which means that more time was required on the questions that they answered incorrectly.
Table 8. Difference between mean SNRT for right and for wrong answers within a test taker.

<table>
<thead>
<tr>
<th></th>
<th>Right</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>Variance</td>
<td>0.16</td>
<td>0.37</td>
</tr>
<tr>
<td>Observations</td>
<td>135</td>
<td>135</td>
</tr>
<tr>
<td>Pearson Correlation</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Hypothesized mean Difference</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>t Stat</td>
<td>-3.03</td>
<td></td>
</tr>
<tr>
<td>t Critical one-tail</td>
<td>1.66</td>
<td></td>
</tr>
<tr>
<td>t Critical two-tail</td>
<td>1.98</td>
<td></td>
</tr>
</tbody>
</table>

These results support the study of Bergstrom et al. (1994) who reported that examinees spent more time on items they got wrong than on items they got right. Hornke (2000), using data from a large study of general mental ability, replicated his previous findings that wrong responses require more time in an adaptive testing environment. His results about a larger spread of time distribution for wrong answers also match the results obtained in this study. However, their study did not address the question of how to determine the cut off line after which the RT of a test taker could be flagged as delayed or how RT may be related to the probability of a wrong answer.

**Response Time distribution pattern**

To examine if there are some values of SNRT that are associated with higher or lower level of errors the data of performance speed and accuracy were summarized. The accumulated distribution chart (Figure 14) accounts for all SNRT of all test takers measured during the test. As it can be seen on the chart, SNRT for wrong answers tends to be shifted to the right, towards positive values, which indicates that in general wrong answers take slightly longer then correct answers.

However, the chart also provides a more detailed picture of SNRT distribution, demonstrating which parts of the distribution are associated with a higher proportion of right or wrong answers. It must be kept in mind that SNRT represents a
logarithmically transformed RT, and all differences between RT values have been compressed on a logarithmic scale. This means that even a small difference between SNRT values corresponds to a much bigger difference in values of response time.

Answers with SNRT higher than 0.4 are characterised with a higher level of errors (Figure 14). The difference is significant in the ranges 0.4 – 1.3. A higher level of errors is also produced by answers with SNRT less than -2.2. On the contrary, answers with SNRT between -1.2 and 0.4 are characterised with a lower level of errors. The difference is significant in the ranges from -1.2 to -0.7 and from -0.1 to +0.3.

*Figure 14. Summary of distribution of SNRT across all entries.*

The analysis of the distribution suggests that some values of SNRT were associated with higher or lower probability of incorrect answers. The strength of this association and its replicability requires further investigation. This distribution summary provides a general picture of some regularity in the distribution of time data which may be used to create a refined test taker’s RT profile based on his/her estimated performance speed and the time difference for his/her right and wrong answers. This requires that most of the major factors responsible for the alteration of
the performance speed are accounted for and the correlations between performance speed and certainty and between performance speed and cognitive domain are established.

**Summary**

The Response Time of a test taker showed high variability both between test takers on a question and within a test taker across different questions. This variability makes it impossible to rank test takers according to their performance speed without factoring in information about their internal state, which was beyond the scope of this study.

No correlation was found between a test taker’s accuracy and the time it took to attempt all questions. This means that performance speed of a test taker does not predict his/her test score when the test does not apply a time pressure on the test taker. However, the magnitude of difference in individual response time suggests that a test taker, who is generally slower on a test, may be discriminated against on a test with high time demands if the time pressure would make him accelerate towards the end of the test, which may result in a higher rate of errors.

The analysis of test takers’ SNRT separately for right and wrong answers suggests that there is some regularity in the data distribution. A delayed question response (with a longer RT) can be suggested as indication of higher probability of a wrong answer that may be caused by insufficient strength of the test taker’s knowledge. The study identified ranges where the probability of a wrong answer was significantly higher. The ranges of the association need to be further refined for different cognitive domains or different subgroups of test takers. Further investigation is required to determine if these ranges may be persistent under different test conditions.
The results of the study suggest that response time exhibits the potential to be used as a supplementary source of information for item selection and scoring modules of adaptive testing systems. High variation of RT needs to be explained before RT can be used to measure the quality of a test taker’s answer.
Chapter 5. Conclusions, limitations of the study and recommendations for further research

Computer-based tests (CBT) are now a common form of test delivery for licensure, certification and admission tests, and the investigation of what additional information can be generated by monitoring how many times test takers return to the same question for review (index of return) or measuring the time spent on a question (Response Time, RT) is of particular interest in the emerging area of computerized testing.

Index of return

A new notion of “the index of return” representing the number of return visits to a question has been introduced by the study and is suggested for test question monitoring. The study found that the index of return measures a construct different from the construct of question difficulty, as it has only weak correlations with the percentage of correct answers and mean question response time. It is suggested that index of return measures the difficulty as perceived by test takers and thus may reveal otherwise hidden aspects of the response elicited by a test question. It may be proposed that questions that have a high proportion of wrong answers and low index of return possess a level of difficulty that was not correctly identified by test takers.

Test item calibration

The adoption of computerized testing by large-scale admission tests has greatly increased the need for item development to compensate item exposure (Wainer & Eignor, 2000). Creating parallel (equal) items to extend a test item pool is
The current study investigated the issue of validating questions equality using measurements of response times.

The index of return has demonstrated very close values for equal questions so it can be suggested as additional tool for item calibration.

The arithmetic mean of a question response time proved to be a reliable parameter of the question. The equal questions, administered to different groups of test takers, yielded very close means with variation of up to 2% for less complex tasks and up to 8% for more complex tasks. The means have also shown to be sufficiently reliable for a smaller number (20-30) of participants. It was discovered that MQRT may expose even very subtle differences that can be underestimated or considered irrelevant by experts. Potentially misleading quasi-equality of questions was found to be effectively revealed by monitoring RT data, which is suggested as a powerful tool for test item calibration procedure.

**Variability of Response Time**

The analysis of the prior research allows crystallisation of factors that may be expected to contribute to RT variation within a test taker and between test takers. The current study aimed at evaluating the variability of a test taker’s RT and investigating probable correlation of the variation with the test score of the test taker and his/her average performance speed. Establishing the proportion of variation that each factor accounts for is beyond the scope of this study and will be a rich area for further research.

Individual RT has been found to have a high variability both between test takers on a question and within a test taker across different questions. High variation between test takers raises concerns about the fairness of test scores on a test with high
time demands. It is suggested that test designers need to monitor RT to make sure that slow test takers are not discriminated against.

The analysis of test takers’ SNRT separately for right and wrong answers suggests that there is some regularity in the data distribution, which allows identifying the ranges where the probability of a wrong answer is significantly higher. The delayed question response can be viewed as an indication of insufficient strength of the test taker’s knowledge associated with higher probability of a wrong answer.

**Use of Response Time data**

Response time exhibits the potential to be used as a supplementary source of information for item selection and scoring modules of adaptive testing systems. The ranges would need to be further refined for different cognitive domains or different subgroups of test takers. Further investigation is required to determine if these ranges would be persistent under different test conditions.

**Limitations of the study**

Due to a small number (<8%) of males, participating in this study, the results of data analysis by gender parameter would not be significant and therefore, the study did not analyse the data separately by gender. Most participants were of the same age and therefore, the data was not separated by age.

In the current study the information about accuracy was based on the final answer submitted by a test taker at the end of the test. It was not possible to determine exactly when the final answer was produced, whether there were intermediate versions, or whether they were altered during the test. Registration of all entries into an answer text box during a test could provide additional insight into the process of reaching a final answer.
Since the focus of the study was on Mathematics the conclusions about temporal patterns are limited to this subject. Further research may be required to investigate the same research questions for other subjects.

**Recommendations for further research**

Further research is required to establish the impact of degree of certainty on an individual’s RT. Logistic regression analysis may be needed to address the issues of regularity in temporal behaviour of test takers across different cognitive domains under different states of certainty. If degree of certainty is confirmed to be the most powerful variable of a person’s RT, its factoring out would allow identifying ability factors across cognitive domains. The comprehensive study of a test taker’s decision making process would help to develop new test scoring models which would assess the quality of a test taker’s answer so increasing the general precision of testing.

Further processing of the data is required to examine test takers’ time allocation for each question with reference to the total test time. It will allow investigation of the proportion of time allocated to easy/difficult and less complex/more complex questions by top achievers and low achievers. This would contribute into the discussion how effective time allocation is related to the knowledge on the subject.

Additional research is required in defining the terms of *question general difficulty*, *question perceived difficulty* and *question complexity*. These question parameters would allow predicting test question functioning and substantiate new test item generation procedures.
Developing a methodology for distinguishing equal, quasi-equal and unequal questions would provide a great input into refining test item calibration procedures.

The accumulated data presents a rich source of information for further analysis of test takers behaviour elicited by different question formats and styles. Response Time measurements can provide a lot more insights into the process of learning and teaching.
## Appendices

### Appendix 1

## Glossary

### Table 9. Glossary and abbreviations.

<table>
<thead>
<tr>
<th>Abbreviation or short name</th>
<th>Full name</th>
<th>Definition and notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT</td>
<td>Response Time</td>
<td>Time between the completion of a question web page download and test taker’s clicking the button “submit the question” or pressing “enter” key on a keyboard.</td>
</tr>
<tr>
<td>NRT</td>
<td>Normalised Response Time</td>
<td>The logarithms of RT instead of the values in order to achieve a normal distribution.</td>
</tr>
<tr>
<td>SNRT</td>
<td>Standardised and Normalised Response Time</td>
<td>A value, representing the difference between a test taker’s NRT and average NRT on a question, in terms of Standard deviation for NRT on that question. It is equal to a standard score, which is calculated by the following formula: [ \frac{\text{NRT}<em>{\text{ind}} - \text{NRT}</em>{\text{avrg}}}{\text{SD}} ] Where ( \text{NRT}<em>{\text{ind}} ) is the logarithm of time spent by a test taker on the question, ( \text{NRT}</em>{\text{avrg}} ) is the logarithm of average time spent by all test takers on the same question, and ( \text{SD} ) is a standard deviation for logarithms of RTs on the question.</td>
</tr>
<tr>
<td>MQRT</td>
<td>mean of Question Response Time</td>
<td>Average RT for all test takers on a question</td>
</tr>
<tr>
<td>MSNRT</td>
<td>Mean of Standardised and Normalised Response Time</td>
<td>Average SNRT for a test taker. It also was calculated separately for right and wrong answers.</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
<td>In this case Standard Deviation is a statistic used as a measure of the dispersion or variation in a distribution, equal to the square root of the arithmetic mean of the squares of the deviations from the arithmetic mean.</td>
</tr>
<tr>
<td>TTFP</td>
<td>Total Time of the First Presentation</td>
<td>The accumulated time for the first approaches to all presented questions</td>
</tr>
<tr>
<td>TTT</td>
<td>Total Time of the Test</td>
<td>Time between the beginning of a test and submitting test results.</td>
</tr>
<tr>
<td>Index of return</td>
<td>Index of return</td>
<td>Average number of approaches to the same questions. It is calculated by dividing the total number of approaches to the question by the number of unique test takers, who have been presented this question.</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Percentage of students who gave right answers on a question</td>
</tr>
<tr>
<td>Experts</td>
<td>Experts</td>
<td>Highly qualified in Mathematics teaching academics</td>
</tr>
<tr>
<td>Entry</td>
<td>Entry</td>
<td>Each input of information</td>
</tr>
<tr>
<td>Pacing pattern</td>
<td>Pacing pattern</td>
<td>Some regularities in changing performance speed across a test</td>
</tr>
<tr>
<td>Test question</td>
<td>Test question</td>
<td>A single test unit, containing instructions and a task to perform</td>
</tr>
<tr>
<td>Test item</td>
<td>Test item</td>
<td>A group of equal questions from which a question is selected for presentation</td>
</tr>
<tr>
<td>CAA</td>
<td>Computer Assisted Assessment</td>
<td>One of the terms for computer-aided testing</td>
</tr>
<tr>
<td>CBT</td>
<td>Computer-Based Testing</td>
<td>The most accepted term for computer-aided testing</td>
</tr>
<tr>
<td>CAT</td>
<td>Computerized adaptive testing</td>
<td>Each answer affects selection of the next question, a test is stopped when a test taker is unable to solve a more difficult question.</td>
</tr>
<tr>
<td>Index of Correctness</td>
<td>Percentage of right answers</td>
<td></td>
</tr>
<tr>
<td>Performance speed</td>
<td>The construct measured by Response Time</td>
<td></td>
</tr>
<tr>
<td>CR</td>
<td>Constructed response</td>
<td>Question format which requires the test-taker to supply an answer to a question</td>
</tr>
<tr>
<td>MC</td>
<td>Multiple-choice</td>
<td>Question format that consists of two parts: (1) the stem, which contains the problem; and (2) a list of suggested answers.</td>
</tr>
</tbody>
</table>
Appendix 2

Plain Language Statement

Project: “Test item response time in computerized assessment“

As a student of subject 485-202 Mathematics 2, you are invited to participate in the above research project, which is being conducted by Dr Dianne Chambers (supervisor) and Mr Eugene Gvozdenko (Masters student) of the Department of Science and Mathematics Education at The University of Melbourne. This project will form part of Mr Gvozdenko’s Masters research, and has been approved by the Human Research Ethics Committee.

The aim of this study is to investigate whether the performance speed of test-takers in computerised testing can be a useful source of information about his/her knowledge. The test has been created with standard software (TestPilot), available via Webraft. The investigation does not interfere with regular testing procedures and complements TestPilot statistics with a time stamp for each item request.

The research will not gather information on your progress or pass/fail results. It will focus only on time it takes you to answer the test questions. It will investigate if it is possible to compare the response times of different test-takers. The accumulated activity log analysis will provide an insight into the relationship between speed and accuracy and help to define the area of uncertain knowledge on the basis of test taker item response speed. It will contribute to the discussion how activity logging can inform item selection and/or scoring process in computerised testing.

You will be asked to complete an on-line test at your normal speed. You can choose time and location at your convenience. The test will take between 30 and 120 minutes depending on your general skills in Mathematics and your computer proficiency. You are greatly encouraged to repeat the test several times to get sufficient training before the hurdle test.

The materials used for the test, have been sourced from previous hurdle pen-and-paper tests and provided by subject co-ordinator. The test item pool will be periodically renewed and enlarged to provide you with training on a wide range of different tasks.
The project does not impose any risk additional to your usual subject participation. Your confidentiality and anonymity will be protected, subject to legal requirements. The students, who choose not to participate in the research, will still have access to the regular version of the test. Information about time-registering and regular versions of the test will be placed on the test web site.

Once the thesis arising from this research has been completed, a brief summary of the findings will be available to you on application at the Department. It is also possible that the results will be presented at academic conferences. The data will be kept securely in the Department for five years from the date of publication, before being destroyed.

Please be advised that your participation in this study is completely voluntary. Should you wish to withdraw at any stage, or to withdraw any unprocessed data you have supplied, you are free to do so without prejudice. If you would like to participate, please indicate that you have read and understood this information by signing the accompanying consent form and returning it in the envelope provided.

Should you require any further information, or have any concerns, please do not hesitate to contact either of the researchers: Dr Chambers: ph. 8344-8556 or Mr Gvozdenko: ph. 0425803295. Should you have any concerns about the conduct of the project, you are welcome to contact the Executive Officer, Human Research Ethics, The University of Melbourne, on ph: 8344 2073, or fax: 9347 6739.

Dr. Dianne Chambers (supervisor)
Mr Evgueni Gvozdenko (Masters student):
Appendix 3

Examples of equal questions

Table 10. Equal questions with a smaller number of test takers.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Simple conversions 2</th>
<th>Simple conversions 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>task</td>
<td>2580 ml = ? (L)</td>
<td>560 ml = ? (L)</td>
</tr>
<tr>
<td>number of test takers</td>
<td>35</td>
<td>22</td>
</tr>
<tr>
<td>% of right answers</td>
<td>100%</td>
<td>95%</td>
</tr>
<tr>
<td>Mean for 1st presentation (in sec)</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>Mean for the total time (in sec)</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>index of return</td>
<td>1.67</td>
<td>1.70</td>
</tr>
</tbody>
</table>

Table 11. Equal questions with longer RT.

<table>
<thead>
<tr>
<th>Questions</th>
<th>volume 270</th>
<th>volume 630</th>
<th>volume 640</th>
</tr>
</thead>
<tbody>
<tr>
<td>task</td>
<td>How many cubic boxes of side length 30 cm can be stored in a shed where the floor measures 2 m by 2.5 m and the available height for storage is 3.2 m?</td>
<td>How many cubic boxes of side length 40 cm can be stored in a shed where the floor measures 3 m by 4.2 m and the available height for storage is 3.2 m?</td>
<td>How many cubic boxes of side length 60 cm can be stored in a shed where the floor measures 5.6 m by 2.5 m and the available height for storage is 3 m?</td>
</tr>
<tr>
<td>number of test takers</td>
<td>40</td>
<td>50</td>
<td>43</td>
</tr>
<tr>
<td>% of right answers</td>
<td>78%</td>
<td>74%</td>
<td>84%</td>
</tr>
<tr>
<td>Mean for 1st presentation (in sec)</td>
<td>162</td>
<td>161</td>
<td>172</td>
</tr>
</tbody>
</table>
Appendix 4

Examples of quasi-equal questions

Q.1190

Match the following statement with the appropriate diagram:

AB'C' is the image of ABC after reflecting about the green line L2.

Figure 15. Quasi-equal questions (Q.1190) about reflections.

Q.1200

Match the following statement with the appropriate diagram:

A'B'C' is the image of ABC after reflecting about the blue line L1.

Figure 16. Quasi-equal questions (Q.1200) about reflections.
Appendix 5

Practice test completion times

Figure 17. Time used by test takers to complete a practice test.
References


94