MEANING AND POTENTIAL OF TEST RESPONSE TIME
AND CERTAINTY DATA: TEACHING PERSPECTIVE

PhD Thesis
Evgueni (Eugene) Gvozdenko

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy

October 2010

Melbourne Graduate School of Education

The University of Melbourne
Computerised testing is becoming a major component of and an increasingly preferable method of assessment. The potential of the vast information generated during a test to increase the precision of evaluation and to extend the utility of a test for test design, teaching and learning are in the focus of current scientific discourse. This study contributes to the discussion by investigating the meaning of response time and certainty data from a teaching perspective. The study also explores the utility of response time measurements to deliver information about certainty, thus capitalising on the benefit of unobtrusiveness of the measurements.

The study utilised a quasi-experimental design, with the data collected from computerised tests in basic mathematics skills administered to tertiary students as part of their curriculum in a University setting. With the focus on the potential of response time to inform teaching, the study examined: (a) how response time could reveal the difference in cognitive load between test items; (b) how response time could reflect solution strategies; (c) how response time could be used to detect cognitive progress and to monitor the impact of teaching on a student cohort across a semester; (d) what meaning is conveyed by different combinations of response time and certainty; and (e) whether there is a relationship between response time and certainty and then whether response time can be used as a proxy for certainty.

The research adopted a data-driven approach based on the use of quantitative data analysis to identify a phenomenon and prompt its exploration through qualitative methods of interviews and discussions with test takers and teaching experts who then provided further interpretation of the phenomenon.

The study found that response time and certainty data deliver the information about underlying cognitive phenomena that are not captured by traditional accuracy statistics. The study established that response time measurements can assist in the following education processes: (a) designing parallel test items; (b) evaluating difference in cognitive load; (c) estimating prevalence of mental solution strategy over written strategy; and (d) monitoring the impact of teaching on specific cohort subgroups across a semester. The study established the relationship between response time and certainty which indicates that response time data can be used instead of certainty for assessing group readiness for new learning. Additionally, the study determined the format for presenting the data that was perceived as the most meaningful by a teacher.

The findings of the study are especially valuable for enhancing the pedagogical practice of Assessment For Learning which stands in the centre of current education assessment reform. Practical implementation of the findings in education testing software can facilitate a change in teaching practice by providing the means for ongoing monitoring of cognitive dynamics of a student cohort to further enhance the teaching and learning process.
DECLARATION

This is to certify that
(i) the thesis comprises only my original work towards the PhD,
(ii) due acknowledgement has been made in the text to all other material used,
(iii) the thesis is less than 100,000 words in length, exclusive of tables, bibliographies and appendixes.
PREFACE

True knowledge starts in knowing where your knowledge ends.
(in response to Socrates)
ACKNOWLEDGMENTS

Successful completion of this thesis is a product of the highly intelligent and stimulating academic environment at The University of Melbourne. The encouragement, patience and firm belief in final success by my supervisors Dr Dianne Chambers and Prof. Kaye Stacey made this journey well guided. The enthusiasm of my peers, local and international postgraduate students, and on-going academic exchange and interactions created a unique atmosphere of connected learning. Thanks also to the School of Graduate Studies for organising various short courses that have a great value especially in the beginning of the course.

My debt is to those students and teachers who were willing to participate in my research. I would like to acknowledge the helpful suggestions made by Dr Vicki Steinle.

I would like to take this opportunity to thank Graham Hepworth, a statistical consultant, from the Statistical Consulting Centre, University of Melbourne, for his assistance in analysing quantitative data of this research and commenting on the Results chapter of this thesis.

This thesis would not be possible without financial assistance by Australian Postgraduate Award scheme that provided funding for the first half of the course.

Finally, as behind every successful person stands an exclusive life partner, behind the thesis is my wife and our family with their loving support, encouragement and understanding.
TABLE OF CONTENTS

ABSTRACT ........................................................................................................................................... I
DECLARATION ........................................................................................................................................ II
PREFACE ................................................................................................................................................ III
ACKNOWLEDGMENTS ........................................................................................................................ IV
TABLE OF CONTENTS ........................................................................................................................ V
LIST OF TABLES ................................................................................................................................ VIII
LIST OF FIGURES ............................................................................................................................. IX
DEDICATION ........................................................................................................................................ X

CHAPTER 1. INTRODUCTION .............................................................................................................. 1
  1.1. PROBLEM STATEMENT ............................................................................................................... 1
  1.2. SIGNIFICANCE ............................................................................................................................ 2
  1.3. OVERVIEW OF THE STUDY ..................................................................................................... 9
      1.3.1. Aim and scope ....................................................................................................................... 9
      1.3.2. Research questions ............................................................................................................. 10
  1.4. OVERVIEW OF THE THESIS ................................................................................................... 11

CHAPTER 2. REVIEW OF THE LITERATURE ..................................................................................... 15
  2.1. PHILOSOPHICAL FOUNDATIONS .......................................................................................... 16
      2.1.1. Dual nature of individual differences ................................................................................. 16
      2.1.2. Education paradigm through the prism of individual differences ................................... 17
  2.2. COMPUTERISED TESTING .................................................................................................... 20
      2.2.1. Types of computer-based testing ....................................................................................... 23
      2.2.2. Assessing problem-solving ............................................................................................... 25
  2.3. RESPONSE TIME MEASUREMENTS IN TESTING .............................................................. 27
      2.3.1. Validity of response time measure ....................................................................................... 28
      2.3.2. Subgroup bias ..................................................................................................................... 29
      2.3.3. Utility of response time ...................................................................................................... 30
        2.3.3.1. Monitoring Temporal Behaviour across the Test ......................................................... 31
        2.3.3.2. Detecting Cheating due to Pre-knowledge ................................................................. 31
        2.3.3.3. Evaluating Examinees’ Effort and Test Motivation ................................................... 33
        2.3.3.4. Complementing Accuracy ........................................................................................... 34
      2.3.4. Relationship between practice, automaticity and response time .................................... 36
  2.4. CERTAINTY MEASUREMENTS IN TESTING ......................................................................... 37
      2.4.1. Scale of measurement ......................................................................................................... 38
      2.4.2. Individual and subgroup differences ................................................................................. 39
      2.4.3. Application of certainty measurements in teaching .......................................................... 41
CHAPTER 3. METHODOLOGY ......................................................................................49

3.1. FRAMEWORK AND GENERAL APPROACH ..........................................................49
   3.1.1. Scientific method .........................................................................................49
3.2. STUDY DESIGN ..................................................................................................50
3.3. APPLICABILITY OF PSYCHOMETRIC AND COGNITIVE MODELS .................51
3.4. DATA DESCRIPTION ............................................................................................52
   3.4.1. Data sources ................................................................................................53
3.5. EQUIPMENT USED IN DATA COLLECTION .......................................................55
3.6. PILOT STUDY ......................................................................................................57
3.7. PROCEDURES FOR DATA COLLECTION ............................................................57
   3.7.1. Summative tests ........................................................................................57
   3.7.2. Formative tests ..........................................................................................58
   3.7.3. Interviews ..................................................................................................59
3.8. ETHICAL ISSUES .................................................................................................60
3.9. STATISTICAL PROCEDURES ..........................................................................61
   3.10. SUMMARY ......................................................................................................62

CHAPTER 4. RESULTS AND DISCUSSION .................................................................63

4.1. MEANING AND POTENTIAL OF RESPONSE TIME MEASUREMENTS FOR TEST DESIGN AND TEACHING ..........................................................................................................................64
   4.1.1. Defining requirements for data processing: dealing with outliers and distribution skewness ..........................................................................................................................64
   4.1.2. Investigating MQRTs on parallel tests ..........................................................67
   4.1.3. Evaluating the difference between parallel items .......................................68
   4.1.4. Detecting difference in question complexity ..............................................70
   4.1.5. Investigating difference in required skills .....................................................73
   4.1.6. Investigating solution strategies used by test takers ..................................75
   4.1.7. Examining rapid-guessing, cheating and distraction in formative testing ....81
   4.1.8. Monitoring learning progress of a group across a learning period ..............84
      4.1.8.1. Dynamics of Student–Expert Performance Speed ..................................87
      4.1.8.2. Novices vs. Student-Experts ..................................................................89
4.2. MEANING OF CERTAINTY DATA FOR TEST DESIGN AND TEACHING .......92
   4.2.1. Defining a scale for certainty measurements: the pilot study .....................92
   4.2.2. Confirming construct validity of certainty measurements and discussing potentially confounding factors .................................................................97
   4.2.3. Exploring subgroup bias in self-evaluation ................................................105
   4.2.4. Discerning the meaning of certainty data for teaching ................................108
   4.2.5. Analysing responses within a test ...............................................................110
   4.2.6. Identifying “hazardous” content areas across curriculum .........................117
4.3. RELATIONSHIP BETWEEN RESPONSE TIME AND CERTAINTY ...............119
CHAPTER 5. CONCLUSIONS...........................................................................................................125

5.1. RESPONSE TIME ...............................................................................................................125
5.2. CERTAINTY ..........................................................................................................................127
5.3. RELATIONSHIP BETWEEN RESPONSE TIME AND CERTAINTY ..............................128
5.4. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH ....................129
5.5. OUTCOMES..........................................................................................................................131

LIST OF PUBLICATIONS PRODUCED OR PRESENTATIONS MADE DURING CANDIDATURE..................................................................................................................133

REFERENCES ..................................................................................................................................135

APPENDICES ..................................................................................................................................149
LIST OF TABLES

Table 1. Data sources..................................................................................................................................54
Table 2. Mean question response time (MQRT) and standard deviation statistics before and after
removing outliers in summative test S2006. ...........................................................................................66
Table 3. Questions requiring conversion from square metres into hectares (S2005) ................................73
Table 4. Questions requiring conversion from square metres to hectares. ..............................................75
Table 5. Descriptive statistics for question 140_S2006............................................................................76
Table 6. Suggested steps of solution for questions 140_S2006 and 140_WOT20_F2006 .........................80
Table 7. Number of fast and slow responses in test WOT4_F2006..........................................................82
Table 8. Number of “too fast” responses in test WOT9_2006 by accuracy ............................................83
Table 9. Questions with negative change in facility across a semester ................................................76
Table 10. Change in MQRT for students-experts in test WOT2_F2006..................................................88
Table 11. Questions exhibiting greatest reduction in MQRT of expert subgroup in WOT2_F2006. ........88
Table 12. Percentage of test takers using given number of certainty categories in summative test S2005
(n=188)................................................................................................................................................93
Table 13. Usage of a five-category certainty scale in summative test S2005 (n=188) ..........................94
Table 14. Percentage of test takers using given number of certainty categories in summative test S2006
(n=177)................................................................................................................................................95
Table 15. Usage of a three-category scale in summative test S2006 (n>140) ..........................................95
Table 16. Usage of a three category scale in formative tests F2006* (n>90)..........................................96
Table 17. Percentage of test takers using given number of certainty categories in formative test series
F2006 (n>90). ........................................................................................................................................96
Table 18. Number of correct and incorrect responses by certainty category from test S2005..............98
Table 19. Number of correct and incorrect responses by certainty category from test S2006..............98
Table 20. Number of correct and incorrect responses by certainty category by achievement from test
S2006. ......................................................................................................................................................107
Table 21. Meaning of six combinations of certainty and accuracy. ....................................................108
Table 22. Questions of formative test WOT4_F2006 and WOT4_F2007..............................................115
Table 23. Questions-outliers identified with regression analysis of certainty and response time .........124
Table 24. Content of test S2006v1.........................................................................................................151
Table 25. Results of ANOVA for three parallel versions of test S2006. ..............................................158
LIST OF FIGURES

Figure 1. Scale of certainty with five categories .......................................................... 55
Figure 2. Scale of certainty with three categories ......................................................... 56
Figure 3. Mean response times (MQRT) for all questions of three parallel tests S2006v1, S2006v2, S2006v3 .......................................................... 68
Figure 4. P-values of ANOVA of three parallel tests S2006v1, S2006v2, S2006v3 ............... 69
Figure 5. Three versions of question 240_S2006 ........................................................... 71
Figure 6. Distribution of response times on question 240_S2006v1 .................................. 72
Figure 7. Distribution of response times on question 240_S2006v2 .................................. 72
Figure 8. Distribution of response times on question 240_S2006v3 .................................. 72
Figure 9. Distribution of response times for question 210_S2005 (n=65) ......................... 74
Figure 10. Distribution of response times for question 530_S2005 (n=91) ....................... 74
Figure 11. Box plots for three versions of question 140_S2006 ...................................... 77
Figure 12. Distribution of test takers’ response times on question 140_S2006v1 ............... 78
Figure 13. Distribution of test takers’ response times on question 140_S2006v2 ............... 78
Figure 14. Distribution of test takers’ response times on question 140_S2006v3 ............... 79
Figure 15. Mental vs. written calculation at each step of solution in question 140_WOT20_F2006 ................................................................. 81
Figure 16. Change in test results between the beginning and the end of the semester (WOT1-WOT4_F2006) ................................................................. 85
Figure 17. Comparing mean question response time for experts vs. novices in WOT2_F2006 ...... 90
Figure 18. Number of correct and incorrect responses in each certainty category by achievement subgroups in test S2006 ................................................................. 106
Figure 19. Interpretation of accuracy across certainty categories on a test item 70 (WOT4_2007). ..... 109
Figure 20. The percentage of all responses that are incorrect by certainty in test WOT4_F2006 ........ 114
Figure 21. The percentage of all responses that are incorrect by certainty in test WOT4_F2007 (the same content as in WOT4_F2006) ................................................................. 114
Figure 22. Number of incorrect responses marked as “Most likely correct” vs. total number of incorrect responses in WOT1-WOT10_2006 test questions ................................................................. 117
Figure 23. Odds of a certain response being incorrect vs. percentage of incorrect responses in WOT1-WOT10_2006 test questions ................................................................. 118
Figure 24. Odds of a fast response and a certain response being incorrect on 104 questions in WOT1-WOT10_F2006 ................................................................. 122
Figure 25. Percentage of incorrect responses vs. odds of a fast response being incorrect .......... 123
DEDICATION

This thesis is dedicated to my wife and my parents.

For their endless love, support and encouragement.
CHAPTER 1. INTRODUCTION

This chapter aims to introduce readers to the thesis by stating the research problem and pointing out the significance of attending the problem, describing the background factors that contribute to the importance of the research in the area, presenting the scope of the research and research sub-questions and delineating the structure of the thesis. A section with the definitions for often used abbreviations or terms is included in the Attachments.

1.1. Problem statement

Computerised testing is becoming a major component and increasingly preferable method of assessment. The potential of the vast information generated during a test to increase the precision of evaluation and to extend the utility of a test for test design, teaching and learning are in the focus of current academic discourse.

The research in response time investigates the utility of the measurements as an additional source of information about (a) student cognitive behaviours; (b) cognitive characteristics of test items. The growing interest to the research field is based on the need to increase test efficiency and improve teaching process. The potential to capitalise on the benefits of unobtrusiveness and complete automation of the data collection process in computerised mode of test delivery provides additional incentives to the research.

Most of the existing research is concentrated on (a) examining validity and reliability of the measurements; (b) revealing individual differences and subgroup bias; (c) informing test scoring; (d) monitoring temporal behaviours across a test; (e) detecting cheating or guessing. Prevalence of research on summative testing can probably be explained by the major demand from education administrators, test developers and policy-makers.

However, with the current shift to Assessment For Learning paradigm the demand for the test information that could assist ongoing teaching and learning raises important research questions about the meaning of what can be extracted from response time measurements and about the format of data presentation that is applicable to the current education setting.

This study contributes to the current body of research by focusing on a teacher’s perception of the potential utility of response time on a subject (group) level. The format of the results of data analysis is confirmed as useful by a teaching practitioner.
Additionally, self-assessed certainty in the correctness of an answer is examined for the utility in summative and formative testing as a means to (a) evaluate the distribution of partial knowledge in a group; (b) assess cognitive progress of a group. Finally, the technique to obtain group certainty information by using response time as a proxy measurement is proposed. The real-life setting of the experiment assures the applicability of findings to the current educational environment.

1.2. Significance

The following section describes the background and highlights the significance of the study.

Worldwide debates about the importance of education in an information society demonstrate increasing public concerns about the impact of current education policies on the future economic and political perspectives of a nation. The ongoing discussion of education philosophies and emerging educational approaches tend to draw evidence from assessment results. The business-like vocabulary, used in discussions, with such terms as targets, range, outcomes, reflects a trend to apply the evidence-based decision making model that has been accepted as a requirement for the policy-making process. This new function of educational assessment as a policy-making informant has generated a new interest in the academic research in assessment tools and approaches, with the specific focus on educational testing.

The mass introduction of computers in schools and homes made computerised testing a viable and increasingly preferable mode of test delivery for the purpose of basic skills assessment, college admission and graduation, and professional certification in OECD countries. In the United States the leading testing authority, the Educational Testing Service (ETS), for example, introduced its computerised Graduate Record Exam (GRE) in 1992. Now ETS administers a number of tests, for example The Praxis I Pre-Professional Skills Tests (measuring basic academic skills for teacher licensure and certification) in the USA, and English-language test (TOEFL) around the world mainly on computers.

An underlying force of a rapid introduction of computerised assessment is a demand for monitoring the outcomes of the current education reform, with the major shift from the industrial model of education that was designed for a uniform population and aimed at providing general skills training, to the new economy that values human intellectual capital over industrial assets or financial structures, that requires highly creative employees with a diverse range of skills. The new economy is globally oriented
and is characterised by a multicultural working environment which promotes ongoing self-directed and collaborative learning, and requires additional sets of non-cognitive skills in collaboration and interpersonal communication.

Furthermore, the understanding of learning has changed dramatically from static behavioural model where an individual was treated as an empty vessel to be filled with knowledge, to a multidimensional, dynamic cognitive model of learning where individuals are informed and actively involved in learning process, determining directions and scopes of their education.

Changing requirements to education in general motivated a significant shift in education philosophy and the assessment paradigm.

**Current shift in education philosophy to a student-centred approach** requires new means of evaluation of system functioning, based on evaluation of the progress of an individual. For successful implementation of a user-centred education model it is important that, on one hand the learning is individualized to meet the needs of an individual, on the other hand the individual has to make an informed decision about the direction of the learning. It means that the provision of such information by the means of advanced computer-based diagnostics is crucial to the success of the approach.

**Significant shift in the testing paradigm** to utilising “testing-for-learning” reflects new demands to educational assessment (Wainer et al., 2000). The essence of the shift was expressed by Stout (2002) in his Presidential Address at the 67th Annual Meeting of the Psychometric Society:

> The long dominant summative-assessment-focused and single-score-based testing paradigm that unidimensional IRT [Item Response Theory] modelling so effectively addresses has begun to be challenged. Summative assessments focus on broad constructs such as mathematics, reading, writing, or physics and typically assess individual educational achievement as examinees exit one system and make the transition to another, such as from high school to college … The new testing paradigm calls for a blending of summative and formative assessments, with sometimes the same test being used for both purposes and sometimes new tests being developed for formative assessment purposes alone. Examinee skills-based formative assessments, often aggregated over classroom, school, school district, state, or other collective, as appropriate, can be used by students, teachers, curriculum planners, government departments of education, and
others involved in the educational process, to facilitate teaching and learning. In particular, using feedback from a test-based formative assessment, each student learns what skills he or she has mastered and what skills need his or her further attention. For example, a student’s algebra test score is supplemented or replaced by a skill profile indicating mastery of linear equations, nonmastery of factoring, and so forth, such information hopefully leading to informed remediation. Used effectively, this can lead to immediate test-driven changes in instruction for the just-tested classroom and future changes in instruction for future classrooms in the same subject.

This approach initiated new advances in educational testing that are focused on evidence-centred design (ECD) combining student modelling, evidence modelling, and task modelling, thus creating a model-based approach to complex assessments.

Equipping a teacher with an instrument for evaluation of latent processes of cognitive development within a class can make a significant positive impact on the efficiency of teaching. Having a real-time feedback from formative online assessment is valuable on its own as it allows adaptation of learning context and teaching methods to a particular cohort. The facility for monitoring student’s temporal behaviour brings another layer of information and can provide the basis for evaluation of the practice element of learning. Thus the study will contribute to the theoretical discussion of adaptive teaching.

Traditional number-correct format of test evaluation has inherent limitations due to the dichotomous nature of the scale (correct-incorrect). For example, when high achieving students consistently get mostly correct answers on a formative test at the beginning of a semester and at the end, there is no currently available indicator of any progress achieved. One of the promises of response time measurements is that response time as a function of practice could provide indication of the strength of the knowledge and of the position of a learner on a learning curve. Providing the means for the evaluation of the progress as a result of practice will be beneficial to forming a favourable conception.

Mathematical thinking as subject of enquiry encompasses the issues that are on intersection of several sciences in the areas of cognition, psychology, neurophysiology and others. All modern societies insist on having mathematics in a school curriculum. Devlin (2001, pp. 272-279) notes that additionally to a general reason that the world is so dependant on science and technology that everybody needs
“to develop a general understanding of mathematics and an ability to acquire special mathematical skills as and when required” there is another strong argument: everybody needs to develop the particular mode of rational thought that is offered by mathematics reasoning. Devlin (p.186) suggests that the ability to do mental calculations in the course of the evolution of humans may have been a precursor to other higher order cognitive abilities including human linguistic ability. This makes the evolution of mathematical thinking of a student an important factor that impacts the development of other metacognitive skills. It can be expected that monitoring the progression of general mathematical reasoning and specific mathematical skills is becoming an essential part of the assessment of the cognitive development of a student.

A clear distinction is drawn between thinking, mathematical thinking and the body of knowledge described as mathematics. If thinking can be defined as the means used by humans to improve their understanding of their environment, mathematical thinking is a style of thinking that relies on mathematical operations and can be applied in any context (Burton, 1984, p. 36). Mathematical thinking involves two major factors: an object of thinking and a method of thinking. An idea, an observation or any stimulus can be an object or an element of thinking. The approach, procedure or strategy used in thinking is a method. Recognizing a pattern of the relationship between elements by means of applying such methods as ordering, inversing, combining, substituting and others, constitutes mathematical thinking.

Sundre and Thelk (2008) refer to mathematical thinking as “quantitative and scientific reasoning”, “the ability to understand and use mathematics and science as ways of knowing” (p. 6). They have developed an instrument for assessment of the progress of tertiary students in developing that ability. An instrument that allows answering the question whether learning happens seems to be in a high demand in the current education system.

The starting point of mathematical thinking is recognising and classifying the element. It can be suggested that the mental operations of comparing a new set of stimuli with existing memories are effectively a lower order mathematical thinking as they involve finding the difference and similarity between new and learned patterns. This process is heavily based on memory retrieval and in the field of mathematical education is referred to as knowledge of mathematical facts.

As far as most mathematical facts are just the memorised results of applying mathematical thinking to a set of lower level facts, the learning process can be viewed as the ongoing process of converting lower level thinking into memories and using them
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

in a higher order thinking. This understanding accords well with a constructivist view on learning as a process of building a new knowledge into an existing scheme or mental representation.

From this perspective individual mathematical thinking at any particular time is a set of mental processes, each having a different combination of memory retrieval sub processes and methods of information processing. It can be suggested that such a model of mental activity can apply to different levels of processing. The basic architecture of processing can be inferred from known physiological mechanisms of neural activities. Simen (2004) demonstrated that control mechanisms involved in problem solving are determined by brain organisation and are computationally universal. This means that the computational load of a particular thinking process of an individual is a function of the composition of higher and lower level sub processes. Those sub processes can be differentiated by the time demands for information processing. As prior research demonstrated, direct memory retrieval is the fastest approach to solving a mathematical problem. De-composing a process by steps and measuring time for each step could potentially provide rich information about the balance between memory retrieval and the use of the procedural knowledge in individual thinking.

**Problem solving strategies and applied procedural knowledge** are in the focus of the study. The current study will investigate the preferences of high achieving students on several tasks to see if their strategies are characterised by a faster response time and accuracy than of other students. This would allow identifying the strategies associated with the fastest and/or most accurate performance. This study intends to examine whether response time can be a valid indication of a level of a strategy applied by a test taker.

**Evolution of methods of thinking**

Researchers, academics, teachers and education administrators hold conflicting beliefs of preference to particular solution strategies and about classifying strategies by such parameters as efficiency or higher/low level.

A set of methods used by a student for solving a particular problem has a dynamic nature and reflects a current quality of mathematical thinking of the student, which is the ability to apply general methods of thinking, such as pattern recognition, inference, or operation under uncertainty, and the knowledge of mathematical context, such as basic facts and procedures. The procedural knowledge of manipulating the facts to solve a mathematical problem can have a different level of efficiency.
Ashcraft (1982) suggested that a developmental trend in the mastery of arithmetic knowledge could be described as a path from initial reliance on procedural knowledge and methods to a gradual shift to retrieval from network representation of arithmetic facts. A concept of knowledge as a dynamic state of ability to retrieve information from memory and to operate it in a meaningful way, with such an ability increasing as a function of training, requires a new approach to evaluating students’ knowledge. The current study aims at investigating whether RT on formative testing tasks can provide indication of a student’s progress.

**Arithmetic is an ideal field** to study the acquisition of cognitive competence, since the learning content, the learning progress and the learning goals can be easily defined. Thus, learning studies in the field of arithmetic may contribute to our general understanding of learning (Delazert *et al.* 2003, p.76).

It is generally accepted that the development of arithmetic knowledge in children is characterised by progressing from primitive procedures to using more and more advanced strategies. There is still a discussion of whether adults use a preferable strategy for a particular task invariantly or their choice is dependant on other circumstances like a risk or current cognitive load which may cause them to retreat to a previously known less efficient but more accurate solution strategy. In the study of mental computations the participants preferred to use a calculator versus memory retrieval for multiplication tables with increasing stakes of a test. LeFevre, Smith-Chant, Hiscock, Daley, and Morris (2003, p. 203) suggested that a comprehensive framework for understanding the range of arithmetic performance among adults should be based on a multiple procedure perspective.

Thus, investigation into strategies that are used by different sub-cohorts of students can be useful for understanding the dynamics of cognitive processes in the group. The current study will investigate the composition of preferred strategies used by students on several arithmetic tasks, whether the strategies used by students change after the material is presented in the classroom, and whether this change can be monitored with response time (RT) measurements.

The current study continues to explore the issue of the usefulness of monitoring response times in the context of testing different cognitive domains in the area of basic mathematics skills. It demonstrates how response time measurements can provide valuable insights for educators, revealing otherwise hidden issues about test takers’ current understandings and learning progress and identifying content areas in which unexpected patterns of time demand prompt further investigation.
The introduction of computers in assessment is a response to the need to improve assessment processes by maximising evaluation precision and fairness, and minimising the costs. Whereas the cost of teaching is more related to the hours per course and is not greatly affected by how many students attend a lecture, the cost of assessment is a cost per student, which grows proportionally with student numbers. The cost of assessment in higher education is the most rapidly growing component of tuition fees (Ricketts, Filmore, Lowry & Wilks, 2003).

E-learning is an emerging reality. The confluence of rapid development of mobile communication technology, computing power and intelligent user interface created a new paradigm of mobile learning. Investigation of the relationship between RT and learning progress makes the results of the study potentially beneficial for the development of intelligent computer-based tutoring systems. Additionally to that, it would open the process for competitiveness which may provide additional motivation similar to the one that drives computer game players.

New techniques of unobtrusive data collection are being developed for the purpose of monitoring otherwise hidden characteristics of learning progress.

Psychometric analysis of test takers’ behaviours was suggested by Masters and Keeves (1999) for the assurance of the quality of assessment measurements. Weiss and Schleisman (1999), Schnipke and Scrams (1997, 1999, 2002); Schnipke and Pashley (1999); Hornke (1997, 2000) and Bergstrom, Gershon and Lunz (1994) considered speed an important component of ability and drew attention to the need to develop testing models that would include test takers response time, that is, the amount of time a test taker spends on a test item in test scoring procedures.

High demand for the research in this area is suggested by the number of academics (Wainer et al., 2000; Schnipke, 2004), with a particular focus on the question of how response time measurements can improve the precision of cognitive tests by offering additional information about the impact of the question on a test taker. Hornke (2000, 2005), Thompson, Yang and Chauvin (2009) conclude that further research is required about the possible diagnostic surplus of item response latencies and testing times. The relationship between test takers knowledge and their RTs in the context of basic mathematics skills is a subject of the present study.

Setzer and Allspach (2007, p. 17) noted that “An additional area of research in which response time could be explored involves the cognitive processing of the examinee. Is it possible to use response time to better understand the interaction between the examinee and the test items? Some research has already begun in this area (e.g., Wise, Pastor & Kong,
2007) but certainly more is needed. With the recent advances in psychometric models that incorporate cognitive processes, this area of research is rich with opportunity.”

Measurements of a test taker-computer interaction are suggested by Luecht and Clauser (2002) to be “a potential source of vast amounts of information where every keystroke or mouse click, every referencing action, every response, and every elapsed unit of time is a possible source of valid information” (p. 69).

van der Linden (2006a, 2006b, 2009) suggests that any of the current applications of IRT can be improved by using response times as an extra source of information on the persons and items. Examples of such applications are: (1) the use of response times as covariates in IRT item calibration, (2) improved item selection in CAT using the response times on the previous items in the test, (3) empirical determination of test speededness, (4) empirical study of the speed-accuracy trade-off in testing, (5) person fit analyses using response times, for example, to detect cheating or pre-knowledge about the items, and (6) more accurate estimates of the extra time to be offered to students who need accommodation.

1.3. **Overview of the study**

1.3.1. **Aim and scope.**

The overall research aim is to examine how Response Time and Certainty measurements in computerised testing can enhance testing, teaching and learning from the perspective of a teaching practitioner. The study will focus on the meaning that can be automatically extracted from Response Time and Certainty data in the context of tertiary education.

A number of simple statistical procedures for data processing and a format of data presentation that are accepted as useful by a teacher will be explored. The feasibility of inferring some aspects of certainty information from response time data will be investigated.

Participants in the study include students, teachers and a subject coordinator. Computerised summative and formative tests in basic mathematic skills are administered in real life settings of an Education Faculty curriculum.
1.3.2. Research questions.

The study investigates three research questions that focus on the meaning of response time, certainty data and the relationship between them. The potential of the data to deliver the information that is not accessible with accuracy statistics is of a special interest.

1. Meaning and potential of response time data for teaching, test design, curriculum development and cognitive research:
   - How can response time data alert test designers and researchers of the difference in cognitive demand between test items?
   - What are the potential and the limits of response time data to predict the range of solution strategies on correct answers?
   - How can response time data alert teaching of rapid-guessing or cheating due to pre-knowledge?
   - Can response time data detect the presence of misconceptions or misperceptions related to incorrect answers?
   - How can response time data provide insights into the dynamics of the group cognitive development through the course?

2. Meaning of certainty data for teaching and curriculum development:
   - What scale can be used as an instrument to measure self-reported certainty?
   - What is the relationship between certainty and accuracy on summative tests?
   - Is there a difference in precision of self-evaluation between high and low achievers?
   - What meaning can different combinations of certainty and accuracy convey to teaching?
   - What kind of information about certainty is considered useful by a teacher?
   - How can certainty be analysed within and across formative tests?

3. Relationship between response time and certainty:
   - What is the relationship between response time and certainty?
   - Can response time be used to obtain information about certainty?
1.4. Overview of the thesis

This section provides guidelines to reading the thesis and describes its composition, development of arguments and interrelations between parts and sections.

Chapter 2 presents background information and a review of the relevant literature. It is setting the stage, presenting the findings of prior research to identify gaps in current knowledge. Opposing views are contrasted to stress the highlights of academic discourse in particular domains. The methods and experimental conditions used by other researchers are mentioned where it is important for critical assessment of the results.

To lay the foundation for further discussion of the two major constructs of the study, response time and certainty, the chapter examines the nature of intra- and inter-individual differences which stand behind the variation in performance speed and self-efficacy.

The current shift in the education paradigm is interpreted from the point of view of accommodating individual differences. Brief analysis of the previous learning theories is undertaken to reconstruct the trend in the evolution of theories of learning in the past century and the relationship between a dominating theory and learning technologies. The mass introduction of computerised testing is shown to be a logical continuation of the trend and an integral part of modern assessment process in an intelligent learning environment.

The demand for extended diagnostic facilities that would utilise newly available psychometric information and incorporate the latest achievements of cognitive sciences is illustrated by a number of recommendations from prior research. Formative testing is demonstrated to be in the focus of the current change in the assessment paradigm, with a strong emphasis on the value of timely cognitive diagnostics in the interests of all parties of the learning process: students, teachers and education administrators. The functionality of different types of computerised tests is discussed to locate the niche for the experiment.

The history of and current developments in research in response time measurements in educational testing are in the heart of the chapter. The discussion of validity of the measure and subgroup bias leads to the central topic of utility of response time for the purpose of monitoring temporal behaviours of test takers, investigating the inequality of parallel test questions, detecting cheating due to pre-knowledge and guessing, and incorporating response time data into scoring procedures.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

The review of the research about the effect of practice on performance speed builds a foundation for an important assumption of the study about the relationship between knowledge, learning progress and response time.

The next major section of the chapter addresses the research questions about certainty. Confidence-based testing is reviewed to distil the potential of certainty to inform teaching, enhance learning and improve test efficiency. The procedures of measuring certainty and scales are highlighted as they directly relate to our research questions.

The issues of the meaning of certainty for teaching, including suggested format of data representation are discussed in regards to individual and group data. The potential threat to validity of the measurements from individual differences and subgroup bias is addressed.

The last section of the chapter discusses prior art in research of the relationship between response time and certainty to situate the final research question in the context.

Chapter 3 describes a theoretical framework and provides a general overview of the methodology of the study. Quantitative and qualitative segments of the study are shown as interweaving methods that provide a holistic picture of an observed cognitive phenomenon.

A statement about two major assumptions of the study’s philosophical position, holistic realism and reliance on correspondent theory, is made with the reference to the scientific method of enquiry. The study design is described in detail to provide readers with all information to form a judgement of the experimental setting. The data section includes the description of major constructs and variables, with special considerations given to the reliability and validity of the data.

The chapter further discusses the research context in regard to the cohort definition, sampling methods and experimental environment. The equipment used in data collection with examples of on-screen appearance of a test is described. The timeline for the phases of the study is presented in a table.

The approach to data analyses is discussed with references to data processing procedures for obtaining descriptive statistics and methods of using inferential statistics.

Chapter 4 reports the findings of the present study and discusses them. The main arguments are summarised and the contribution this study has made to the understanding of the research questions is outlined.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

While validating the measurements scale the study explores the impact of outliers on mean and standard deviation of response time in order to determine applicable statistical procedures for data analysis. Lognormal transformation is discussed to determine whether its benefits warrant its application.

The variance of response times between parallel test versions is examined with the aim of determining a technique for flagging out extreme data points for further analysis. Visual analysis of response time distribution charts is tested as one of the instruments for identifying difference in cognitive demands between test questions.

Mental calculations used in solution strategies are compared with written calculations in respect to time demands. The sensitivity of response time measurements in detecting difference in solution strategies is further examined on another task to define the limitations.

The potential of response time statistics to inform teaching about learners’ progress across a learning period is investigated in the context of ongoing formative assessment. Subgroup difference between expert students who had previously mastered the material and novices who have just learned it is analysed.

Certainty is the focus of the second section of the chapter. As with response time, the study determines the scale for certainty measurements and confirms the reliability and validity of the measurements before proceeding to the discussion of the meaning.

Relationship between certainty and accuracy is examined with the breakdown by three subgroups according to their achievement scores to investigate for a bias in precision of self-evaluation.

Interpretation of different combinations of certainty and accuracy by a teacher is distilled with the emphasis on the usefulness of the data and the best format of the data presentation. The insights of a curriculum coordinator about the meaning of certainty data when analysed within a test and across a sequence of formative tests are also presented.

The last section of the chapter reports findings about the relationship between response time and certainty. Capitalising on the established dependencies the study sets forward a proposal to utilise unobtrusive response time measurements as proxy for certainty to infer distribution of certainty in a group and to detect test items that require special attention. The following summary of the chapter synthesises findings across the three sections.
Finally, in **chapter 5** there is a consideration of the achievements of this study and recommendations for future research. It attempts to put together all the threads that have been followed in the process of this study and suggests some conclusions and implications of the findings for the theory and practice of educational assessment.

While preparing the thesis for publishing, every attempt has been made to adhere to APA (American Psychological Association) style recommendations where practicable. APA is most commonly used to cite sources within the social sciences and its application to the current publication was considered appropriate.
CHAPTER 2. REVIEW OF THE LITERATURE

This chapter describes and summarises primary scholarship that is relevant to the current study. It identifies gaps in current knowledge and opposing views, establishes the grounds for the research questions under investigation, critically assesses the methods used by prior researchers and puts the project into perspective in relation to other work.

In the beginning, the chapter addresses several foundational theoretical works in educational philosophy and psychology to delineate the origin and scope of the phenomenon of individual differences in general. The role of this phenomenon in teaching and learning, and its recognition by a dominant education paradigm are reviewed to lay a foundation for the discussion of the difference in performance speed.

After establishing the philosophical standpoint the chapter describes the current place of computerised testing in education and the main types of testing. Theoretical and practical issues of response time measurements on different types of computerised tests are treated in a more detailed manner. Prior research in validity of the measure, subgroup bias and confounding factors are examined. This puts the research questions concerning individual differences in response times in the context of educational testing.

Further, research in mathematical thinking, and in particular in the relationship between solution strategies and the time required to execute the strategy, is reviewed to provide the background to and justification for the subject of the current enquiry.

Studies in the utility of response time measurements are the focus of the chapter as they form the basis for discussing research outcomes of the current study. Different areas of response time investigation are presented: modelling temporal behaviour across a test, detecting cheating and low effort, informing item selection and scoring processes, and evaluating impact of teaching and learning.

Seminal works in the utility of certainty measurements in educational assessment which is another research question of the current study are reviewed, setting the stage for the discussion on the relationship between response time and certainty.

The summary of the chapter provides an overview of prior research by research question, placing the study in the context of current theory and practice.
2.1. Philosophical foundations

This section determines the grounds of the gnostic position of the author of the thesis and the philosophical stand that will guide this work. As the central point of the study is about differences in cognitive behaviour on an educational test, addressing the definition and the source of individual differences in learning is essential for further discussion.

2.1.1. Dual nature of individual differences.

The discussion of the nature of individual differences in physical and behavioural traits has a long history. The debate about the interaction between “nature” and “nurture” has seen a wide range of opinions, from the deterministic view that most of human behaviours are hard-wired genetically and would prevail over learned behaviours, to an idealistic view that child upbringing can override any inherited personality trait. New findings are coming recently from the research in behavioural genetics which is the interdisciplinary field that studies the role of genetics in animal and human behaviour.

Fuller and Thompson (1960) pointed out that it is the variance in traits among individuals, not the traits themselves that can sometimes be partitioned into genetic and environmental sources. A number of studies (Brooks, Fulker and DeFries, 1989; Baker, Vernon and Ho, 1991) examined the genetic and environmental influences upon general intelligence and specific cognitive abilities. Recent findings provide strong evidence that there is a general factor that accounts for much of the total variance of inter-individual performance on tests of general cognitive ability (Carrol, 1993). The heritability, that is the proportion of phenotypic variation in a population that is attributable to genetic variance, was found to vary from 40% to 80% (Singer, 2006). According to the twin study of 11 thousand twin pairs done by Haworth et al (2009), genetic influence on high cognitive ability (15% of top of the distribution) was in the range from 0.41 to 0.60, which coincides with a general agreement that heritability accounts for about half of the variation in cognitive abilities (Petrill, Thompson & Detterman, 1995).

Jensen’s (1987) theory of mental speed refers to intelligent behavior as mainly the cumulative result of a general biological speed factor which, in the case of computerised testing, is reflected by response times. Bates and Stough (1997) suggested that explicit control of spatial attention requirements indicates that individual
differences in intelligence are substantially underpinned by differences in the speed of information processing.

Since our study is concerned with the variation in performance speed on a test item, the findings outlined above establish the base line for our approach to treating the difference in individual response times as a result of the convolution of genetic and environmental factors. It is assumed that genetic difference would explain inter-individual difference in the case when other cognitive factors are equal. The intra-individual differences in performance speed on test items would then be assumed a function of cultural impact which is a sum of the formal and informal learning experiences of an individual. Major developments in investigation of the relationships between response time and certainty, solution strategies, practice, automaticity, correctness of response, motivation and effort will be covered in more detail later in the chapter.

2.1.2. Education paradigm through the prism of individual differences.

In order for the study to situate the research questions in educational context and to discuss findings in relation to their usability for teaching and learning, this section needs to determine the role and place of individual differences in the current pedagogical paradigm. The present state of the recognition of individual differences as an important factor for teaching and learning by the paradigm is shown from an historical perspective which suggests that the accommodation of individual differences as a driving force of the advancement of education systems in general and of education technology in particular is a logical outcome of the sequence of past education theories.

Learning theories that were dominant in the 20th century can be arbitrarily grouped into one of several paradigms: behaviourism, cognitivism, or constructivism. In the first part of the century, behaviourism led by John Watson and Ivan Pavlov were the dominant psychological and educational frameworks. Learning was viewed as a direct product of operant conditioning where practice would shape appropriate responses to particular stimuli. The role of an instructor was in supplying new information, managing the exercises on remembering and fast retrieval of it and in providing an assessment feedback. Role of a student as a passive recipient of knowledge was reflected in curriculum and teaching methods. Individual differences were considered more a deviation from standard than an integral feature of the population. The school curriculum was aiming at raising the workforce that would suit the environment of an industrial enterprise to satisfy the demands of rapidly growing mass production.
Individual differences were neither recognised as a source of enrichment of the cultural environment nor as the source of creativity at a workplace.

In the 1970s, it became evident that focusing on functional relations between stimulus and response without considering the mechanism of knowledge internalisation was a serious limitation of behaviourism. Noam Chomsky suggested a theory of generative grammar which investigated internal representations and set a starting point to a new paradigm called cognitivism. Capitalising on the availability of computing power and inspired by Information Processing Theory, the researchers were looking at human mental processing, such as thinking, memory, knowing or problem-solving, from the perspective of a computational machine. Brain imaging and computational modelling are among the tools of investigations in cognitive science. There is a prominent body of research in Artificial Intelligence and the philosophies of language and mathematics. However, the gap between the mostly theoretical inquiry and the needs of educational practice required new school of thought.

The constructivist theory of learning proposed by Piaget offered a new paradigm with the focus on scaffolding learning experiences, which rapidly gained popularity. The transition to constructivism and social constructivism by the end of the century witnessed a dramatic change in the prevailing views of the nature of cognition, general purpose of education, place of instruction in learning and the role of technology in the process. The shift was driven by the advancements in the field of educational psychology, with the most prominent theory of child development created by Piaget (1970) who attempted to apply evolutionary perspective of biogenesis to ontogenesis by defining discrete stages in cognitive development.

A paradigm shift to a learner-centred approach based on Piaget’s findings about children’s learning, generated a new pedagogy of constructivism, that recognises an individual as an active builder of their knowledge. It is the declaration of the value of an individual approach as an alternative to “mass-production” or “factory” education that makes Piaget’s theory so important. Using statistical terms, Piaget found that within-group variation in individual needs is highly significant and has to be considered by official education policies.

Piaget greatly advanced the theory of developmental stages, first suggested by Baldwin (1906), which states that there are definite patterns of cognition that are specific to particular age groups. According to the theory, investigating existing cognitive structures and providing the feed for continual balancing ("equilibration" by Piaget) of existing structures with new experiences would result in a most effective
curve of child development. One of the major implications of those stipulations for teaching practice is in increasing emphasis on formative testing as a means to feed diagnostic information to the teaching and learning processes.

Piaget found that informal learning that happens without formal teaching can be even more effective because the control over the learning process is fully obtained by an individual. Learning in such circumstances is ideally suited to the current needs of that individual. This led Piaget to suggest that the content of learning should be selected with the perspective of the evolution of the knowledge structures in the minds of students.

Application of Piaget’s principles in teaching mathematics was developed by Seymour Papert who proposed an educational theory of constructionism. Papert found that learning can happen most efficiently when people are actively involved in physical manipulation with tangible objects in the real world. He applied his expertise in maths to restructure the content of maths teaching to accommodate the natural tendencies of a child. Papert also developed the computing language “Logo” that would stimulate inherent natural abilities and allow the gradual building of mathematical understanding. Mathematical thinking and problem solving are discussed later in the chapter.

Another prominent theory that shaped the current educational paradigm is Socio-cultural theory, first developed by Lev Vygotsky (1978). Though being just 3 months younger than Piaget, he created all his major works within 10 years before his death in 1934. It was not until 1978 when the compilation of his works was published (Vygotskij, 1978) and the socio-cultural framework got recognition in epistemology.

The central point of the framework is the assertion that the mechanism of individual development is determined by society and culture. By extending the example of development of tools in human history to the development of language and number system, Vygotsky demonstrated the dialectical relationship between a cultural system and individual learning. As with a mechanical tool, where learning or internalizing brings behavioural changes and in it’s turn enables a learner to further advance the tool, the same is true with learning a culturally created sign system. Learning changes the mental state and behaviour of a learner and enables a creative input into further development of the system.

Though some theorists do not recognise a connection between Vygotsky and Piaget, pointing out that “Vygotsky as a Marxist was apposite to the view that human nature is given”(Pea, 2004, p. 426), however Piaget’s notion of learning as the process of constructing internal representations through individual interactions with the physical
and representational worlds seems to be encompassing socio-cultural interactions and thus would not contradict Vygotsky’s approach.

To summarise the essence of the currently dominant education paradigm, it is the learner-centred approach, based on constructivist pedagogy and facilitated by information technology, which forms a structure and guidance for education policies. Individual differences are no longer viewed as limiting factors that must be eliminated to achieve consistency in abilities but are considered mutually enriching components of an interactive learning environment. This approach underpins the value of the research in individual response times as a means to fully employ the potential of the data to inform teaching and learning of otherwise hidden cognitive processes, underlying the learning progression of an individual.

2.2. Computerised testing

Computerised testing, also called computer-based assessment (CBA) or computer-assisted assessment (CAA), is an emerging educational technology (Wainer, Dorans, Green, Mislevy, Steinberg & Thissen, 2000) which reflects the rapidly changing education environment in an information economy. The most promising aspects of computer-based assessments are:

- the capability to provide students with complex tasks;
- the feasibility to unobtrusively measure student-task interactions;
- use of statistical methods to explore underlying cognitive phenomena;
- infer learning and problem-solving processes in real-time to provide immediate feedback into the learning process.

Though sometimes computers are viewed mainly as a vehicle for delivery of rich audio-visual content, the real novelty of computerised assessment is in offering an environment for connected learning:

- connectivity when students can use global internet resources to perform a task;
- computational power when students are required to use a software to solve the problem;
- online collaboration facility when a team task is offered.

However, a number of researchers (Lawson, 2001; Bull & McKenna, 2003) point at some disadvantages of CAA and suggest that a caution is required in
implementation. Lawson draws attention to potential pitfalls with computer-aided assessment, particularly in mathematics. The author is concerned about the effect of CAA on learning outcomes and suggests that “care should be taken to ensure that using computer-aided assessment does not introduce extra outcomes to the assessment task or allow the assessment to be completed satisfactorily in a way which does not require the achievement of the outcome being assessed”.

A wide usage of multiple-choice question format is named as another reason for caution due to high risk of a guessing strategy. This leads to constructing complex questions to ensure that students cannot get good marks by a combination of partial knowledge and guesswork (Gardner-Medwin, 1995).

Negative marking which reduces the total score with each incorrect response has been examined for its potential to limit “lucky guesses” on multiple-choice questions. Bush (2001) suggested that examinees should be allowed selecting more than one answer to a question. Though a test-taker is penalised for an incorrect selection, this arrangement was shown to be beneficial to the learners who possess partial knowledge. Another attempt to overcome the limitations of multiple choice format was made by Gadnery-Medwin (1995) and Davies (2002) who introduced confidence in a scoring procedure. Their studies will be reviewed later in section 2.4. A summary of alternative methods that discourage guessing and give examinees the benefit of partial knowledge are well presented in Ng & Chan (2009).

Difficulty with assessing a method used by a test-taker is also in the list of CAA limitations. In paper-and-pen version of a test sometimes a working paper or a requirement to demonstrate a solution path could provide insights into test-takers’ thinking. Giving partial credit or a method mark in CAA is a very difficult issue which requires additional research.

Assessing higher level outcomes by means of CAA is considered another challenge for test design. A vast number of different ways to express a correct logical proof to an equation or a theorem makes it virtually impossible to recognise all correct versions. This either leads to limiting the range by offering a list of fixed expressions as operators, or results in an error in response evaluation.

The introduction of computers in assessment is motivated by the need to improve assessment processes by maximising evaluation precision and fairness, and minimising the costs. Whereas the cost of teaching is more related to the hours per course and is not greatly affected by how many students attend a lecture, the cost of assessment is a cost per student, which grows proportionally with student numbers. During the last 10-15
years the cost of assessment in higher education was the most rapidly growing component of tuition fees (Ricketts, Filmore, Lowry & Wilks, 2003). In spite of the hopes for reduced costs, the overall cost of CBA exceeds the cost of a paper-and-pen test due to high test design and development expenses. The cost-effectiveness of CBA can be achieved, however, with the economies of scale, which in turn stimulates the prevalence of large-scale summative tests.

Overall, most researchers agree that computerised assessment has more pedagogical and administrative advantages than disadvantages and properly constructed and implemented CAA schemes can be highly beneficial to institutions.

Considering additional sets of meta-cognitive skills that are involved in computer-based activities, such as internet information search skills, the software user skills, teamwork skills, and a set of particular skills that are specific to the learning objectives of the subject, the evaluation of performance is to comprise new dimensions of human behaviour. Most current Learning Management Systems, like Moodle or Blackboard, offer statistics of students’ usage of online resources, such as participation in an online forum or discussion, number of visited pages with the course content, download of learning objects and other parameters of a person’s activities additionally to results of testing. Testing, in its turn, is to expand its diagnostic facility to provide additional information about the current cognitive state of a person and about longitudinal dynamics of cognitive development. The search for the most efficient forms of test content delivery, scoring procedures and the scope of evaluation motivates further research in existing and the development of new types of testing.

There has been a broad discussion about equivalence of test score on paper-and-pencil vs. computerised test with some very contradictory conclusions. Clariana and Wallace (2002) found that gender, competitiveness and computer familiarity were not related to the performance difference, though higher-attaining students benefitted more under computerised versions of the test. Akdemir and Ogus (2008) examined the scores of 47 undergraduate students on a test that was delivered in computerised and paper-and-pencil format. The findings suggest that there was no statistically significant difference between the scores.

In general, there seems to be an agreement that a well-developed computerised test meets the requirements for educational testing and provides measurements at least as valid as a paper-and-pencil one. While choosing a testing method, one has to take into consideration a number of factors, such as efficiency, precision, innovation and
other benefits of a computerised test on one hand, and its cost and complexity of implementation on the other.

2.2.1. Types of computer-based testing.

There are several classifications of types of computer-based testing (CBT) with a different focus on one of the principle features of testing: item content, item selection, scoring system and provided feedback.


Computerised Fixed Tests (CFT) are the most direct analogue to paper-and-pencil tests as the content and the length of a test are firmly fixed. In some literature the term CBT (computer-based test) is used to refer to that kind of test. Other authors (Visspoel, Rocklin & Wang, 1994; Wise & DeMars, 2005) prefer the term Fixed-Item Test (FIT) to contrast the mode of test content selection with adaptive tests.

As in a paper-and-pencil test, in CFT test takers are usually allowed to review and revise their answers. There is a fixed test time limit for all test takers. CFT can benefit from enriched item content when innovative items, based on multimedia or interactivity are included in the test. In general, CFT development and implementation is easier and less costly than of other types of a test (Parshall, Spray, Kalohn & Davey, 2001, p 92). The authors suggest that the tests are suitable for short exams where the individualised content selection is not required, for low-stake exams where security is of minor concern, and for small scale testing where there is a small number of skills to test or a limited number of test takers.

Automated Test Assembly (ATA) for Online Administration extends features of CFT by introducing variation into the content within the constraints of a test, such as test length, targeted cognitive domains or equitable form difficulty. There are two different mathematical concepts to approach the content optimisation problem: Linear Programming concept and Heuristic concept (van der Linden, 1998).

Scoring procedure employed by ATA usually use a number correct approach in a fixed length test. This type of score is easily interpretable; however the validity of a score is greatly determined by the extent of equality of parallel versions. Contrary to CFTs which use parallel items to provide parallel versions of a test, ATA defines parallel tests from the point of view of distribution of individual answers on some scale of difficulty. A similar mean difficulty and variability for each test taker is regarded as a
proof of the test versions being parallel. With this major departure from classical
definition of parallel tests, test reliability becomes a function of internal consistency
rather than objective criteria, external to the testing system, and thus heavily relies on
statistical methods for its validation.

Computerised Adaptive Tests (CATs) are radically different from the previous
type of tests in that they dynamically determine the following item depending on the
answer to the previous item. Test content and test length will vary from one test taker to
another. The most common approach to item selection and scoring is based on Item
Response Theory (IRT), with the underlying assumption that a test taker’s response can
inform the system about the position of the test taker on some scale of one-dimensional
latent ability. There is an ongoing debate about the method’s efficiency and reliability
that involves polar suggestions. There are some claims that under ideal conditions IRT-
based CAT can produce a reliable measurement of the construct under testing with
about the half the items of a fixed-form, non-adaptive test (Parshall, Spray, Kalohn &
Davey, 2001).

A sub-type of CAT is a self-adapted test (S-AT), which was initially developed
by Rocklin and O’Donnell (1995). An item pool is divided into several difficulty levels
and an examinee begins a S-AT by choosing the difficulty level. Feedback on the item
is given immediately and then the examinee chooses the difficulty level of the next
item. The examinee’s overall test performance is calculated using an IRT-based
proficiency estimation method on the completion of the test. S-AT is reported to have
an advantage over CAT in increasing the motivation of examinees and diminishing their
anxiety toward the test.

Computerised Classification Test (CCT) may be considered a sub-type of a test
as it can use any of the test approaches described above and is different only by a
scoring outcome which uses categories (pass/fail) instead of a numerical scale.
Classification tests are mainly used for certification purposes and usually involve high
stakes. Other terms that are used to describe a similar test situation are Criterion-
Referenced Measurement (CRM).

Another major distinction between test types is based on the purpose of the test.
Assessment Of Learning in the form of testing for external reporting purpose is usually
referred to as summative. As a rule, the testing is carried out at the end of a learning unit
and its evaluation is resulted in a score that attempts to quantify the learner’s
knowledge.
A major shift in learning theories from a static behavioural model where an individual was treated as an empty vessel to be filled with knowledge, to a multidimensional, dynamic cognitive model of learning where individuals are informed and actively involved in a learning process, motivated a change in the assessment paradigm from Assessment Of Learning to Assessment For Learning, and also As Learning (Gardner, 2005). Summative tests as a means for providing the accountability of an individual, a school or an educational policy are gradually giving way to on-going formative assessment that would provide parents, teachers, and principals with a tool that can help to understand and address the specific academic needs of students. There has been a wide discussion about the ways to obtain a diagnostic value from subscores on a summative test. Sinharay (2010) indicates that subscores have to satisfy strict standards of reliability and correlation. Based on Sinharay’s results, the subscores have to consist of at least 20 items and have to be sufficiently distinct from each other. Subscores with 10 or less items were not of any added value. This demonstrates that by their nature summative tests cannot provide individual examinees with diagnostics of their strengths and weaknesses in a particular cognitive domain.

An alternative method for testing, formative testing, is concerned with specific skills rather than a general score and aims at developing a skill rather than evaluating. Skills-based formative assessment aims at supplementing test scores with a skills profile that indicates mastery of particular skills, which could produce interpretive, descriptive, and diagnostic reports for each individual student. This can lead to immediate test-driven changes in instruction for the just-tested classroom and future changes in instruction for future classrooms in the same subject (Stout, 2003).

The significance of the current study for advancing the field is also in examining the potential of the measurements of temporal behaviours in computerised formative tests, as well as summative, to inform teaching about different aspects of cognitive process.

2.2.2. Assessing problem-solving.

Inquiry into problem-solving processes has long been a focus of cognitive research. With the broad introduction of ICT in education the feasibility of the measurement of underlying cognitive processes advances the research from theoretical discussion to practical issues of implementation.
An approach to teaching students the knowledge of standard algorithms for solving particular types of tasks is gradually shifting to recognising a value of adaptive expertise which is the ability of individuals to flexibly and creatively produce a meaningful solution strategy (Baroody & Dowker, 2003). The purpose of instruction in this case is to enable students to use what they learned to invent effective procedures for solving new problems. Baroody concluded that adaptive expertise depends on conceptual understanding and its integration with procedural knowledge.

The importance of learning a metacognitive skill of recognising the applicability of a skill outside a given domain is getting a wider popularity among educators. It can be noted that this trend may be a reflection of the rise of industry demands that staff be flexible and creative. The transferability of knowledge and skills between domains becomes a significant factor in the quality of the workforce.

The lack of agreement in understanding the theoretical basis of the adaptive expertise notion can be demonstrated by analysis of the study by Heinse and Lipowsky (2008) who attempted to examine the relationship between accuracy and adaptivity, which they consider the two dimensions of strategy competence. The study evaluated efficiency of the strategy chosen by a third grade student on the scale of being inappropriate, inefficient or efficient according to normative rating. They found that on average a student was able to find an appropriate strategy for 74% of the items. The authors attempted to correlate accuracy with efficiency of the strategy. The approach to treating different strategies on a dichotomous scale seems to be very indicative of a gap in the research in mathematical thinking. It could be expected that the existence of a number of different strategies to solve a problem would prompt an attempt to build a classification system with a finer gradation then just two: efficient/inefficient. Situating a strategy in a multidimensional space of individual abilities, computational skills and learning environment would be a fruitful area of future research. Additionally, it can be argued that adaptivity skill can be defined as the ability to choose the best strategy out of the pool of available strategies given the task nature and learning environment demands. In this case accuracy should be viewed as one of the indications of the right choice and so be a dependant variable of adaptivity.

In recent years, there has been a significant shift in understanding of adults’ approach to simple and complex mental arithmetic problem solving. Contrary to past theories that postulated the fixed choice of a solution strategy in an adult and a major prevalence of memory retrieval over solution strategy use for simple arithmetic (e.g. Ashcraft & Battaglia, 1978), new evidence suggests that adults may maintain a mixture
of procedures similar to children. This new understanding would allow for theoretical models that stress the continuity of development (Shrager & Siegler, 1998, LeFevre, Smith-Chant, Hiscock, Daley & Morris, 2003).

LeFevre, Smith-Chant, Hiscock, Daley and Morris (2003, p. 219) suggested a Multiple Procedure Perspective for understanding mental arithmetic. This perspective attempts to account for the large and persistent individual differences that exist in arithmetic performance. It suggests that to understand the sources of individual variability in arithmetic performance we need to investigate procedure selection.

Chung and Baker (2003) explored the feasibility of using a “click-through” interface to measure the “generate-and-test” problem-solving process for a design problem. The participants were offered a set of on-screen controls, interaction with which would lead to composing a required design of a mechanism. The interface allowed measuring intentional acts that would reflect thinking. The study measured the interaction details and inferred problem-solving strategies. The study reported on observing stable behaviour sets and suggested that monitoring problem-solving processes by means of measurements of student-simulation interactions is feasible and promising.

Thus the overwhelming evidence suggests that development for both children and adults can be defined as a gradual transition from the use of strategies to perform mental computations to the prevalent use of memory retrieval. As non-retrieval procedures are slower than retrieval and produce a higher rate of errors the transition provides a basic mechanism for automaticity and allows achieving an expert level of knowledge. This study depends on those findings as justification for the use of response time as an indication of the position of a learner on the skill development continuum.

2.3. Response time measurements in testing

Response time (RT) is considered the most important measure used to investigate the hypotheses about mental processing in cognitive psychology (Eysenck & Keane, 2005). Response time measurements are also collected routinely in empirical investigations in biological, social, developmental, and clinical psychology for many reasons, including the usability of the measurement which means that response time can be processed as a scaled measure by available statistical analysis procedures.

In the field of education psychology substantial research has been focused on defining the general attributes of RT and examining the nature of the within-person and
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

between-person variance through establishing the relationships with other phenomena, such as general intelligence, accuracy, abilities, attention, memory and emotional state.

In educational testing RT was investigated mostly in respect to the potential to increase the efficacy of evaluation and to inform test design, as a means to improve summative assessment. Relationships between RT and gender, age, education level, language proficiency and nationality are being well explored.

Less number of studies concerned RT in formative assessment, which can be explained by a former dominant role of summative tests. Very few studies focused on the meaning of RT for teaching practice, or on teachers’ perception of the useful content of the output of data analysis. The potential of monitoring changes in response time of an individual and a group of learners over a learning period to inform teaching or curriculum design seems to be a fruitful area for future research.

A brief overview of the research in response time where it is relevant to the current study is presented below.

2.3.1. **Validity of response time measure.**

The utility of response time for analysing data from criterion-referenced tests for the purpose of assessing the quality of the tests and for improving the efficiency of evaluation of examinees’ abilities was put on the research agenda by Tatsuoka and Tatsuoka (1978). The researchers examined regularities of the time it takes examinees to respond to items on matrix algebra tests. The results suggested that RT data can be a valuable addition to performance scores. The study also found that RT data can be described by a Weibull distribution.

Schnipke and Scrams (1997, 1999), using Thissen’s Timed-Testing model as a framework for their research, examined the data from computer-based tests of verbal, quantitative and reasoning skills involving 7,000 examinees. They observed the correlation between examinee ability and slowness being relatively high for verbal- and quantitative-reasoning tests ($r^2 = .39$ and .33, respectively) but non-existent for the analytical-reasoning test. It was suggested that the relationship between speed and accuracy depends on test context and content. The study concluded that question response time and accuracy statistics measure separate constructs of learning and should be used as separate measures of question functioning.

The results of a previous study (Gvozdenko, 2005) also indicated that response time is a separate variable with a weak relationship with accuracy (Pearson correlation = -0.33). This means that questions that required a longer time to complete also tended to
result in a slightly higher percentage of incorrect answers. Descriptive statistics of response times on a test question were suggested as a stable parameter of the question functioning when the number of test takers exceeded 30 people. The potential of using Mean Question response time (MQRT) in combination with response times distribution analysis to deliver additional information in the course of monitoring test takers’ temporal behaviours and evaluating the equality of test questions is further examined in this thesis.

The results of Embretson and Prenovost (2000, p. 861) strongly support response time measures as providing new information about individual differences. They further suggest that response times probably should not be combined with test accuracy. Because response times provide distinct information, combining them with ability would provide a confounded index.

### 2.3.2. Subgroup bias.

Schnipke and Pashley (1999) investigated methodology for assessing subgroup differences in response-time rates on the basis of response times distribution analysis using survival-analysis methodology. For a number of items in this study, significant differences between the response-time distributions of the two subgroups investigated were found. This could be expected since having low English skills caused longer answering times. The authors suggested that future research could examine the response-time curves produced to determine whether any preset difference levels between the curves have been exceeded. This recommendation was considered by the current study which employed the analysis of distributions to compare equality of parallel items.

Chang, Plake and Ferdous (2005) examined whether the time the examinees spend on a test item is gender-biased. Analysis of response time as a function of ability level and item score for male and female examinees did not find any significant difference in respect of gender.

Wallace and Clariana (2005) examined student performance on computer and paper tests. Pre-test/post-test design allowed controlling for general cognitive abilities and investigating the impact of the course as a covariate. Data analysis indicated that students scored slightly higher on computer versus paper administration. Further breakdown by gender revealed that though the female group, whether tested on paper or online, scored below the males on the first test, on the final exam the female students in the computer-administered test group over performed the male counterparts. The study
suggested that initial performance deficits may have resulted from previous lack of exposure to computers and software.

The relationship between race and the time spent on a test item has been a subject of a number of studies since mass introduction of testing in education. The concern about the race-dependant impact of time constraints on the test score was expressed by Schmitt and Bleistein (1987) who investigated factors, affecting differential item functioning for black examinees on scholastic aptitude test analogy items. The study found that Black students appear to need more time to complete the SAT verbal sections than White students with comparable SAT verbal scores. A set of factors that had a significant impact on time included item position, item difficulty, subject matter content and level of abstractness.

Hoffman and Schraw (2009) investigated the influence of self-efficacy beliefs and working memory capacity on mathematical problem-solving performance, response time, and efficiency. They found a significant effect for self-efficacy on problem-solving performance and efficiency, but limited effects for time. According to their findings, response time, though directly correlated with self-efficacy, was dependant on increased problem-solving efficiency through strategic performance. This supports other research that found no relationship between higher motivation and higher performance speed, when solution strategies are controlled.

2.3.3. Utility of response time.

Responding to the need for more efficient approaches to testing a number of studies in the area of test takers’ behaviours was done with the focus on the potential of RT to inform test design and teaching of underlying learning processes that were unobservable in the pen-and-paper mode of test administration. The value of RT for monitoring general progress of test-takers through a test to determine a fair test length, test-takers effort and motivation; individual RT on a test question to detect pre-knowledge or guessing; and the potential use of mean RT to confirm equality of parallel items were in the focus of the studies. The potential of RT to inform item selection process in adaptive IRT-based testing received the most emphasis of the research but with a very limited practical implementation, which can be partly explained by the significant technical complexity of the issue.


2.3.3.1. Monitoring Temporal Behaviour across the Test.

Schnipke and Scrams (1996) investigated speededness of a test which is measured by the proportion of test takers who do not reach each item on the test. Time limits directly affect test takers' performance, especially on a high-stake test. Additional strategies can be used by test takers on a number-right scored test because if no points are subtracted for incorrect responses, test takers tend to guess on items rather than leave them blank. Development of the procedures that could detect that kind of behaviour and discourage it would have a positive impact on test efficiency. Schnipke and Scrams found that the "rapid-guessing" component of speededness can be estimated by modelling response-times with a two-state mixture model.

Bridgeman and Cline (2004) investigates time demands on a time restricted, fixed-length CAT and expressed concerns that a test can be differentially speeded for high able candidates due to the item selection algorithm’s delivery of cognitively complex, more time demanding items to these examinees.

The conclusions of Chang, Plake and Ferdous (2005) are consistent with that concern. However, they found that high ability examinees spend more average time not only on operational items, which could be explained by higher cognitive demands, but also on the pre-test items, which are not tailored to the examinees’ ability level than do lower ability examinees. Therefore, the authors suggest that higher ability examinees might have a higher persistence with test questions, regardless of the item’s difficulty level which results in a longer response time.

Gvozdenko (2005) examined patterns of behaviour and found four distinct pacing patterns: nearly a quarter of test takers were accelerating during the test, another twenty-two percent of students were decelerating, thirty-one percent of test takers were going faster in the middle, and the last quarter were going slower in the middle. A similar proportion of accelerating and decelerating test takers had been also found by Schnipke and Scrams (1999).

2.3.3.2. Detecting Cheating due to Pre-knowledge.

Linden and Krimpen-Stoop (2003) used a lognormal model for response times to check response times for possible irregularities. The model considered responses and response times as independent variables for a test taker. One of the aberrances that were found is related to pre-knowledge of an item. The author suggested that the detection
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

rates for the Bayesian checks can outperform those for the classical checks, but at the cost of higher false-alarm rates.

Meijer (2006) suggested employing response time data analysis to prevent cheating on a test when some test takers have “advance item knowledge”. He proposed to compare RT of a test taker with “effective response time” that is the time required for a test taker to answer an item correctly. An unusually short response time relative to the expected effective response time was suggested as an indicator of item pre-knowledge. Sensitivity of the detection rate to the magnitude of the reduction in RT was found to impose a limitation on the method.

There were prior attempts (McLeod and Lewis, 1999) to use RT to investigate the possibility for detection of item pre-knowledge and item memorization. However, the low detection rates and high false-alarms rates would prevent the method from being widely used.

Another concern of test administrators is a form of cheating when one test taker attempts to copy the answers from another. It may happen as a part of pre-arrangements between test takers or, as sometimes tests are administered in a computer room with a short distance between computer screens, it is possible for a test taker to peek at another’s answers without getting noticed by a source or a test supervisor.

Currently used analysis of the agreement between the response patterns of the suspected test takers (Sotaridona, van der Linden, & Meijer, 2006) uses the procedures that match correct and incorrect answers separately. It is quite obvious that matching correct answers is limited by the argument that correct answers of two high-abilities test takers are supposed to match and cannot be considered a proof of collusion. Matching incorrect answers has a higher validity but allows catching mostly unsuccessful cheaters who would probably get a low score anyway. Besides, it is a viable option that some test takers may have produced similar patterns of wrong answers because of similar misinformation through common instruction.

The latest research suggests that RT can be used to flag occurrences of collusion. van der Linden (2009) found that analysis of the joint distribution of the response times on a test can be used to check for possible collusion between test takers. He suggested a bivariate lognormal model that could capture the effects of the differences in time intensity between the items and the speed at which the test takers operate during the test.
2.3.3.3. **Evaluating Examinees’ Effort and Test Motivation.**

The administration of low-stakes achievement tests and formative tests often faces the issue of low motivation of test takers. If there are no consequences for a test-taker’s performance record their effort might be insufficient to demonstrate the actual level of proficiency. There is a body of research investigating the ways to measure the efforts and evaluate the validity of the test results accordingly.

Haladyna and Downing (2004) suggested that low test-taking effort introduces construct-irrelevant variance that threatens interpretation of test scores and so constitutes a threat to test score validity. Wise, Owens, Yang, Weiss, Kissel, Kong and Horst (2005) expressed concern about the validity of low-stakes testing results as determinants of the performance of an institution due to the conflict between serious potential consequences of test results for an institution and the lack of motivation for an individual examinee. The study pointed out that examinee effort should be taken into account when high-stakes testing items are trialled in low-stakes testing environment. Another warning was directed to researchers in measurement area who usually rely on volunteers and should be cautious about the validity of the measurements. Wise and DeMars (2005) synthesized the findings from a set of studies and found a mean motivation effect of 0.58 standard deviations.

Wise and Kong (2005) and Kong (2007) showed that item response time can be a useful measure of examinee effort. If RT on the item is too short to allow for required cognitive processes to happen, namely for reading the task, applying a solution strategy and filling in the answer, then the behaviour is diagnosed as rapid-guessing behaviour, which is one manifestation of lack of effort. The authors proposed to calculate examinee’s RTE score as the proportion of items for which a test taker’s RT was consistent with the time required to achieve a solution (solution behaviour). In their study a response time threshold of 10 seconds was used for all items for differentiating rapid-guessing from solution behaviour.

Kong, Wise and Bholak (2007) also investigated RT threshold to differentiate solution behaviour (SB) from rapid-guessing behaviour. This study compared four methods for setting item response time thresholds to differentiate rapid-guessing behaviour from solution behaviour: (a) setting up the same threshold for all items; (b) conditioning the threshold by an individual item content; (c) visually inspecting distribution of response times; and (d) using multi-state mixture modelling. The four methods yielded very similar results, with simple examination of response time distribution being sufficient means for determining the threshold values.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

Setzer and Allspach (2007) described potential effects caused by low-effort and the analytic techniques that have recently been developed to compensate for such effects. The study used the data from CBT version of the Major Field Test developed by ETS. The data included the response time (in whole seconds) for each examinee-by-item interaction. Thresholds for assessing SB were determined by examining the response time distribution for each item as had been suggested by prior research. The threshold was placed at the point on the scale where a rapid-guessing distribution intersected with a solution-behaviour distribution. In one of the examples the threshold was equal 6 seconds.

In most studies that engaged the concept of RTE measure the solution behaviour was treated in a dichotomous manner. Setzer and Allspach suggested further exploration of these phenomena to determine the usefulness of measuring solution behaviour on a continuous scale of effort. They suggested that an additional area of research in which response time could be explored should focus on the cognitive processing of the examinee, which would enrich our understanding of the interaction between the examinee and the test items.

2.3.3.4. Complementing Accuracy.

A number of studies investigated the relationship between response time and the probability of a wrong answer. The evidence was quite contradictory, with some studies establishing a strong relationship between a delayed response time and the probability of a wrong answer (Bergstrom, Gershon & Lunz, 1994; Hornke, 1997; Rammsayer, 1999, Hornke, 2000) and other studies (Gavira, 2005) having mixed results.

Hornke (2000) found that examinees spent more time in a CAT on items they answered incorrectly than on items they answered correctly. Chang, Plake and Ferdous (2005) investigated response times for correct and incorrect item responses on computerised adaptive tests. The study used 10,000 examinee records from international CAT admission tests administered worldwide in 2002. As most operational adaptive tests contain two groups of items, operational and pre-test, with only the former being tailored to an examinee’s ability level, the study was able to correlate response times on pre-test items with ability levels and background characteristics, such as gender and US citizenship. Examinees were grouped into 6 performance categories based on their overall proficiency estimation. Average response times by ability groups were then computed and compared for correct and incorrect responses. They found that “lower ability candidates systematically spent nearly equivalent time on items that they answered correctly and incorrectly. Higher ability candidates, in general, spent more
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

time averaged across items they answered incorrectly than ones they answered correctly” (p. 6). There was a significant trend across all six levels of ability that more able students spent more time on all items, regardless of whether the items are answered correctly or not. These results were consistent for male and female examinees and for US and non-US administration sites. In the interpretation of the results the study suggests that more able candidates are more likely to persevere on test questions, whereas less able students may recognize that the difficulty level of the question exceeds their knowledge and more quickly respond and move on to the next test question.

Gavira (2005) investigated the association between response time and the relative difficulty of items for the person examined under the framework of adaptive IRT-based testing. The study obtained data from 634 subjects whose age ranged from 15 to 33 years. The test content comprised 21 items in the cognitive area of “Rotation of Solid Figures”. The study developed a psychometric model that relates RT for each item with the capacity that the item measures and with its difficulty and discrimination parameters, and tested the model with the simulated data and real data from the experiment. The study found that the model is a feasible instrument for the analysis and that including response time data into the model increases correlation with the original data. The results suggest that “the use of response time provides as much information or more than the use of very informative *a priori* distributions” (Gavira, 2005, p. 67).

González-Espada and Bullock (2007) examined the relationship between RT and accuracy on answering Multiple Choice (MC) format questions. This research focused on items that were used for reinforcement and review during lectures. The results were then correlated with other variables such as gender, grade point average (GPA), America College Test (ACT) scores, and final course grade. The study found that (a) for one third of the participants there was a significant difference between the average time for items answered correctly and incorrectly; (b) average time on correctly answered items was negatively correlated with ACT scores; (c) average time on incorrectly answered items was not significantly correlated with other variables.

It can be argued that the results of the study, described above, were impacted by the mode of test delivery, as the experiment was using a classroom response system which implies a nearly synchronous communication and forces the test takers to submit an item. In this case incorrect responses might be submitted within the same timeframe as correct responses and thus did not bear the temporal characteristics that could be expected in the asynchronous mode of a regular test.
2.3.4. **Relationship between practice, automaticity and response time.**

In order to provide justification for operationalising the quality of mathematical knowledge with RT and certainty measurements it is necessary to review theory and prior research related to the relationship of practice, strategy use and automaticity with performance speed and confidence.

Investigation of the relationship of responses and response times goes as far back as the beginning of psychometrics with the works of Thurstone (1937). Thurstone’s enquiry attempted to answer the question “whether high speed in the performance of easy tasks can be taken as indicative of the ability to perform more difficult tasks without pressure of time, the form of ability that regarded as the socially more valuable” (p. 249). In a broader sense, Thurstone expressed an idea that the ability to generate a correct answer can have different quality of knowledge, underlying the ability, and thus the approach to measuring the performance on a dichotomous scale (correct/incorrect) greatly limits the amount of information obtained from testing. Practice, according to Thurstone, improves performance, ensuring correctness and decreasing response time.

Practice is universally recognised as an integral part of learning. Practice advances a novice to an expert level by increasing accuracy and reducing the time required for task completion. There is no doubt in the direct relationship between practice and knowledge improvement and there is a commonly accepted agreement on the inverse relationship between practice and performance speed on most tasks. However, the rate and the shape of improvement in different cognitive domains is still under investigation.

The empirical law of practice, also known as log-linear learning law, appears to be first introduced by Snoddy (1926, as cited by Newell and Rosenbloom, 1981). The power law of practice stipulates that the logarithm of the reaction time for a particular task decreases linearly with the logarithm of the amount of practice. Newell and Rosenbloom (1981) investigated the occurrences and potential explanations for the law as it applies both for group and individual data. They concluded that “the major phenomenon is the ubiquity of the learning data (p. 52)” with a single family of generalised power function learning curves applicable to all types of cognitive behaviours. The authors further suggest that if humans learn by creating and storing chunks, then the environmental exhaustion effect would also operate to produce power-law learning. Though, they mentioned that “obvious deviations at the ends of empirical curves were eliminated before the fits … were computed. The equations therefore primarily represent this linear portion of the curve.” (p. 5), they insisted that the
empirical curves do not fit the exponential family as their tails are genuinely slower than exponential learning.

The fit of the model was later questioned by Heathcote, Brown and Mewhort (2000) who suggested that the resulting fit could be merely an artefact of statistical averaging. According to their mathematical modelling, exponential curves provide a better fit to describe the data.

From the point of view of our study, the discussion about the best fit model is also important because it creates a fundamental difference in our understanding of learning. If learning follows an exponential curve, then learning progress should be viewed and accordingly evaluated under an assumption that it is based on a fixed percentage of what remains to be learnt. If learning follows a power law, then learning just slows down according to a logarithmic representation of learning attempts. An investigation of the limits of learning in different domains and the role of individual differences in determining the limits would be a very fruitful avenue for further research.

A desirable effect of practice is manifested in reduced time and increased accuracy which is referred to as automaticity. The development of automaticity in simple arithmetic and the relationship between automaticity and response times were examined by a number of studies. Klapp, Porter-Graham and Hoifjeld (1991) used the term automaticity to define the switch from counting to memory retrieval to solve problems. The evidence suggests that automaticity is also characterised by different levels reflecting a growing strength of association between a stimulus and a response. The authors used the term “training beyond automaticity” to reflect effects of the practice after retrieval (automaticity) had been established.

Those findings establish the grounds for treating response time reduction as a valid indication of learning progress thus justifying the use of RT measurements as an operationalized representation of the progress.

2.4. Certainty Measurements in Testing

One of the long-standing problems in educational measurement is the assessment of partial knowledge which is the knowledge that allows a student to diagnose some alternative responses as incorrect but is insufficient to reach a firm conclusion about the correct answer. The problem arises from limitations of the dichotomous accuracy evaluation in the format correct/incorrect being incapable of differentiating between incomplete knowledge and ignorance (Jacobs, 1971; Bennet and Gitomer, 2009, p.43).
There are two distinctive schools of thought approaching the problem from different directions: inferring partial knowledge from a combination of correct and incorrect answers (Coombs, 1953), or asking a student to self-assess their certainty (confidence) in the correctness of the answer (Trow, 1923). The latter is relevant to the current study and will be discussed in more detail in the following text.

The nature of self-confidence, the reliability of confidence measurements, the relationship between decision time, accuracy and confidence of individuals, and conditions where people exhibit underconfidence or overconfidence have for decades been a subject of vigorous academic discourse (Trow, 1923; Johnson, 1939; Jacobs, 1971; Kleitman and Stankov, 2007).

2.4.1. Scale of measurement.

The issue of selecting the right scale of certainty measurements was addressed in a number of works (Shuford, 1968; Leclercq, 1983; Sinkavich, 1995; Bruttomesso et al., 2003) that produced a range of diverse recommendations.

The first discussion of the requirements to successful implementation of confidence testing was initiated by Shuford (1968, p. 2) who “spelled out the conditions under which confidence testing could yield large gains in personnel selection, classification, training, and education”. Shuford administered a confidence test to military personnel and examined additional value created by introducing the element of self-evaluation. The scale that was used in the study was comprised of all letters of the English alphabet, with letter Z assigned to the highest certainty in correctness of an answer. Shuford found that more answers marked with Z certainty were correct than the answers marked with other letters and that there was a correlation between the degree of certainty and correctness, which confirmed internal validity of the measurements. However, the researcher also observed a range of individual differences in the usage of the certainty scale, with some test takers using only two letters: A and Z, to indicate their certainty. The small sample size of the study did not allow making generalisable conclusions but mostly acted as a pilot study which re-ignited the interest in this type of testing.

Further investigation into the most efficient scale of certainty management was attempted by Leclercq (1993) who examined the questions of the human sensitivity (how many degrees should be used), realism (validity of self-estimation) and reliability (stability in time) as determinants for the scale. A more recent study (Bruttomesso et al., 2003) applied the scale that was initially developed by Leclercq (1983), to measuring
the certainty of medical patients. According to Leclercq, human perception has a logarithmic distribution at the extremes of the scale and is limited to 7 degrees. In the procedure of measuring certainty, patients were asked to indicate their degree of certainty using a 7 degrees scale ranging from total doubt to complete certainty as following: totally unsure (2%), not sure (10%), moderately sure (25%), more or less sure (50%), sure (75%), strongly sure (90%), perfectly sure (98%).

An alternative scale on the basis of five categories (Likert scale) was proposed by Sinkavich (1995) who used it in educational settings similar to our experiment. Students of a university educational psychology section were asked to select their multiple choice answer and then to rate their confidence in that answer by marking as "not correct" (-2) to "correct" (+2). The midpoint of the scale was interpreted as “maybe it is correct; maybe it is not correct”. Sinkavich found that confidence varied with performance and under some conditions the correlations between confidence and performance could be as high as 0.65. It was observed that when students were actually correct, they indicate that they were very confident 45% of the time and somewhat confident 32% of the time. Sinkavich concluded that students might be using the rating scale differently. He suggested that a scale rating of +2 to one student might mean "this is definitely correct"; while to another it might mean, "I'm as sure as I am for any item".

Overall, the disagreement between the scales used by different studies resulted in the need for the present study to address the issue as one of the research questions.

2.4.2. Individual and subgroup differences.

The evaluation of the impact of individual differences on the choice of certainty categories has been a long standing issue of certainty measurements. Understanding the role of an individual self-confidence factor grew from the initial suggestion of Trow (1923) that when accuracy is controlled “no evidence to support the notion that certain individuals were generally more or less confident across tasks” (as cited by Blais, Thompson & Baranski, 2005) to an opposite view of most of the recent researchers, for example Kleitman and Stankov (2007) suggesting a general self-confidence factor as a personality trait.

Subgroup differences are another subject of investigation. According to Sieber (1979), secondary school students exhibited consistent subgroup differences in confidence estimation on spelling, mathematics and reasoning tests. High achieving students demonstrated a higher accuracy of confidence estimation and slightly tended to underconfidence. Additionally, the study found that the participants who were trained
by a teacher to express warranted uncertainty and extensively used a confidence estimation procedure performed better on two tests of spelling, and improved more from Test 1 to Test 2, than did participants who merely used or did not use the confidence estimation procedure. The author also found that expression of certainty depends partly on whether it is costly or rewarded to do so. This finding should be taken as a warning about potential threats to the reliability of certainty measurements. As with any measure based on self-assessment, certainty may be vulnerable to the bias introduced by student motivation.

Subgroup differences were also reported by Newman (1984) who examined the relationship between basic numerical skill and confidence on a task of quantitative estimation. They found that counting skill was related not only to accuracy but also to appropriateness of the confidence, with the relationship being stronger for skilful test takers.

One of the most prominent papers about the relationship between knowledge of a domain and ability to assess one’s own skills was published by Kruger and Dunning (1999). The study examined the links between competence, metacognitive ability and inflated self-assessment in such diverse domains as logical reasoning, English Grammar and humour in the cohort of undergraduate students from university courses in psychology. The ability to distinguish what one has answered correctly from what one has answered incorrectly and the ability to recognise the competence in others was explored in a series of four studies. Across all studies the participants who were in the bottom quartile by overall test score not only overestimated their own performance in terms of a number of correct answers but persistently marked their ability as above-average. Those participants were also found less successful than top-quartile participants in the metacognitive tasks of discerning what one had answered incorrectly even after controlling for objective performance.

Another discovery by Kruger and Dunning was that top quartile participants also showed some systematic bias, however of the opposite direction. Across the four studies highly competent individuals tended to underestimate their ability and test performance relative to their peers.

In the section on limitations of the study the authors suggest that “in order for the incompetent to overestimate themselves, they must satisfy a minimal threshold of knowledge, theory, or experience that suggest to themselves that they can generate correct answers”. However, there may be another possibility, that this limitation arises from the use of the dichotomous scale (correct/incorrect) engaged by the study for self-
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

assessment. It seems that the design of the study would have benefitted from the use of a wider range of certainty categories, which would allow filtering the lack of knowledge from partial knowledge.

In the study by Kruger and Dunning a gender factor was found insignificant, however, gender bias in expression of confidence has been reported by a number of other studies (Stankov, Lee, Paek, 2009) that found males being generally more over-confident than females.

2.4.3. Application of certainty measurements in teaching.

One of the most discussed applications of confidence measurements for testing is called Confidence-Based Testing (CBA) or just “confidence testing” which incorporates the subjective probability of correctness obtained through self-evaluation into a scoring procedure along with the objective response. The issue of the validity of the subjective self-evaluation is crucial for successful application and there have been a number of different attempts to create a balanced system of stimulus that would encourage an honest expression of the degree of confidence and discourage misleading.

A number of recent studies (Gardner-Medwin & Gahan, 2003; Davies, 2000) suggested that CBA demonstrates some advantages over traditional accuracy-based assessment approaches for a learner, such as:

- developing general study skills of self-monitoring when a student should be able to assess his/her progress in solving a problem;
- developing the metacognitive skill of being able to justify the answer;
- offering students benefit of learning that there are different ways to solve problems, and that efficient learning and rigorous knowledge involves the skill of always testing one idea against another;
- encouraging learners to relate the answer to other knowledge;
- discouraging learners from using a lucky guess option;
- in case of a confident wrong answer, alerting the student to being misinformed.

New affordances offered by certainty measurements in computerised testing were examined by Gardner-Medwin (2006) who suggested that monitoring certainty information can be an invaluable tool for facilitators mapping a route for lesson plans and adjusting subject curriculum to tailor it to the needs of the particular cohort. The author stressed that distinguishing between well-informed, poorly-informed, and
misinformed answers provides opportunities to reveal and target specific misconceptions that would otherwise be a derailing agent to further learning.

Advances in metacognitive and behavioural skills as a result of CBA-based training were reported by the study of Strater et al. (2004). Trained cadets seemed more likely to have the confidence to trust their judgment and analysis of the situation enough to make independent decisions. Developing confidence that is based on careful, critical analysis of the information available was found likely to produce superior performance, especially at the vital higher levels of comprehension and projection.

Additionally to the use of certainty data to classify individual knowledge Bruttomesso et al. (2003) suggested summarising the knowledge in the group. From the combination between degree of certainty and correctness, each answer was assigned to either the area of mastered knowledge (answers given with at least 90% certainty and correct in at least 90% of cases), the area of hazardous knowledge (answers given with at least 90% certainty but correct in less than 50% of cases), the area of uncertain knowledge (answers given with less than 50% certainty but correct in over 90% of cases) or the area of residual knowledge obtained by subtracting the previous areas to the total area of knowledge. The authors developed a graphical format of the three-dimensional raised map for presentation of the data which they called “group certainty topography” that illustrated the distribution of the certainty within a group. The study suggested that certainty measurements can help educators to understand their students and modify their teaching.

2.5. Relationship between response time and certainty

As the present study is operating with three major constructs: accuracy, response time and certainty, and aims at quantifying certainty in order to correlate it with response time and accuracy, it has to be grounded in a solid theory of mental measurements that focuses on the same constructs and operates the same variables as our study, to be able to justify its methodological approach. In our case, The Theory of Judgment (Theory of Choice) provided a framework for investigating the relationship between knowledge, certainty and response time.

The classical psychological theory of judgement, originated with Gustav Fechner’s Psychophysics, underwent continual development from measuring body sensations to measuring psychological objects such as attitudes, preferences, and meanings (Link, 1992). The three primary measures of discrimination used in the field,
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

are response frequency, response time and response confidence which is called certainty in our study.

The relation between response time and response certainty was first systematically investigated and formally described by Volkmann (1934) who empirically derived the formula for relation between “judgment time and response certainty”:

\[ t = \frac{a}{2v - 1} + b \]

where \( t \) is the time of judgement (RT), \( a \) and \( b \) are constants, and \( v \) is the amount of certainty with which the judgment is given.

As can be seen from the formula, if the denominator, containing \( v \), is growing then response time is decreasing.

If to rewrite the equation for solving \( v \) then it will look:

\[ v = 0.5 + 0.5\left(\frac{a}{t - b}\right) \]

From the equation above it follows that if the constant \( a \) represents some nominal time required to solve a problem, certainty (\( v \)) can be inferred from response time (\( t \)).

Johnson (1939, as cited by Petrusic and Baranski, 2008) confirmed Volkmann’s findings of the inverse relation and proposed an alternative, logarithmic relation between confidence and decision time, expressed by the following formula:

\[ T = a(10)^{b(1-C)} \]

where \( T \) is the decision time, \( C \) is confidence, and \( a \) and \( b \) are intercept and slope constants, respectively.

Relationship between certainty and response time is a topic of on-going scientific discourse. Overcoming the limitations of self-reported certainty, such as individual inconsistency, possibility of manipulation, disturbing effects of motivation by inferring certainty from RT which is a more objective measure, would have significant implication for assessment practice. However, the methodology of the use of RT to estimate different levels of certainty appears almost nonexistent. One of the few known works is by Feinstein (2000) who investigated the merits of considering response time as a replacement for certainty indications for the purpose of assessing an expert’s feelings of confidence in expert systems. The subjects in that study reported the level of
certainty on the 1-to-9 scale for each choice-pair. Feinstein found that the response latency method detects consistency of the responses better than the self-report method. The findings suggest that response time can be useful in detecting finer variations in certainty between responses that were reported as equally certain by a subject. The authors concluded that the response latency method might offer a more accurate, more rapid, easier, less obtrusive and less costly approach to eliciting knowledge from experts. However, due to low sample size (n = 21) and the specifics of the population (adults, experts) the findings can hardly be generalised to other settings where populations are usually represented by a mixture of experts and novices.

In the study of response times and misconceptions in science education, Heckler, Scaife & Sayre (2010) found that students answering with misconception-like responses did so consistently and more rapidly than those who answered correctly. In their experiment the prior knowledge about comparing heights on a chart curve was overgeneralised by some students to comparing slopes of the curve. As correctly comparing heights was on average quicker than correctly comparing slopes, those students who had the misconception demonstrated quicker response times on the incorrect responses. The authors used a decision-criterion model to explain the results. According to the model, decision-making happens when accumulated evidence for the choice reaches some criterion value. It is not surprising that, in the absence of more complex knowledge, the prior simpler knowledge is used to infer the previous patterns to a new situation. Interestingly, in the Heckler, Scaife & Sayre (2010) experiment, more students were able to find a correct response when asked to spend more time on the questions. This suggests that some students had at least partial knowledge which enabled them to produce the correct answer.

2.6. Summary

The process of gathering relevant literature and composing this literature review was carried out intensively throughout the course of the study to ensure that the subject of investigation and the approach adopted are in line with current practice in grounded research work. The review of the previous work in the area allowed refining the focus of this study and maintaining the perspective through all stages of the study from study design to data collection to discussing results. A range of secondary data sources were used for identifying relevant work, with the most significant a University electronic library that provided access to online collections of research journal publications. A number of monograms and edited books provided guidance in reviewing fundamental
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

theoretical works in the area. Due to the interdisciplinary nature of the study, relevant publications were found in different academic domains including cognitive science, mathematics and others.

A synthesis of the earlier work aims to provide a holistic view of the approach that uses the philosophy of evolutionary biology as a theoretical guidance for navigating other life sciences. The origin of learning as adaptation ability and the role of individual differences as a driving force of evolutionary advancement for an individual and a community of individuals were reviewed on the primary level of life organisation which is a living unicellular organism. The postulate was advanced that the evolutionary selection and promotion of effective organisational structures and functions is a key to observed uniformity of current organic forms and underlying chemical processes and biological mechanisms.

Individual differences within human specie were shown to have a dual nature, with approximately equal impact of genetic and environmental factors. The place of individual differences in education paradigms of the last half a century was discussed to demonstrate the grounds for the shift from considering individual differences a disturbance to the uniformity of the education process to growing encompassing individual differences as an important beneficial factor in learner-centred education.

The role of computerised testing in education was discussed with a brief introduction to and discussion of the benefits and limitations of major types of tests. New affordances of computerised delivery were highlighted in respect to the availability of automatically generated data of the interactions between a test taker and a test software.

The review of the research in mathematical problem-solving demonstrated a lack of agreement in understanding the theoretical basis of adaptive expertise. The theories of fixed choice of solution strategy in adults are criticised by the proponents of the theory of continuity of development which stipulates that adults may maintain a mixture of procedures based on both memory retrieval and solution strategies.

Studies of response time (RT) and certainty in testing were reviewed in a more detailed way as they form the subject of the present scientific enquiry. Firstly, the research that examined response time and certainty as separate variables was discussed. Then the research about the relationship between those variables was reviewed.

The review demonstrated that RT is generally accepted as an additional measure of performance. The incorporation of RT in summative testing models as an informing agent for increasing test precision was found to be a major motivation behind RT
research. There seems to be an overall agreement about the high potential of RT measurements for informing all stake holders in educational testing. Most authors also pointed at the need for further research in the area.

The evidence about the relationship between accuracy and RT was found to be contradictory, with some studies establishing a strong positive correlation between delayed RT and the probability of a wrong answer and other studies observing no at all or a negative correlation. Additional investigation of the relationship with breakdown by the level of achievement seems to be highly warranted by the discussion.

To establish the significance of other subgroup differences and potential sources of inter-personal variation in RT, such as gender, race, and ethnicity a review of relevant prior research was included in this chapter. The research about the use of RT measurements for monitoring temporal behaviour across a test, detecting cheating due to pre-knowledge, evaluating examinee’s effort and test motivation was acknowledged as an important contribution to the field.

The relationship between RT and the automaticity resulting from practice was another area of prior research that is important to the current study. The major agreement seems to be that RT is negatively correlated with practice. Automaticity is viewed as the shift to prevalent use of memory retrieval as opposite to solution strategies. However, most research is located in the area of cognitive psychology which investigates the relationship on rather simple tasks. Very few works focus on arithmetic tasks and none was found to examine the impact of practice on RT in complex mathematics on a longitudinal scale.

Prior enquiries into the use of certainty in assessment were reviewed to situate our research questions in the context and establish the methodology for certainty measurements that are engaged in our study. It was demonstrated that certainty is generally recognised as an important characteristic of knowledge, and certainty-based assessment was suggested to have some benefits over regular testing, including the ability to inform teachers of the quality of student’s responses and to reveal hidden misconceptions. Certainty was found to be correlated with response probability, which makes it a potentially useful tool, however, due to a number of limitations, such as certainty self-reporting being perceived as intrusive and time-consuming by some students, a very limited use of certainty-based assessment is found in education practice.

Finally, prior research in the relationship between RT and certainty was examined. The equations suggested by The Theory of Judgment were the only expression of the relationship that could be found. Though the formulas were suggested
to be generalisable for other situations of choice, we could not locate any report of the research that would attempt to apply those formulas to education testing settings.

Overall, this chapter provided a literature review for the purpose of establishing the philosophical position of the author and providing a foundation for the research questions under investigation. It offered justification for the choice of the methodology which is discussed in the following chapter, and indicated some productive avenues for critical engagement that will take place in the Results and Discussion chapter.
CHAPTER 3. METHODOLOGY


3.1.1. Scientific method.

A long history of natural sciences enquiries has produced the scientific method for obtaining new knowledge. This method is not disputed by anyone when it is applied to the physical natural world. However, the difficulties in measuring the variables concerning human behaviours led to the discussion whether scientific method can be applied to development of successful social sciences and especially psychology. For example, Kline (1998, p. 3) questions the viewpoints and assumptions of the scientific method, namely realism and reliance on correspondence theory, as being challenged in philosophy. Kline further points at representational weakness of measurement in psychometrics (p. 195). He states that as the phenomenon of intelligence is unclear then the scales of intelligence and other psychometric tests are meaningless. His call to introduce some fundamental measures in psychology to make it equal to other sciences is accompanied with doubt that it will ever be possible. This work is remarkable as a symbol of disappointment in a scientific method. Impatience with limitations of the method sometimes drives researchers from it to a more speculative field of philosophy which casts doubts and questions the foundation of science – holistic realism.

Fortunately, there are other researchers that continue the best traditions of scientific approach to examining the world, both in hard and soft science. A well-known scientist Eysenck (1986) expressed his viewpoint in the following way:

My belief is that psychology is a science in the same sense that physics or chemistry or astronomy are sciences. I believe that psychology should follow the same path as other sciences, namely the construction of theories, the deduction of consequences from these theories and the testing of these deductions…. In other words I believe that we must follow the hypothetico-deductive method of all other sciences if we are to prosper.

(p. 397)

The current study can be categorised as an empirical study that utilises scientific method of investigation based on a data-driven approach to hypothesis generating and further testing by the means of statistical analysis.
3.2. Study design

The investigation of the research questions (a) the meaning and potential of response time data for teaching, test design, curriculum development and cognitive research, (b) the meaning of certainty data for teaching and curriculum development, and (c) the relationship between response time and certainty, and utilizes the quasi-experimental type of design justified by the importance of studying cognitive processes in a natural setting (Cook & Campbell, 1979). Quasi-experimentation is getting wider recognition as part of a general logical and epistemological framework for research (Shadish, Cook and Campbell, 2002).

The current study is classified as quasi-experimental as it was conducted in a real educational setting which is characterised by a dynamic environment with many variables beyond the control of experimenters. The study recognises the potential shortcomings of the quasi-experimental design and has taken a number of steps to maximise the validity and generalisability of the results: the study utilised random assignment principles to avoid selection bias, it maintained an adequate sample size where it was feasible, it validated the measurement scales and the statistics for major constructs to ensure reliability of instrumentation and it followed recognised procedures for statistical analysis.

The population for the purpose of the study can be defined as pre-service teacher education students enrolled in two mathematics subjects. Every year for three years a new sample was compiled which included all students who consented to making their individual data available for the study. Sample size differed from year to year, being in the range 70-90 per cent of the corresponding population and can be viewed for each test in Table 1.

The tests that were used for data collection were a part of a university subject, which ensured that students were motivated to participate in testing. However, the motivation on a summative test could be assumed to be higher than on a formative test due to the higher stakes involved for a student. Reaching a pass score on a summative (hurdle) test was a pre-requisite for successful completion of the subject. On the other hand, just a timely participation in 80% of weekly online tests was expected from a student on a formative test.

To compensate for the potentially distorting effect of anxiety on performance, the students were offered a practice website two weeks before summative testing rounds. A website for practice tests provided unlimited access from campus and from home and, among other benefits, was supposed to help students alleviate any test
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

anxiety. On completion of a practice test the students received the feedback, containing correct answers and a total score. The level of computer anxiety was assumed to be low as all students had taken a computer skills subject in the previous semester.

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

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

3.3. Applicability of psychometric and cognitive models

Specific research questions of the study determined the approach to the study design. The applicability of a number of cognitive diagnostic models or psychometric approaches to the particular setting of the experiment was considered. As the study was conducted in a real-life classroom, with learning as the subject of the investigation, learning theory and cognitive psychology were most closely related to the study. In the discussion of prior attempts by theoreticians and practitioners to incorporate results from learning theory and cognitive psychology more directly into educational measurement DiBello, Stout and Roussos (1995) state that inherent limitations of psychometric theory, such as the latent space approach based on one or only a few latent traits, make the psychometric modelling applicable only to broad composite abilities, like verbal, mathematics, and so forth. The authors argue that “the factor analytic methods are of little help in trying to determine specific cognitive characteristics of examinees for the purpose of instruction, like ‘proficiency with algebraic rules of exponents’.” (p. 363). As the present study was investigating specific cognitive characteristics like skills in mental computation or decimal conversions, psychometric techniques for modelling broad abilities did not match the needs of the study.
Some researchers have attempted to produce blended models that link psychometric and cognitive theories. An example is given by DiBello, Stout and Roussos (1995) who suggested that Fischer’s (1973) attempt to bring cognitive features into psychometrics by introducing Linear Logistic Latent Trait Model (LLTM) was hindered by viewing the complexity factors as ‘compensatory’ in the sense that a lack of one factor can be compensated by the extra weight of the other. A non-compensatory General Component Latent Trait Model, proposed by Embretson (1984) offered a platform for building fixed dependencies between task performance, component task difficulty and component person abilities. This permitted the stimulus features of tasks to be correlated with component task difficulties, thus linking the psychometric model to cognitive theory.

Another step to create a blended model happened with the appearance of the unified model suggested by DiBello, Stout and Roussos, (1993) to bring together deterministic aspects of cognition and stochastic aspects of test response behaviours (DiBello, Stout & Roussos, 1995). The goal of the model was to enable teachers and testers to extract information from tests about cognitive strengths and weaknesses relevant to instructional design and assignment. The authors stressed the need for a test analysis method that could be useful to the practitioner. Likelihood-based diagnostic classification technique was developed on the principles of the unified model. As put by the authors (DiBello, Stout, Roussos, p. 362, 1995), the fundamental difference between their approach and that of Tatsuoka (1985, 1990) was in modelling the sources of systematic deviations from the response behaviour predicted by the deterministic Q-matrix of cognitive attributes.

All the models above require a large data set. Due to a smaller scale and specific research questions the present study could not utilise the full power of the unified model. However, the deterministic approach to an item as an object possessing a set of cognitive attributes as described in the previous paragraph shaped the present approach. The focus on practical utility of the measurements determined the directions for the present experiment.

3.4. Data Description

To answer the research questions the study obtained such data as response times and indications of self-assessed confidence (certainty) in addition to accuracy statistics. The study also interviewed experts in teaching Mathematics and test takers.
Response times were measured as the time between the loading of a web page with a test question and clicking on a multiple choice selector or typing the last symbol in an entry form of a ‘fill-in’ question.

Self-assessed certainty levels were collected by asking test takers to indicate their certainty in the correctness of their answer before moving to a next question. The scale of certainty offered five categories (Figure 1) in summative test S2005 and three categories (Figure 2) in all following summative and formative tests.

Accuracy was measured by comparing a test taker’s answer with alternatives provided by a test designer. It contains the information about correct and incorrect answers of each test taker on each test. Both multiple choice and open-ended “fill in” types of answers were offered in the tests.

Indication of a problem solving strategy was collected on six questions from 20 formative tests by inserting a special research screen in the sequence of test questions. Test takers were there asked to explain how they solved a previous question by selecting from a supplied list the strategy that they had just used. Additionally, there was an option to explain an individual strategy if different from the ones on the list.

Teachers’ thinking about meaning of response times and certainty was collected from 3 teachers, one of whom was also a subject co-ordinator, through a number of formal and informal meetings and discussions.

3.4.1. Data sources.

The data were collected from summative and formative computerised tests in basic mathematics skills in 2005 – 2007 from the following sources:

- Summative test (S2005) of second year education students (n = 186) in August-September 2005
- Summative test (S2006) of second year education students (n = 204) in September 2006
- A series of 20 formative tests (F2006) of first year Faculty of Education students (n=164) in semesters 1 and 2 in 2006
- A series of 10 formative tests (F2007) of first year Faculty of Education students in semester 1, 2007 (n = 125)
- Interviews with second year education students (n = 5) in 2007
- Ongoing communication with teaching experts (n = 3) in 2005 – 2007
Table 1.
Data sources.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of test</td>
<td>Summative</td>
<td>Summative</td>
<td>Formative</td>
<td>Formative</td>
</tr>
<tr>
<td>Student cohort</td>
<td>Second year</td>
<td>Second year</td>
<td>First year</td>
<td>First year</td>
</tr>
<tr>
<td>When administered</td>
<td>Aug 2005</td>
<td>Sep 2006</td>
<td>20 tests in 2006</td>
<td>10 tests in 2007</td>
</tr>
<tr>
<td>Test versions</td>
<td>Some items randomly selected</td>
<td>Three parallel versions</td>
<td>Fixed versions</td>
<td>Fixed versions</td>
</tr>
<tr>
<td>Response times,</td>
<td>v. 1 n = 72</td>
<td>v. 2 n = 69</td>
<td>n = 164</td>
<td>n = 125</td>
</tr>
<tr>
<td>Certainty,</td>
<td>n = 186</td>
<td>v. 3 n = 62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>n = 186</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student interviews</td>
<td>--</td>
<td></td>
<td>n = 3</td>
<td>n = 5</td>
</tr>
<tr>
<td>Teachers’ interviews</td>
<td>ongoing</td>
<td></td>
<td>ongoing</td>
<td>ongoing</td>
</tr>
</tbody>
</table>
3.5. Equipment Used in Data Collection

The study used the testing authoring software TestPilot (McGraw-Hill Test Pilot Enterprise (v4), http://www.clearlearning.com), that was available for any subject across the university to administer online testing. An additional Java-Script module and graphical interface were developed and built in the software to measure time between a test taker’s interactions with a computer and to give students an option of using a scale of certainty. The module collected data about input values and timestamps for each mouse click or key press performed by a test taker.

Figure 1 shows the interface of a summative test S2005. On the left side of the screen test takers were presented with an additional control panel with a timer and a set of colour-coded buttons that showed their progress through the test and the certainty/confidence they indicated on attempted questions. On the right, test takers were presented with the task, input fields and the scale of certainty with colour-coded buttons.

The choice of descriptors in the certainty scale was determined by the review of the success of implementation of other certainty scales in prior research (see section 2.4.1) and the results of the pilot study. The scale was approved by a subject coordinator and the participating teacher for use in summative and formative tests.

![Figure 1. Scale of certainty with five categories.](image)

After analysing the actual usage of the scale the study changed it to contain only three alternatives starting with summative test S2006 as demonstrated in Figure 2. The justification for the change is provided in the Results chapter, section 4.2.1.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

Figure 2. Scale of certainty with three categories.

It needs to be noted that test takers were asked to indicate their certainty in the correctness of their response but they were not penalised for not indicating or using only one grade for all responses. Though the literature review (section 2.4.3) recognises potential benefits of certainty/confidence based assessment (CBA) as a viable alternative for which implementation would likely result in a higher motivation for test takers to express certainty more precisely, the quasi-experimental nature of the study and the specific context of the real-life class-room settings imposed a number of conditions that were beyond the control of the author of the thesis. The decision to follow traditional scoring procedure and not to implement CBA was made by participating teachers and thus defined the constraints of the present study in regard to the scoring rules and certainty descriptors.

Each question was presented on a separate web page, which allowed measuring the time spent on each question – a measurement technique employed by Bergstrom, Gershon and Lunz (1994). The response time and certainty data were later combined with the TestPilot log data to allow a comprehensive record of test taker-computer interaction.

Summative (Hurdle) tests were taken under supervised conditions in computer laboratories on-campus. There was no significant difference in equipment or other testing conditions between test takers. All formative tests were taken by students online in asynchronous mode at unspecified locations. There was no time limit imposed on test takers for completing a test session.
3.6. Pilot Study

A pilot study was done in August-September 2005 on a hurdle test administered to second year education students (n=186). Results of the pilot study allowed trialling of technology and administration issues. Based on the pilot study, the scale of confidence levels was changed from five levels to three. Technical issues concerning reliability of data transmissions have been identified and resolved.

3.7. Procedures for data collection

3.7.1. Summative tests.

For two years the study collected data from summative tests (S2005 and S2006) that were taken by second year students of the Faculty of Education at the end of a year. The tests were a part of the course requirements and passing the test was compulsory for the course successful completion. Thus, the study assumed that test takers had sufficient motivation to produce a good effort.

Testing was conducted in a computer lab under supervision with strict test time limits. The test takers scheduled themselves to one of ten groups of up to twenty-five people who sat the test in a computer laboratory within one week. There was no difference in hardware or software performance within each testing session. The results were aggregated across groups.

As for some students it was the first test administered electronically, the arrangement for practice before the test were made. For 2 weeks before the test the students had unlimited access to a practice test online. Almost all students attempted the practice test at least once. It was expected that practice tests would help to reduce test anxiety and computer anxiety, and help students familiarise themselves with the testing interface. Feedback comprising correct answers and a total score was automatically generated immediately following the completion of a practice test.

Each of 27 test questions was presented on a single webpage, a sample of which can be seen in Figure 1. Accuracy data were determined by comparing a test taker’s input with a standard value suggested by a content expert. Additionally to answering a task question, the test takers were asked to indicate their certainty in correctness of the answer before they could move to a next question. Response time data were measured as the time between loading a page and clicking a next/submit button. Due to occasional technical problems some entries failed to register. Consequently, the number of registered certainty indications or response times on a question were slightly smaller.
than the total number of responses on that question. As the technical malfunctions were randomly distributed among test takers and test questions, the study did not consider them a threat to the validity.

S2005 test was used for piloting a certainty scale and overall data collection technology. For the purpose of test security, half of the questions were generated automatically from preset values to provide difference in the content between individual test versions.

S2006 test had three fixed parallel versions, with the test takers being randomly assigned to one of the versions. Each version had 27 questions that had been prepared by a mathematics content specialist and the versions were expected to be equal. S2006 was equipped with a different scale of certainty (Figure 2) on the basis of the findings from pilot study in S2005.

3.7.2. Formative tests.

For two years the study collected data from formative tests (20 tests in 2006 and 10 tests in 2007) that were taken by first year students of the Faculty of Education on a weekly basis. As a subject requirement, students were to attempt at least 80% of all tests. Each test had to be taken within the allocated time frame, usually 4-5 days, to be considered “timely” and counted towards the subject requirements.

The tests had been prepared by a mathematics content specialist and included 8-12 questions, some of those relying on prior knowledge and others testing knew knowledge that was learned at lectures and workshops. This resulted in a wide range of question facilities.

Though successful completion of a test was not compulsory, the students had a positive attitude towards the tests, considering them helpful for learning. Besides, the students realised that their teacher was able to monitor individual responses. Thus, the study assumes that test takers had sufficient motivation to produce a good effort.

Each test question was presented on a single webpage. Accuracy data were determined by comparing a test taker’s input with a standard value suggested by a content expert. Additionally to answering a task question, test takers were asked to indicate their certainty in correctness of the answer before they could move to the next question. Response time data were measured as the time between loading a page and clicking a next/submit button.

Tests were delivered online and students were allowed unlimited access, both on campus and from home or any other location, to the tests. A diversity in hardware or
software performance within testing sessions caused occasional technical problems with client-side JavaScript-based timestamp registration, leading to the loss of a few response time or certainty data points. Overall, the number of such cases was about 5% and they were randomly distributed among test takers and test questions.

There were two levels of feedback from formative tests. Firstly, the feedback comprising correct answers and a total score was automatically generated and displayed to the test taker immediately after the test was completed. This aimed at refreshing the knowledge of a specific cognitive domain or raising self-awareness about the lack of particular skills. Secondly, aggregated data of mean response time and distribution of certainty indications were made available to a teacher within a few days. This aimed at providing a teacher with guidance to plan the lesson tailored to the needs of the particular student cohort.

3.7.3. Interviews.

The focus of the study on the meaning of response time and certainty measurements required investigation of teachers’ perception of the data. The study engaged interviews as an instrument that can provide insights on how data are interpreted.

Over the course of data collection in 2005-2007 the study conducted over a hundred interviews with three teachers, one of whom was also a subject co-ordinator. The interviews were usually conducted with a teacher individually before a test to discuss general arrangements for data collection and after each test to discuss the meaning of the collected data. The interviews were informal and semi-structured, aiming at elicitation of the teacher’s perspective on the data.

The semi-structured interview method was chosen because it allows more flexibility in discussing the topics and encourages a more creative exploration than a structured interview. The post-test interviews focused on the following open-ended questions: “What do the data tell you?”; “What information do you find most useful?” and “What data format would you prefer?”. The length of the interviews varied from 15 to 45 minutes depending upon the participants willingness to speak about the topic.

The on-going nature of the interviews served two purposes: it reduced the bias of temporal individual knowledge when a teacher is strongly affected by a current event, and it facilitated gradual evolution of understanding of the meaning of response time and certainty measurements. The practical use of the data in an on-going teaching process also allowed the format of data presentation to be refined.
The final findings concerning the meaning of the data, the aspects that were perceived as the most useful and the preferable data format were formalised in two joint seminars that the researchers and one of the participating teachers conducted for the University Faculties.

Additionally, the study interviewed five students within a week after they sat the summative test S2007. The purpose of the interviews was to confirm the findings from the teachers’ interviews in regard to the suggested steps or approaches to problem solving on several test questions. The interviews with students were semi-structured, with open-ended questions, for example: “How did you solve the task?”, “Did you use mental or written calculations?”, “Tell me about your thinking when you were doing the task on conversion between millimetres and metres?”. The length of the interviews varied from 15 to 30 minutes depending upon the participants.

3.8. Ethical issues

In regard to ethical issues the study followed the standards and procedures prescribed in the Code of Conduct for Research, Regulation 17.1.R8 of the University of Melbourne (http://www.unimelb.edu.au/ExecServ/Statutes/r171r8.html) and in the National Statement on Ethical Conduct in Human Research (NHMRC) (http://www.nhmrc.gov.au/publications/ethics/2007_humans/contents.htm) which sets out the Australian national guidelines for ethical conduct in research involving human participants.

According to the stipulations of the regulations, the study lodged an application to and received a written approval from the Human Ethics office of the University of Melbourne. As the study was conducted in an established educational setting and the study’s data collection methods were mainly observational, with no elements of intervention that would be outside normal educational practices, the study was recognised as “a minimum risk project”. The study used only anonymised data from server logs that were generated during online testing in existing subjects, so the project did not impose any additional risk to student participants.

A Plain Language Statement (PLS) describing the project with the disclosure of all relevant issues was handed to each student of the subjects. A copy of the PLS was published on the subject web page. A consent form where a student could confirm or reject participation was also published on the subject web page.

All participants were informed that their participation was completely voluntary and they could withdraw at any stage or withdraw any unprocessed data at any time.
Confidentiality and anonymity of students-participants was protected, subject to legal requirements.

3.9. Statistical procedures

The study calculated and interpreted descriptive statistics, such as mean and standard deviation, to quantify typicalities, diversity and relationships among response time and accuracy variables. An alpha level of .05 was used for all statistical tests.

Effect size

The strength of the relationship between two variables was calculated to express the significance of the primary outcomes as suggested by Wilkinson et al. (1999). Cohen’s $d$ is a preferable measure throughout the study. It was chosen to define the difference between two means because it calculates the value in terms of a standard deviation for the data, which makes the significance of the results from different tests comparable.

Measuring reliability

Cronbach’s alpha or index of internal consistency was used to take account of measurement error. Alpha reliability is the average of all possible split-half reliability coefficients for the indicators of a variable, taking into account the number of indicators in split-halves. Alpha reliability coefficients range from zero to one. Values above .80 are considered to indicate a reliable measure (Cramer, 2003).

Calculating standardized response times

Standardized RT of an individual indicates his/her performance speed on the question relatively to other test takers. Considering the concerns that traditional RT analysis, when restricted to mean RT, could be misleading due to positive skewness of the data (Heathcote, Popiel, & Mewhort, 1991; Brown & Heathcote, 2003) the study also engaged distribution analysis to characterise the spread of response times among test takers and to locate and delete outlier RTs (Ratcliff, 1979).

Distributional analysis

Hohle (1965) offered a mathematical model based on a decomposition of RT into principal and additional components where the principal (residual) component has a normal distribution and additional components have an exponential distribution. The convolution of the normal and the exponential functions resulted in the ex-Gaussian distribution that fits actual RT distribution. Though the rationale underlining such decomposition remains unclear, this is still the most useful model of the RT distribution.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

(Heathcote, Popiel & Mewhort, 1991). The study intends to use components of the ex-Gaussian distribution model to obtain indications of a group RT variance.

**Generalized linear mixed model**

Examining the relationship between certainty and response time the study used the assumptions that linear responses, constant variance, and normality are questionable. Generalized linear mixed models provide a means of modelling these deviations from the usual linear mixed model.

3.10. Summary

This chapter provided an overview of the method and approaches utilized in the course of the research. It described the study design as quasi-experimental and told about the steps to maximise the validity and generalisability of the results. Further, it defined the population and sampling techniques.

The data description section contained a brief specification for 5 types of data: RT, certainty, accuracy, student problem solving strategies and teachers’ perceptions that were collected and processed by this study. Data sources were specified by time, test type, level of education and number of participants.

Then the chapter described the equipment used in data collection and provided more details on technical issues of test delivery. The change in the testing interface after the pilot study was illustrated in two figures.

Procedures for data collection were described separately for summative and formative tests due to a crucial difference in test arrangements. The use of interviews as an instrument for qualitative data collection was described in detail. Ethical issues were discussed with a reference to active standards and procedures stipulated by the university code of conduct and by the national statement on ethical conduct. Statistical procedures employed by the study were described briefly with the aim of providing references to the procedures where they were relevant to the current study.
CHAPTER 4. RESULTS AND DISCUSSION

In this chapter the study presents results and discusses findings. These are three major parts of the chapter: (a) meaning of response time statistics (section 4.1); (b) meaning of certainty statistics (section 4.2); and (c) relationship between response time and certainty (section 4.3).

The review of both constructs under investigation (response time and certainty) is preceded with validation of the statistics as a pre-requisite to any meaningful discussion. The requirements of the scale of measurement and of data preparation and processing procedures are addressed to lay a foundation for discussion of the meaning of response time (RT) and certainty for teaching and learning.

The results are presented with a reference to the method of investigation (quantitative or qualitative) and with a short description of experimental settings; they are initially reported by research question and then synthesized for trends across the study. Charts and tables are extensively used to illustrate the trends, often using a single test as an example of a phenomenon that has been observed on a larger scale where it is helpful for the explanation.

The study employed a number of techniques for data analysis, recommended for social research (Cramer, 2003; Popham & Sirotnik, 1973) but also following the advice of Fisher (1935) to choose a minimally sufficient statistic in a particular analysis that matches the research design and answers a research question. It was the intention of the study to use a simpler method if the assumptions and the strength of it were reasonable for the data and research problem, as was suggested by Wilkinson & Task Force on Statistical Inference (1999).

Special consideration was given throughout the chapter to the relationship between the two variables (response time and certainty) and the accuracy statistics that are commonly used in assessment procedures. To further synthesize the findings, the study examined the relationship between response time and certainty with the focus on the potential of RT to replace measurement of certainty, thus capitalising on the benefits of unobtrusiveness of response time measurements.

For each finding the study explored whether RT and certainty data provided additional information about student learning to what can be inferred from accuracy data.
4.1. **Meaning and Potential of Response Time Measurements for Test Design and Teaching**

In this section the study reports on the results of the investigation of the potential utility of response time measurements. While some subsections of the chapter (4.1.1 - 4.1.5) continue and expand earlier works of the author (Gvozdenko, 2005; Gvozdenko & Chambers, 2006), other subsections (4.1.6 - 4.1.7) explore the issues currently under active academic discourse in the field. The last subsections (4.1.8.1 - 4.1.8.2) report the results of the investigation that aims at making a contribution to knowledge by presenting new facts and an original approach to the issue.

At the start the study addresses the validity of the statistics and the requirements to data processing. Then the study proceeds to establish the potential of response time measurements to be used (a) for monitoring potentially parallel test items for different functioning due to mismatched cognitive load, (b) for investigating solution strategies engaged by test takers, (c) for monitoring rapid-guessing or cheating behaviours, and (d) for evaluating learners’ progress through a learning period.

### 4.1.1. Defining requirements for data processing: dealing with outliers and distribution skewness.

Approaching the issues of the usability of Response Time (RT) measurements in general, the study had to determine the requirements for data pre-processing, and in particular to examine the impact of outliers on averaging RTs on a test item. It is crucial to determine whether averaging without first deleting outliers would impair the usage of the Mean Question Response Time (MQRT) statistic for investigating some of the research questions.

A number of studies (Ratcliff, 1993; Anderson & Tweney, 1997; Heathcote, Popiel & Mewhort, 1991, Brown, 2002) were concerned about potential computational and inference problems caused by outliers. For example, estimated regression coefficients that minimize the Sum of Squares for Error (SSE) are very sensitive to outliers, and standard deviation plays a crucial role in determining a sufficient sample size that would reliably represent the population. On the other hand, deleting outliers can potentially distort the distribution of the responses and could lead to a loss of valuable information about learners.

The study examined response time data from a summative test S2006 (for the content, see Table 25) for outliers to determine the need to apply deletion, transformation or accommodation techniques. The means and standard deviations were
calculated before and after cleaning the outliers to establish whether the procedure significantly altered the statistics. This is illustrated in Table 2. For each of 27 questions of the test the MQRTs and standard deviations for response times of all test takers \( (n = 203) \) were computed, then after visual examination of RT distribution the values for cutting off too long or too short responses were determined and the outliers thus defined were removed. For short responses a gap in the distribution within the first 10 sec was taken as an indication of the cut-off time. Additionally, a factor of accuracy was taken into consideration. For longer responses a rule of three standard deviations was used to guide the cut-off.

As it can be seen from Table 2, the number of outliers varied between test questions in the range from 5 (question 160) up to 15 (question 290). Reduction in MQRT due to the deletion of outliers also varied between test questions in the range 2\% up to 16\%. Interestingly, the largest reduction was observed on the questions about translation and rotation of geometric shapes (questions 280 and 290), and the smallest reduction happened on the questions that involve calculations rather than visualisation and can be solved by simple application of a formula (questions 120 - 180). Due to the general trend of response time distribution to be positively skewed, the elimination of too long responses had a greater effect on MQRT than the elimination of too short responses, which resulted in reduction of MQRT on all questions.

The reduction in standard deviation, as it could be expected due to the commonly recognised impact of outliers on standard deviation, was more significant, in the range 21\% (question 170, about finding the volume of a cylinder) up to 70\% (question 280, about translation of a geometric figure).

Overall, deleting outliers resulted in average reduction of MQRT by 7.2 percent while standard deviation was affected more, with an average reduction of 34.1 percent. As the reduction in MQRT does not seem to be substantial, it can be suggested that for the purpose of this study there is no need for a special processing of response time data when mean response time on a question is required. In addition, when MQRT is used to compare parallel test versions, it can be assumed that outliers are randomly distributed across parallel versions of a test due to the randomised assignment of test versions and thus the impact of outliers on investigation of between-version variation would be negligible.
Table 2
Mean question response time (MQRT) and standard deviation statistics before and after removing outliers in summative test S2006.

| Question ID | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 | 120 | 130 | 140 | 150 | 160 | 170 | 180 | 190 | 200 | 210 | 220 | 230 | 240 | 250 | 260 | 270 | 280 | 290 | 300 |
|-------------|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Before deleting outliers | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Number of responses | 190 | 192 | 198 | 196 | 199 | 200 | 199 | 200 | 199 | 198 | 197 | 199 | 194 | 190 | 194 | 191 | 190 | 191 | 192 | 198 | 197 | 191 | 194 | 189 | 162 | |
| MQRT | 39 | 41 | 41 | 74 | 78 | 65 | 48 | 59 | 74 | 203 | 154 | 154 | 69 | 113 | 118 | 38 | 38 | 61 | 43 | 28 | 25 | 23 | 36 | 30 | 22 | 62 | 111 | |
| SD | 29 | 34 | 29 | 35 | 39 | 36 | 34 | 35 | 39 | 111 | 91 | 99 | 35 | 64 | 65 | 30 | 28 | 42 | 41 | 24 | 28 | 24 | 35 | 22 | 30 | 52 | 74 | |
| After deleting outliers | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Number of responses | 181 | 180 | 183 | 185 | 190 | 192 | 191 | 194 | 193 | 192 | 188 | 189 | 192 | 193 | 188 | 183 | 185 | 184 | 184 | 186 | 191 | 188 | 186 | 186 | 174 | 156 | |
| MQRT | 34 | 35 | 36 | 70 | 74 | 62 | 45 | 56 | 72 | 198 | 148 | 151 | 67 | 111 | 115 | 35 | 35 | 57 | 39 | 26 | 23 | 21 | 31 | 29 | 19 | 52 | 105 | |
| SD | 18 | 17 | 19 | 26 | 28 | 28 | 25 | 26 | 29 | 83 | 69 | 72 | 27 | 50 | 49 | 19 | 17 | 21 | 26 | 16 | 17 | 14 | 15 | 16 | 9 | 27 | 54 | |

Resulting decrease in mean RT and standard deviation

| Reduction in MQRT (in %) | 13% | 14% | 14% | 5% | 5% | 3% | 7% | 4% | 3% | 2% | 4% | 2% | 2% | 2% | 3% | 9% | 9% | 6% | 9% | 7% | 10% | 10% | 14% | 3% | 16% | 16% | 5% |
| Reduction in SD (in %) | 36% | 50% | 37% | 25% | 27% | 22% | 27% | 26% | 26% | 25% | 24% | 27% | 23% | 21% | 24% | 36% | 38% | 50% | 38% | 31% | 40% | 40% | 56% | 28% | 70% | 49% | 26% |
A number of researchers had suggested using lognormal transformation of response time (Thissen, 1983; Schnipke and Scrams, 1997; van der Linden, Scrams, and Schnipke, 1999) before applying further statistical procedures because RT distributions are usually positively skewed. Each of those studies demonstrated a good fit of response times to a lognormal distribution. However, normalisation of distribution through lognormal transformation of data in a pilot study was found to affect the shape of distribution by creating left skewness on some questions, thus making interpretation more difficult. As analysis of the distribution was in the focus of most research questions of this study, the study chose to use original values (not transformed) to preserve RT distributions that were used as evidence for making judgements about the difference in group response times or test item functioning. When inferential statistical procedures are involved in other research questions in this study, the statistical techniques are selected with attention to robustness to distribution skewness.

4.1.2. Investigating MQRTs on parallel tests.

For practical purposes of test administration test S2006 was subdivided into three parallel versions S2006v1, S2006v2 and S2006v3. The questions (see Table 25 for content) were attempted to be designed equal in respect to testing the same cognitive constructs and carrying a similar cognitive load. MQRTs of parallel questions (Table 26) were then examined to confirm the requirements of equality. Two examples of how parallel versions of a question differed can be found in Table 5 and Figure 5.

Figure 3 illustrates the comparison of the three versions of the test by MQRT on each test question. It can be seen (Figure 3, Table 26.) that on most test questions MQRTs of all versions of test questions are close. However, on a few questions the versions differed in MQRT substantially. The top three questions with the greatest difference were identified:

- MQRT of question 240_S2006v3 (n = 58, M = 18, SD = 16) was 11 sec (37%) smaller than MQRT of question 240_S2006v2 (n = 64, M = 30, SD = 17), t(120) = -3.60, p < .001, d = -.73;
- MQRT of question 140_S2006v3 (n = 58, M = 119, SD = 59) was 51 sec (30%) smaller than MQRT of question 140_S2006v2 (n = 66, M = 171, SD = 69), t(122) = -4.49, p < .001, d = -.81;
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

- MQRT of question 220_S2006v1 (n = 67, M = 34, SD = 24) was 14 sec (29%) smaller than MQRT of question 220_S2006v3 (n = 51, M = 48, SD = 29), \( t(116) = -2.82, p = .005, d = -.54 \);

![Mean response times (MQRT) for all questions of three parallel tests S2006v1, S2006v2, S2006v3.](image)

**Figure 3.** Mean response times (MQRT) for all questions of three parallel tests S2006v1, S2006v2, S2006v3.

This section demonstrated that from the visual analysis of Figure 3 it was possible to detect the difference in item functioning between versions by comparing MQRTs. A few statistical procedures for estimating a magnitude of the discrepancies will be explored later in the chapter.

### 4.1.3. Evaluating the difference between parallel items.

The study explored how analysis of individual response times could elaborate on the difference in item functioning revealed by MQRT. Effect size (for example, Cohen’s \( d \)) could be a good measure of the difference, however it would require calculations for each pair of versions. Besides, it is not available in commonly used software, such as SPSS (PASW). This study employed the use of analysis of variance (ANOVA) as it is one of the common procedures for evaluating the difference between samples from different populations.

Figure 4 presents results of ANOVA of the same summative test S2006 as Figure 3. The vertical axis demonstrates statistical significance of the difference between the means for the three versions of each test item. The values on the vertical axis are placed...
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

on inverted logarithmical scale to highlight the items with the most significant difference between the versions of the questions.

As can be seen in Figure 4 analysis of variance returned p-values > 0.05 on most items. However, all three questions (240, 140 and 220), which were flagged by MQRT also stood out in ANOVA (Figure 4) as it was expected. Two of them, questions 140 and 240, demonstrated also the most statistically significant difference between the test versions \((p < 0.001)\) in the analysis of variance of individual RTs. Those questions will be a subject for further investigation into probable causes of the differences. The questions will serve as examples of the utility of response time measurements to inform teaching and cognitive research of underlying cognitive phenomena.

![Figure 4. P-values of ANOVA of three parallel tests S2006v1, S2006v2, S2006v3.](image)

As our study aimed to demonstrate how relatively simple statistical procedures can provide very rich information about test item performance, a distribution analysis was examined for its utility to demonstrate visually the difference in item functioning and to evaluate the magnitude of the difference. The assumption behind it is that a difference in cognitive load of a question may be reflected in a distribution of individual RTs in a number of ways: (a) the distribution can be shifted to reflect a longer/shorter time required to solve the problem, or (b) the difference in cognitive load may affect high and low achievers in a different way, or (c) some test takers would choose a different strategy to solve the problem, which can be shorter or longer than average and will be reflected in a changed tail of the distribution. Examples of applying distribution
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

analysis to examine individual RTs on parallel test questions are presented in the following section.

4.1.4. Detecting difference in question complexity.

The study continued investigation of the items that were flagged by the difference in MQRT and confirmed by ANOVA of response times. Such difference can imply that the questions are not truly parallel, being miscalibrated or quasi-equal (Gvozdenko, 2005). One of the questions, selected for further investigation in the previous part of the chapter (Figure 3), was question 240_S2006 which had the largest (relative) difference between versions 3 (test S2006v3) and 2 (test S2006v2). MQRT of version 3 of question 240 was 11 sec (37%) smaller than MQRT of version 2.

The questions are given in Figure 5. From visual examination of the images in Figure 5 it is clear that the image of version 2 is the most complex for mental representation of rotational symmetry. It has 2 different rotated elements and the highest order of rotational symmetry (5) so it would be reasonable to expect a higher mental load and longer processing time, which is supported by both accuracy and response time data. Interestingly, though the percentage of correct answers on this version was not substantially smaller (by 7%) than on version 3, time demands differed significantly as it took test takers on average 60% longer to solve version 2 than version 3.

When the case was referred to content experts who examined the three versions of the question they agreed that the most likely cause of the difference is a higher complexity of version 2, however the magnitude of the difference in MQRT was a surprise to them.

Examination of the histograms of response time distributions in Figure 6, Figure 7 and Figure 8 revealed additional details about the difference between the questions. As could be expected from the difference in MQRT, the distribution of RTs on version 2 (Figure 7) clearly differed from other versions, with longer response times. By the time when more than two thirds of test takers completed version 3 (Figure 8) and a half of test takers had completed version 1 (Figure 6), less than a quarter of test takers had finished version 2.
Task: **For the given shape, give the order of rotational symmetry.**

<table>
<thead>
<tr>
<th>Version 1</th>
<th>Version 2</th>
<th>Version 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>240_S2006v1</td>
<td>240_S2006v2</td>
<td>240_S2006v3</td>
</tr>
<tr>
<td>![Shape 1]</td>
<td>![Shape 2]</td>
<td>![Shape 3]</td>
</tr>
</tbody>
</table>

**Answer:**

- **Version 1:**
  - n = 64
  - MQRT = 19.6 sec
  - SD = 15.2
  - Facility = 92%

- **Version 2:**
  - n = 64
  - MQRT = 29.4 sec
  - SD = 17.4
  - Facility = 86%

- **Version 3:**
  - n = 58
  - MQRT = 18.4 sec
  - SD = 16.1
  - Facility = 93%

*Figure 5. Three versions of question 240_S2006.*

However, the observed difference in the shape of the distributions between version 1 (Figure 6) and version 3 (Figure 8) was a surprise. It can be seen that distribution shape of version 3 is close to “a normal distribution plus a right-hand tail”, with about 80% of test takers answering the question within 24 seconds. The distribution of RTs on version 1, in contrast is characterised with a positive skewness and no embedded normal distribution. About 80% of students completed the question within 27 seconds which is over 10% longer than for version 3. Considering that there was little difference in the number of correct responses, it can be suggested that for some test takers the geometric shape that was used in version 1 requires more time to be processed, as this shape is more complex and there are more visual elements.

Thus, the analysis of MQRT, followed up with visual examination of histograms of RTs distributions, offered additional clues to detect the difference between versions of the question. In this case the use of RT measurements magnified the difference, first detected by the difference in question facility on version 2, and provided insights into the difference between version 1 and version 3, which would not be available from the statistic of question facility only.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

Figure 6. Distribution of response times on question 240_S2006v1.

Figure 7. Distribution of response times on question 240_S2006v2.

Figure 8. Distribution of response times on question 240_S2006v3.
4.1.5. Investigating difference in required skills.

The following section presents another example of using RTs to detect a difference between the questions that were intended to be parallel. In this case the source of the difference was identified as a mismatch in required skills.

A summative test S2005 included an item that was testing the skill of conversion from square metres into hectares (Table 3). Each of the test takers was offered one of two versions of the item, randomly drawn from the test pool. Those questions were expected to be parallel which would not be challenged if only the usual measure of question facility was used to monitor question functioning.

However, the values of MQRT indicated that time demands for those two questions differed by nearly 30%, which flagged the item for further investigation. One-way ANOVA found that MQRT of question 530_S2005 significantly differed from MQRT of question 210_S2005 ($F_{1,154}=6.38$, $p<.05$).

Table 3. Questions requiring conversion from square metres into hectares (S2005).

<table>
<thead>
<tr>
<th>Question ID</th>
<th>210_S2005</th>
<th>530_S2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>27500 m² = ? (ha)</td>
<td>690 m² = ? (ha)</td>
</tr>
<tr>
<td>Answer</td>
<td>2.75</td>
<td>0.069</td>
</tr>
<tr>
<td>Number of responses</td>
<td>65</td>
<td>91</td>
</tr>
<tr>
<td>Facility (% of correct answers)</td>
<td>69%</td>
<td>72%</td>
</tr>
<tr>
<td>MQRT (in sec)</td>
<td>48</td>
<td>62</td>
</tr>
<tr>
<td>SD (in sec)</td>
<td>31</td>
<td>37</td>
</tr>
</tbody>
</table>

The observed difference in MQRT prompted further investigation by means of visual examination of the distribution histograms for each task (Figure 9, Figure 10). Cumulative frequency graphs in figures Figure 9 and Figure 10 demonstrates that while half of the test takers completed question 210_S2005 within 40 seconds and 90% of test takers did it within 87 seconds, for question 530_S2005 it took 60 seconds for the first half of the test takers and up to 105 seconds for 90% of the test takers to complete the question. From the charts it is clear that those two questions are substantially different in terms of solution time demands. This triggered further investigation into probable causes.
Exploring the hypothesis that the content of question 530_2005 contains an additional cognitive task, the study examined other questions that required similar calculations. In the previous work (Gvozdenko, 2005, p.51) similar (but not identical) questions (210 and 530, Table 4) were used in the 2004 experiment on a similar student population. The difference between means of those questions in 2004 was not statistically significant. Comparison between 2004 and 2005 questions (Table 4) led to the proposal that the identified difference may be attributed to the different number of digits in the value being transformed from square metres to hectares.
Table 4. *Questions requiring conversion from square metres to hectares.*

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Facility (% of correct)</th>
<th>MQRT (in sec)</th>
<th>Task</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>210 (2004)*</td>
<td>77</td>
<td>44</td>
<td>12560 m² = ? (ha)</td>
<td>1.256</td>
</tr>
<tr>
<td>530 (2004)*</td>
<td>78</td>
<td>43</td>
<td>9570 m² = ? (ha)</td>
<td>0.957</td>
</tr>
<tr>
<td>210_S2005</td>
<td>69</td>
<td>48</td>
<td>27500 m² = ? (ha)</td>
<td>2.75</td>
</tr>
<tr>
<td>530_2005</td>
<td>72</td>
<td>62</td>
<td>690 m² = ? (ha)</td>
<td>0.069</td>
</tr>
</tbody>
</table>

Note: *Question IDs are referenced according to Gvozdenko (2005).*

It was hypothesized that question 530_2005 is different from other questions because dividing by 10000, as required for conversion from m² to ha, involves moving a decimal point to the left by four digits, and some test takers may encounter an additional problem when there are not enough digits just to move the decimal point. They have to recall the procedure of inserting an additional zero after the decimal point in question 530 (2005), which increases the time spent on the question.

Further analysis of test takers' errors confirmed the association of wrong answers involving a misplaced decimal point with longer RTs. Consultation with Dr Vicki Steinle whose PhD thesis (Steinle, 2004) was in the area of decimal content confirmed the explanation. This example demonstrates how RT measurements can initiate an enquiry into the test content and generate a deeper understanding of the cognitive demands of a test question.

4.1.6. **Investigating solution strategies used by test takers.**

This section reports on the investigation of the research question: “What is the potential and what are the limitations of response time data in predicting the range of solution strategies on correct answers?” To demonstrate the approach, the section continues examining the test questions that were flagged by MQRT due to different time demands on the versions that were supposed to be parallel. As it has been described earlier in the chapter, analysis of MQRT of summative test S2006 (Figure 3) detected the most significant differences between parallel versions of question 240_S2006 and question 140_S2006.

In this section the study investigates the origin of discrepancies in response times between the three versions of question 140_S2006: 140_S2006v1, 140_S2006v2 and 140_S2006v3 (Table 5). After detecting the difference in MQRT, examination of distributions of RT on the three versions was performed. The study hypothesised that
the difference between the questions was caused by the difference in cognitive demand and the RT difference reflects a proportion of students who used a written method of calculation as an alternative to mental calculations. To test the hypothesis the study administered the same questions as part of a formative test WOT20_F2006 to a similar cohort of students and asked them to indicate which method of calculation they had used for each step of the solution.

As it can be seen in Table 5, MQRT of version 2 (140_S2006v2) is considerably longer (43%) than of version 3 though the difference in facility is only 5 points. Standard deviation of version 2 is also larger by 19% and the 20th percentile of RT is larger by 62% than the corresponding statistic for version 3. This means that a considerable number of test takers were experiencing additional cognitive load.

Table 5.
Descriptive statistics for question 140_S2006.

<table>
<thead>
<tr>
<th>Test version</th>
<th>Task</th>
<th>Facility</th>
<th>MQRT (in sec)</th>
<th>SD</th>
<th>Percentile 20th</th>
<th>Percentile 80th</th>
</tr>
</thead>
<tbody>
<tr>
<td>V. 1 n=72</td>
<td>Izabella is washing windows. It takes her 5 min to wash and polish 1 square metre area of glass. The room she is cleaning has 8 rectangular windows, 50 cm wide and 1.8 m high. How long will it take her (please, round your answer in minutes)?</td>
<td>70%</td>
<td>151</td>
<td>67</td>
<td>93</td>
<td>208</td>
</tr>
<tr>
<td>V. 2 n=66</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 8 rectangular windows, 75 cm wide and 6 m high. How long will it take her (please, round your answer in minutes)?</td>
<td>82%</td>
<td>171</td>
<td>69</td>
<td>110</td>
<td>191</td>
</tr>
<tr>
<td>V. 3 n=58</td>
<td>Izabella is washing windows. It takes her 4 min to wash and polish 1 square metre area of glass. The room she is cleaning has 5 rectangular windows, 50 cm wide and 2.4 m high. How long will it take her (please, round your answer in minutes)?</td>
<td>77%</td>
<td>119</td>
<td>58</td>
<td>68</td>
<td>160</td>
</tr>
</tbody>
</table>
Figure 11 presents a box plot for each version of the question to demonstrate the difference between the medians, the upper and lower quartiles and the interquartile ranges of the questions. Further investigation of RT distributions was considered warranted by the observed difference in descriptive statistics between the versions.

Figure 11. Box plots for three versions of question 140_S2006.

Histograms presented in Figure 12, Figure 13 and Figure 14 demonstrate in more detail that the three versions of question 140_S2006 produced different distributions of test takers’ RTs. The right-hand side vertical axis represents accumulated percentage of students and helps to illustrate the time range within which most students completed the task. The left-hand side vertical axis represents the number of students for each time bar.

The histogram in Figure 12 shows that RT of version 1 ranges between 30 sec and 290 sec with a gap between 240 sec and 270 sec. The distribution shape (Skewness = .25; Kurtosis = -.41) is slightly positively skewed, which is typical for RT distributions (Heathcote, 2003). Over 60% of RTs are in less than 170 sec. The data looks reasonably symmetric, with the modes at 110, 130, 160 and 190 quite evenly spread within the mid-range.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

Version 2 (Figure 13), however, is characterised by a shift of the distribution to the right compared to Version 2 (Skewness = .07; Kurtosis = -.32). Only 47% of test takers did the question within 170 sec, which suggests that for some test takers the content of the question required more processing time than on other versions.

Conversely, Version 3 (Figure 14) (Skewness = .96; Kurtosis = .69) demonstrates a shift of responses to the left resulting in positive skewness, which means that many test takers answered the question quickly. Over 80% of test takers completed the question within 170 sec.

Figure 12. Distribution of test takers’ response times on question 140_S2006v1.

Figure 13. Distribution of test takers’ response times on question 140_S2006v2.
After establishing and quantifying the differences in RT distributions of the versions of the test question, the study proceeded to investigate the causes of the variation. The study hypothesised that the variation could be attributed to the difference in cognitive demand arising from the version tasks requiring calculations of different degrees of difficulty.

A content expert suggested that most test takers would follow four steps of the solution where each of the steps could be done as mental calculation or written calculation, and that the difference in time demands between the two methods would generate difference in response time. The hypothesis was then re-stated that a longer MQRT of a version would be associated with the prevalence of a written strategy for calculations over a mental one.

To test this hypothesis another cohort of the same population was given the same question versions as part of formative testing in question 140_WOT20_F2006. In the following question of the test they were given a set of possible steps that had been provided by a content expert, and asked to indicate the type of calculation they had employed at each step (Table 6) to solve the problem. Their indications are illustrated in Figure 15.

---

Figure 14. Distribution of test takers’ response times on question 140_S2006v3.
Table 6.  
*Suggested steps of solution for questions 140_S2006 and 140_WOT20_F2006.*

<table>
<thead>
<tr>
<th>Step No.</th>
<th>Description of a step</th>
<th>Suggested expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>V. 1</td>
</tr>
<tr>
<td>1</td>
<td>Conversion of cm to m</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 cm = 0.5 m</td>
<td></td>
</tr>
<tr>
<td></td>
<td>or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 cm = ½ m</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Calculating area of 1 window</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5 x 1.8 = 0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>½ x 1.8 = 0.9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Calculating area of all windows</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9 x 8 = 7.2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Calculating time to clean all windows</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.2 x 5 = 36</td>
<td></td>
</tr>
</tbody>
</table>

Figure 15 illustrates the proportions of mental calculations employed by the students at each step by tabulating the responses described in Table 6. It can be seen that step 1 is characterised by a prevalence of mental calculations over written in all three versions, which could be expected as the conversion from centimetres to metres is very straightforward. On the contrary, step 2 clearly indicates the preference of a written method of calculation. The percentage of the test takers who chose written strategy is between 60 and 70 per cent for all three versions at that step. At the third step version 1 produced more answers by mental calculation than other versions. However, at the fourth step version 3 had two-thirds of test takers using mental calculation strategy while version 2 had only one-third.

It is clear from the data that version 2 is different from other versions as the test takers consistently demonstrated a preference for the written method from step 2 onwards. The study suggests that this explains the longest MQRT of the version.

Interestingly, the version also produced the highest facility (82%), perhaps because some of the test takers who chose mental method on version 3 (facility = 77%) overestimated their proficiency in mental calculations.
This section demonstrated how the analysis of response time data enabled the detection of the difference and offered substantial grounds for generating a hypothesis about the cognitive nature of the phenomenon. In this case the difference in accuracy (5%) was only marginal and on its own would hardly trigger further investigation into the cognitive process. Response time statistics provided an instrument for evaluating the difference between the versions which was useful for test design and teaching.

From the perspective of informing teaching of the current cognitive state of a group it can be suggested that comparing the distribution of response times with pre-determined reference values may provide insights into the quality of calculation skills.

However, the detection facility of RT measurements was found to be effective only when the solution strategies differed significantly in time requirements, which should be considered as one of the limitations of the procedure.

4.1.7. Examining rapid-guessing, cheating and distraction in formative testing.

This section focuses on detection of anomalies in test takers’ behaviour that are likely to be associated with such phenomena as rapid-guessing, cheating or distraction. Those phenomena are explored in the context of formative testing due to an urgent need
for the tools that would inform an educator about the underlying student behaviour and a limited number of studies in this area.

The study examined unusually fast or slow responses on formative tests to establish a prevalence and the magnitude of the events and their meaning for teaching in the context of tertiary education settings. The study assumed that the presence of too fast responses could reflect a rapid-guessing behaviour, associated with low effort or pre-knowledge of an answer due to cheating. Extremely slow responses were assumed to reflect a distraction experienced by some students during formative testing.

The current study collected and analysed RTs data from 20 formative tests (WOT1_F2006 – WOT20_F2006) to investigate the presence of extremely fast and extremely slow responses. Following the method proposed by Wise and King (2005), after visually examining histograms of RTs distributions for each of the test questions the study determined an arbitrary value of 9 seconds that would be a minimum required to read a task and come up with a solution. All responses that were faster than 9 seconds were flagged as too fast answers for further investigation. A similar approach was also taken by Kong, Wise and Bhola (2007). An example from a test administered in the fourth week of a semester is given in Table 7 (WOT4_F2006). This example demonstrates that few students produced too fast responses (only 2% of total responses).

<table>
<thead>
<tr>
<th>Question ID</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQRT</td>
<td>107</td>
<td>87</td>
<td>33</td>
<td>68</td>
<td>25</td>
<td>32</td>
<td>30</td>
<td>40</td>
<td>23</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>SD</td>
<td>71</td>
<td>68</td>
<td>23</td>
<td>54</td>
<td>24</td>
<td>23</td>
<td>24</td>
<td>29</td>
<td>17</td>
<td>21</td>
<td>35</td>
</tr>
</tbody>
</table>

Number of responses

<table>
<thead>
<tr>
<th>Attempted</th>
<th>103</th>
<th>102</th>
<th>103</th>
<th>104</th>
<th>104</th>
<th>100</th>
<th>100</th>
<th>104</th>
<th>100</th>
<th>104</th>
<th>103</th>
<th>104</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too fast*</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Too slow*</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

* Too fast RT < 9 sec; too slow RT > MQRT + 2SD

By the end of the semester the number of too fast responses slightly increased (to 3%), as can be seen in Table 8. A breakdown by accuracy reveals that most of the fast responses are also correct which suggests the possibility of cheating.
The information about the number of too fast responses and the proportion of correct answers in the responses was presented to a teacher and perceived as useful for monitoring the process of formative test administration.

Another observation concerns the responses that were considered too slow. Too slow responses were defined as the ones for which response times exceeded two and a half standard deviations above the mean on the question. Surprisingly, the study found few such responses (< 0.5%). This could be partially explained by teachers’ instructions to students that advised them to take a best guess in the situation of missing knowledge and move on. Another reason may be that the students were not generally predisposed to spend too much time on a formative test.

Overall, the data suggest that behaviour of test takers on a formative test is not characterised by any significant increase in a number of rapid-guessing cases or a number of too long responses when compared to summative testing. This finding will contribute to the discussion of the concerns about test validity and the factor of student motivation (Haladyna and Downing, 2004) and rapid-guessing behaviour (Setzer & Allspach, 2007; Kong, Wise & Bhola, 2007) in low stakes testing. As all referenced studies examined response times on a low-stake test that was a copy of a summative test it was not clear how the findings could be extended to response time measurements on a formative test in tertiary settings.

Analysis of the data of the current study demonstrated that information about the magnitude of rapid-guessing behaviour is important to a teacher as an indicator of low effort, and RT measurements can deliver that information. The study also demonstrated that providing RT statistics for too fast correct responses has a potential to alert a teacher to cheating as a probable cause of quick, however correct, responses.
4.1.8. Monitoring learning progress of a group across a learning period.

One of the aims of the present study was to investigate the potential of the use of response time statistics for the purpose of informing teaching about learners’ progress in the context of formative assessment. The study examined the dynamics of response time on four tests which were taken at the beginning and at the end of a 12 week period (semester) to establish: (a) whether mean response time changed over the semester, (b) how response time of the students who answered correctly both times was impacted by teaching, and (c) how different was response time of the students who answered correctly only at the end of the period from response time of the students who answered correctly both times.

The four tests under investigation were taken in March-April 2006 and re-taken in June 2006. Sufficient time between the tests allowed an assumption that the impact of teaching and learning that happened within that period could reasonably be expected to exceed the impact of pre-knowledge of the answers due to the feedback that had been received after the first test attempts. For each student the data included their responses to test questions and time stamps for the key presses. The descriptive statistics for accuracy and response time, presented further in this section, were calculated for those students who took both tests, to examine the progress of a student across the semester.

The question formats were of two major types: multiple choice and free response. An analysis of variance showed that the effect of the format type on question facility was significant, $F_{1,40} = 7.62, p = .009$. Average question facility was significantly higher in the condition of multiple choice format (27 questions) ($M = 76\%$, $SD = 17\%$) than in the condition of open response format (15 questions) ($M = 59\%$, $SD = 24\%$), with a large effect size (Cohen’s $d = 0.88$). However, the effect of the format type on the change in question facility was not significant ($F_{1,40} = 3.15$, $p = .083$), and for the purpose of investigating the change in the facilities the questions of different formats are grouped together.

Figure 16 presents a scatter plot describing how question facility and mean question response time (MQRT) of all forty-two questions in tests WOT1_F2006 – WOT4_F2006 changed between initial test taking at the beginning of the semester and the repeat at the end of the semester. It can be seen that the percentage of correct answers increased for nearly all questions between the two rounds of testing, showing an improvement across the semester. It should be noted that some questions like 20_WOT4_F2006 or 60_WOT3_F2006 which marks are displayed utmost right, have a large increase in facility because the initial facility was low.
Another set of findings concerns response time. Most of the points in Figure 16 are concentrated in the negative area of the vertical scale, representing MQRT, which shows that for most test takers it took less time to do the same test by the end of the semester. The study assumes that this observation can be treated as evidence of some impact of teaching and learning on students between the test sessions. Interestingly, most of the seven questions that did not fit the trend of MQRT reduction, still demonstrated increased question facilities.

Exploration of the relationship between the change in MQRT and the change in facility suggested that mean question response time, as it can be seen in Figure 16, shows a weak tendency to reduce while the facility increases, however the correlation between those two variables was not statistically significant (Pearson $r = -.14; p = .35$), which may suggest that the variables are dependent on different cognitive phenomena.

The study next uses WOT2_F2006 as an example of the approach taken, as this test happens to contain the three questions that displayed a negative change in question facility and one item (50_WOT2_F2006) that was characterised with both increased response time and decreased question facility. Table 9 presents the content of those questions and

Table 10 presents additional statistics about all questions of WOT2_F2006 including question type, initial and final question facility, and initial and final MQRT.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

That information is shown for the reference only as the study did not intend to
investigate the reasons why some of the students who had responded correctly at the
beginning were not able to produce a correct answer by the end of the semester.

Table 9.  
*Questions with negative change in facility across a semester.*

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Question task</th>
<th>Choices</th>
<th>Correct answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>40_WOT2_F2006</td>
<td>The Roman numeral CIX represents which of these numbers?</td>
<td>109 119 104 1009</td>
<td>109</td>
</tr>
<tr>
<td>50_WOT2_F2006</td>
<td>I bundle 10 pop-sticks to make a new unit, and I bundle 10 of these bundles to make a “big” bundle. What MAB piece would this “big” bundle be equivalent to?</td>
<td>Mini Long Flat Cube</td>
<td>Flat</td>
</tr>
<tr>
<td>110_WOT2_F2006</td>
<td>If I multiply a prime number and a composite number then the result is a composite number</td>
<td>sometimes always never</td>
<td>always</td>
</tr>
</tbody>
</table>

Table 10.  
*Question facilities in WOT2_F2006 at the beginning and end of semester (n>100).*

<table>
<thead>
<tr>
<th>Question ID</th>
<th>20 30 40 50 60 70 80 90 100 110 120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question type</td>
<td>MC MC MC MC MC OR TF MC MC MC OR</td>
</tr>
<tr>
<td>Facility (in %)</td>
<td>Average</td>
</tr>
<tr>
<td>beginning</td>
<td>36 77 90 86 73 28 47 69 35 59 83 62</td>
</tr>
<tr>
<td>end</td>
<td>57 87 77 82 82 34 66 76 53 57 89 69</td>
</tr>
<tr>
<td>Increase</td>
<td>21 10 -13 -4 9 6 19 7 18 -2 6 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MQRT (in sec)</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>beginning</td>
<td>36 27 35 29 62 39 17 32 24 27 37 33</td>
</tr>
<tr>
<td>end</td>
<td>23 25 31 31 48 36 10 25 17 15 27 26</td>
</tr>
<tr>
<td>Increase (in %)</td>
<td>-36 -8 -9 6 -23 -6 -37 -23 -31 -42 -27 -21</td>
</tr>
</tbody>
</table>

Note: MC=multiple choice; OR=open response; TF=True/False

Overall, after examining general trends in question facility and MQRT the study established that question facility tends to increase and mean response time tends to decrease over a learning period. The correlation between the change in question facility
and the change in MQRT was weak which suggests that the variables are impacted by
different aspects of the teaching and learning process and thus the information delivered
by the variables can be complementary. The magnitude of the change differed in
different learning areas and was not associated with a question format.


After exploring the data for the whole cohort, the study proceeded to
investigating subgroup differences in response time when a factor of initial level of
expertise is taken into account. The purpose of this exploration was to establish whether
response time could be used to measure an impact of teaching and learning on the high-
achievers who already had some knowledge at the beginning of the semester. This
impact could not be measured with accuracy statistics because of a ceiling effect. The
increase in question facility would not reflect any progress of those high achievers who
produced correct responses both at the beginning and at the end of the learning period.

In the four formative tests (WOT1_F2006, WOT2_F2006, WOT3_F2006, WOT4_F2006) that were previously examined for general trends in the change in question facility and mean question response time in section 4.1.8, for each of 42 test questions the study identified the students who took the question at the beginning of the semester and then re-took it at the end of the semester and who were correct both times. Those students will be further referred to as “student-experts” in a particular question.

Table 11 presents the data from test WOT2_F2006 to demonstrate a trend observed across all four tests that were used to investigate the dynamics of RT across the semester. Average response time for student-experts was calculated for each question of the test. A paired-samples t-test was conducted to compare response times of student-experts for the difference in means on each question of this test in beginning and end conditions. There was a statistically significant difference in mean response time for beginning (\( M = 31, SD = 11 \)) and end (\( M = 23, SD = 11 \)) conditions; \( t(10) = 5.63, p < .001 \). The results suggest that on most questions it took the students less time to answer the same questions at the end of the semester than at the beginning.

From Table 11 it is also evident that the amount of reduction in response time was different on different questions. The greatest reduction in relative terms happened on questions 20, 80, 100 and 110, whose content can be found in Table 12. However, a paired samples t-test for means of RTs for each individual question (Table 11) found that only on five test questions (20, 60, 90, 100, 110) the difference in means was statistically significant at level .05.
Table 11.
*Change in MQRT for students-experts in test WOT2_F2006.*

<table>
<thead>
<tr>
<th>Question ID</th>
<th>MQRT (in sec)</th>
<th>SD (in sec)</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At the beginning of the semester</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Students (n)</td>
<td>26</td>
<td>66</td>
<td>67</td>
<td>65</td>
<td>49</td>
</tr>
<tr>
<td>MQRT (in sec)</td>
<td>29</td>
<td>23</td>
<td>33</td>
<td>30</td>
<td>55</td>
</tr>
<tr>
<td>SD (in sec)</td>
<td>29</td>
<td>18</td>
<td>25</td>
<td>18</td>
<td>36</td>
</tr>
<tr>
<td>At the end of the semester</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Students (n)</td>
<td>26</td>
<td>66</td>
<td>67</td>
<td>65</td>
<td>49</td>
</tr>
<tr>
<td>MQRT (in sec)</td>
<td>15</td>
<td>23</td>
<td>29</td>
<td>28</td>
<td>45</td>
</tr>
<tr>
<td>SD (in sec)</td>
<td>7</td>
<td>16</td>
<td>20</td>
<td>20</td>
<td>27</td>
</tr>
<tr>
<td>Paired two sample test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>2.5</td>
<td>0.1</td>
<td>1.1</td>
<td>0.7</td>
<td>2.0</td>
</tr>
<tr>
<td>df</td>
<td>25</td>
<td>65</td>
<td>65</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>0.02</td>
<td>0.89</td>
<td>0.27</td>
<td>0.47</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 12.
*Questions exhibiting greatest reduction in MQRT of expert subgroup in WOT2_F2006.*

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Question task</th>
<th>Choices</th>
<th>Correct answer</th>
<th>Increase in MQRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>20_WOT2_F2006</td>
<td>Letters are building blocks for words. What are the building blocks for numbers?</td>
<td>digits</td>
<td>-49%</td>
<td></td>
</tr>
<tr>
<td>80_WOT2_F2006</td>
<td>Is 1 a prime number?</td>
<td>True False</td>
<td>False</td>
<td>-39%</td>
</tr>
<tr>
<td>90_WOT2_F2006</td>
<td>If I multiply two composite numbers then the result is a composite number</td>
<td>sometimes always never</td>
<td>always</td>
<td>-30%</td>
</tr>
<tr>
<td>100_WOT2_F2006</td>
<td>If I multiply two prime numbers then the result is a prime number.</td>
<td>sometimes always never</td>
<td>never</td>
<td>-43%</td>
</tr>
<tr>
<td>110_WOT2_F2006</td>
<td>If I multiply a prime number and a composite number then the result is a composite number</td>
<td>sometimes always never</td>
<td>always</td>
<td>-42%</td>
</tr>
</tbody>
</table>
The observed trend of MQRT of student-experts to reduce over the learning period is in accordance with Thurstone’s (1937) idea that the ability to generate a correct answer can have a different quality which is a function of practice. The results confirmed that the inverse relationship between practice and performance speed can be observed and quantified by means of RT measurements under conditions of formative testing in tertiary education settings. Because this particular test was on basic mathematics skills relevant to the elementary school curriculum, it was possible for students to be expert at the beginning of the course, but to be further exposed to the relevant ideas throughout the subject on teaching mathematics. Students were also expected to undertake private practice of elementary school mathematics during the semester.

It can be suggested that response time can be a useful instrument to monitor an impact of teaching & learning on those students who answer consistently correctly. In the situation where the facility of a question has not changed within a learning period, a change in response time may provide an input into evaluation of the cognitive progress.

4.1.8.2. Novices vs. Student-Experts.

As a next step in the investigation of the impact of teaching and learning progress on performance speed, the study compared end-of-semester response times (RTs) of the students who got a question wrong at the beginning with start-of-semester and end-of-semester RTs of student-experts. Those students will be subsequently referred to as “novices”. The study assumed that wrong answers indicated a lack of knowledge, and in order to produce a correct answer at the end of the semester the students had to learn the material during the semester. In this section the study continues using test WOT2_F2006 to demonstrate the trends that were observed across four formative tests (WOT1_F2006, WOT2_F2006, WOT3_F2006, WOT4_F2006).

Figure 17 presents the data to illustrate the dynamics of performance speed for student-experts as compared to novices. It contains three series that demonstrate MQRT for each question of WOT2_F2006, measured at the beginning of the semester for student-experts and at the end of the semester for student-experts and novices.

As can be seen in Figure 17, the end-of-semester MQRTs of the novices are longer on all but one question than those of the student-experts. On average it took novices \((M = 27, SD = 12.50)\) 17% longer than experts \((M = 23, SD = 10.78)\) to answer the test questions. On the other hand, the end-of-semester MQRTs of the novices were shorter on most questions than the beginning-of-semester MQRTs of the student-
experts. On average the novices were 12% faster at the end of the semester ($M = 27$, $SD = 12.50$) than student-experts ($M = 31$, $SD = 10.62$) at the beginning of the semester.

The data suggest that novices who had to learn new material within the semester could not usually catch up with student-experts who worked on advancing their knowledge during the semester. However, the novices, once they learned the material, were faster than the student-experts at the beginning when their knowledge had not been recently refreshed.

Interestingly, questions 20 and 80 which showed little difference between end-of-semester student-experts and novices are the same questions that exhibited the largest increase in test facility (see Table 12 for the content and Table 11 for facility). This suggests the possibility of teaching and learning impact being so powerful that it advanced novices to the level of student-experts in terms of both accuracy and performance speed.

A surprising outlier in Figure 17 is question 40 (see Table 11, Table 12), with RT of novices being shorter than of experts, but with the facility at the end of the semester being 13% lower than at the beginning, would need further investigation by a teaching expert, which was beyond the scope of this study.

![Figure 17. Comparing mean question response time for experts vs. novices in WOT2_F2006.](image)

The results, presented above, demonstrate that combining accuracy data with RT for analysis on a subgroup level can provide new insights into hidden cognitive
phenomena of teaching and learning across a learning period that are not available from the accuracy data only. Contrasting RT of student-experts with novices allowed the conclusion that on most questions a novice followed the path of a student-expert, where an average RT of a novice on a correct answer was faster than an average RT of a student-expert at the beginning of the semester but longer that an average RT of a student-expert at the end of the semester.

Consequently, two approaches to data analysis can be suggested for further investigation of the potential of RT measurements to inform teaching and learning: (a) analysis of the magnitude of the reduction in RT of a novice at the end of a semester relative to the initial (beginning-of-semester) RT of a student-expert, and (b) analysis of the difference between average RT of a novice and average RT of a student-expert at the end of a semester.

Overall, the results assisted in answering the research question “How can response time data provide insights into dynamics of the group cognitive development through the course?” by proposing that monitoring dynamics of MQRT on parallel test items administered at the beginning and at the end of a learning period can generate additional information about cognitive progress of the group and its learning.
4.2. Meaning of certainty data for test design and teaching

This is the second part of the chapter, which presents results of the investigation into the meaning of certainty data. The following sections will report trialling and refining a scale for certainty measurements in a pilot study, defining directions for data analysis, and investigating the teacher’s perspective on usefulness of certainty data and a preferred format of data presentation.

For two years the study collected certainty, accuracy and response time data on formative and summative testing in real life settings in a university. Collecting certainty data was done by asking students to indicate their certainty in the correctness of their own answer on a scale (for more detailed description see Chapter 3) before moving to another test question. Indicating certainty did not affect a test score or any formal record of a student.

Within a few days after testing, the summary of the data was presented to participating teachers and a subject-coordinator to investigate the potential usefulness of the data to inform teaching through identification of latent cognitive states and processes. The feedback on the meaning of different combinations of certainty and accuracy that was obtained from the data and the most useful options for presentation of these data, such as a selection of variables, a type of a chart or a table, was collected through a series of formal and informal discussions over the two years.

As a part of the investigation into the certainty-accuracy relationship, the study explored the difference in the self-reporting behaviour of low and high achievers (as defined by test results) in the context of the particular setting of the experiment.

4.2.1. Defining a scale for certainty measurements: the pilot study.

In order to determine the specification for a scale for certainty measurements and applicable measuring procedures, the pilot study was conducted on a summative test in 2005. To answer the question of how many categories (grades) of certainty would provide sufficient granularity of the measurement, the pilot experiment used a certainty scale with five categories, following the experience of Sinkavich (1995) who used the scale in educational settings similar to this experiment. The scale, as demonstrated in Figure 1, was built into an interface of a computerised test which prompted test takers to indicate their certainty before they would move to another question.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

Table 13 demonstrates the use of categories of certainty by test takers in summative test S2005. It can be seen that only 9% of test takers chose to use all five grades of certainty across the test. Another 11% of test takers did not discriminate their answers by certainty, using only one grade to advance between questions. It is not clear whether they tried to avoid self-assessment as they felt that it slows down their performance, or they were unwilling to expose their self-assessment. Both explanations were suggested in follow-up interviews, though the participants had been assured that their certainty indication would not affect scoring. Most students preferred to use either 4 or 3 categories. Overall, the data suggested that the range of a five category scale was not fully utilised by the test takers.

Table 13.

<table>
<thead>
<tr>
<th>Number of levels of certainty used</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test takers</td>
<td>11%</td>
<td>23%</td>
<td>28%</td>
<td>30%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Further, the study explored how often the test takers used each of the categories and if there were preferences to using a particular range of categories. Table 14 illustrates the distribution of certainty indications between categories. Considering that the overall test facility was 79%, it is not surprising that the distribution of occurrences of different categories of certainty is characterised by the prevailing use of the positive part of the certainty scale over the use of the negative part, as in most cases test takers had reasonable grounds to believe that they were more likely right than wrong. It could be expected that a significant proportion of certainty indications on the remaining 21% of incorrect answers would fall under the two right hand-side categories of the scale (“probably wrong” and “most likely wrong”), however those categories accumulated only up to 6.4% of indications. It can be concluded that in most cases a doubt in correctness was expressed with choosing the category “maybe right maybe wrong” as a more positive alternative.

The concentration of certainty indications in three major categories that was observed by this study matches the results of Shuford (1968) who attempted to use a certainty scale based on all 26 letters of English alphabet, but the most commonly used letters turned out to be “A”, “Z”, “F”, (in order of frequency) that accounted for 83
percent of all uses. Three other letters, “G”, “E”, “M” accounted for a further 7 percent, with the remaining letters having an insignificant number of selections.

In contrast to the Shuford’s study which found that the distribution of confidence scores was shifted more to the negative side of the scale, this study observed prevalence of positive categories, which might be related to the particular instructions to the participants or their knowledge of the domain.

Table 14.
Usage of a five-category certainty scale in summative test S2005 (n=188).

<table>
<thead>
<tr>
<th>Certainty category</th>
<th>Most likely right</th>
<th>Probably right</th>
<th>Maybe right, maybe wrong</th>
<th>Probably wrong</th>
<th>Most likely wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responses</td>
<td>35%</td>
<td>34%</td>
<td>24%</td>
<td>3%</td>
<td>3%</td>
</tr>
</tbody>
</table>

In light of the analysis of the usage of the five certainty categories in the 2005 pilot experiment (test S2005), the decision was made to limit the scale of certainty to three categories in further experiments (see Figure 2, Chapter 3). The two positive categories “most likely right” and “probably right” were joined into one category “most likely correct”, the two negative categories “most likely wrong” and “probably wrong” were joined into the category “probably wrong”, thus reflecting a general trend to use a more positive category. The category “maybe right maybe wrong” was replaced with “not sure” that was supposed to convey the same meaning.

Another alteration of test taking procedure was in making the use of the certainty scale voluntary. The students did not have to indicate their certainty in order to advance to the next page, though they were encouraged to make the best use of the scale by the researcher and the teacher. The usage of the new scale was then re-examined in the analysis of test S2006 data to investigate the impact of the change.

From 203 students who took test S2006, 26 students chose not to indicate their certainty at all. Half of them are high achievers (by total test score) who probably felt that using the certainty scale would slow them down, as some of them mentioned on a follow-up interview. The study did not specifically investigate the motivation behind students decisions to use the certainty scale.

Table 15 describes the range of categories that was used by students who decided to engage the scale of certainty in test S2006 which had similar content and administration arrangements to S2005. As we can see, 18% of students used the full range of the scale, up from 8.5% in S2005 (Table 13). Additionally, 27% of the students used only one category, with two thirds of them using only the category “Most likely
correct”. The interpretation is that those students tended to mark the questions they were confident with, to be able to review other questions at a later stage of the test. This pattern of behaviour, involving dichotomous use of the scale with some students avoiding marking an answer as “Probably wrong” by omitting the certainty indication, was observed mostly with high achievers.

Table 15. *Percentage of test takers using given number of certainty categories in summative test S2006 (n=177).*

<table>
<thead>
<tr>
<th>Number of levels of certainty used</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test takers</td>
<td>27%</td>
<td>56%</td>
<td>18%</td>
</tr>
</tbody>
</table>

Summary of the usage with the breakdown by occurrences of the categories in summative test S2006 is presented in Table 16. It can be seen that the two categories: “Most likely correct” and “Not sure” account for 97% of certainty indications, despite the question facility being on average 84%. Test takers again made disproportionally little use of the category with a negative meaning “probably wrong” in the same way as they did with the two negative categories of the five category scale in the pilot study in 2005. A possible explanation of the preference to indicate “not sure” on incorrect answers due to miscalibration of self-assessment under a condition of partial knowledge will be further explored later in the chapter.

Table 16. *Usage of a three-category scale in summative test S2006 (n>140).*

<table>
<thead>
<tr>
<th>Categories of certainty</th>
<th>“Most likely correct”</th>
<th>“Not sure”</th>
<th>“Probably wrong”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responses</td>
<td>71%</td>
<td>26%</td>
<td>3.2%</td>
</tr>
</tbody>
</table>

To investigate whether the use of certainty categories in a formative test was different from a summative test, the study offered the same scale in ten formative tests (WOT1-WOT10_F2006) in 2006. Additionally, the students were specifically instructed to use the full range of the scale as a means to inform a teacher about their needs. The content of the questions included some tasks based on prior learning and other tasks from the areas that would be covered in coming lessons. Such selection of questions resulted in a wider range of question facilities (from 21% to 99%), which is reflected in lower mean question facility (72%) and larger standard deviation (19%) than on summative test S2006 as can be seen in Table 17.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

Table 17.
*Usage of a three category scale in formative tests F2006* (n>90).

<table>
<thead>
<tr>
<th>Categories of certainty</th>
<th>“Most likely correct”</th>
<th>“Not sure”</th>
<th>“Probably wrong”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responses</td>
<td>58%</td>
<td>35%</td>
<td>7%</td>
</tr>
</tbody>
</table>

* Average question facility (M=72%, SD=19%)

The wider range of question facilities in F2006 tests was also associated with a wider range of certainty indications, with some questions getting up to 62% of “not sure” and up to 25% of “Probably wrong” certainty indications.

Further analysis of the usage in Table 18 demonstrates that the number of the students who utilised the full scale (34%) in the formative tests was nearly twice as high as the number in summative test S2006 (18%). Additionally, the number of the students who used only one category was substantially smaller (5%) which supports the suggestion that most of the students were engaged in usage of the certainty scale.

Table 18.
*Percentage of test takers using given number of certainty categories in formative test series F2006 (n>90).*

<table>
<thead>
<tr>
<th>Number of levels of certainty used</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test takers</td>
<td>5%</td>
<td>61%</td>
<td>34%</td>
</tr>
</tbody>
</table>

The results of the usage of the scale suggest that the granularity of the scale based on three categories (degrees) is sufficient to convey self-assessed certainty values. A three category scale had previously been reported as a sufficiently sensitive instrument by Gardner-Medwin (2006) and is supported by the findings from this study. This is in agreement with the results of Sinkavich (1995) who attempted to use a five category scale and found that only a third of test takers used three or more categories. Increasing the scale up to seven degrees as proposed by Leclercq (1983) and Bruttomesso et al. (2003) does not seem to be warranted by student responses in the context of testing mathematics.

Formative tests demonstrated a greater engagement of students with the certainty scale and a higher discrimination of the responses between test questions than summative tests. This can be partially accounted for by a higher motivation to use certainty categories as an instrument to pass the feedback to a teacher as an input in tailoring tutoring to the students’ needs.

96
Overall, the data suggest that for the purpose of informing a teacher about students’ perception of their knowledge a three category scale offers sufficient utility to provide a range for expressing self-evaluation for most students.

4.2.2. Confirming construct validity of certainty measurements and discussing potentially confounding factors.

The study examined proportions of correct answers in each certainty category to confirm the construct validity of the scale. It can be expected that if the certainty indications reflected actual knowledge then there should be a significant association between correct responses and the certainty indications on most questions.

The summary of the data from pilot test S2005 which used a five category scale for certainty indications is presented in Table 19. The data from test S2006 which used a three category scale is given in Table 20.

As can be seen in Table 19 the proportion of correct answers increases with the increase in confidence, starting with 47% in the category “Most likely wrong” and reaching 82% in the category “Most likely right”. This trend was also observed in S2006 test (Table 20) which data was examined to reconfirm the association between accuracy and certainty indications in connection with the use of a three-category scale. The association was found statistically significant, \( \chi^2(3) = 143.7 \) \( p < .001 \), which confirms that in general self-assessed certainty was positively associated with actual performance and thus certainty measurements have internal construct validity.

Interestingly, the correlation between the proportions of correct answers and the proportion of the responses marked as "Most likely correct" (high confidence), when calculated for each question of S2005, was weaker \( \text{Pearson } r = 0.34, p < 0.001 \) than the correlation between the proportion of incorrect answers and the proportion of the answers marked as “Probably wrong” \( \text{Pearson } r = 0.48, p < 0.001 \). It has to be noted that the proportion of correct responses that were marked as “Most likely wrong” was surprisingly high in S2005 (47%) and even higher when a modified scale was offered in S2006 (61%).
Table 19.  
*Number of correct and incorrect responses by certainty category from test S2005.*

<table>
<thead>
<tr>
<th>Certainty category</th>
<th>Responses by accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>incorrect</td>
</tr>
<tr>
<td>Most likely wrong</td>
<td>57</td>
</tr>
<tr>
<td>Probably wrong</td>
<td>92</td>
</tr>
<tr>
<td>Maybe right, maybe wrong</td>
<td>415</td>
</tr>
<tr>
<td>Probably right</td>
<td>308</td>
</tr>
<tr>
<td>Most likely right</td>
<td>295</td>
</tr>
<tr>
<td><strong>Total count</strong></td>
<td><strong>1167</strong></td>
</tr>
</tbody>
</table>

Table 20.  
*Number of correct and incorrect responses by certainty category from test S2006.*

<table>
<thead>
<tr>
<th>Certainty category</th>
<th>Responses by accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>incorrect</td>
</tr>
<tr>
<td>Probably wrong</td>
<td>30</td>
</tr>
<tr>
<td>Not sure</td>
<td>150</td>
</tr>
<tr>
<td>Most likely right</td>
<td>177</td>
</tr>
<tr>
<td><strong>Total count</strong></td>
<td><strong>357</strong></td>
</tr>
</tbody>
</table>

The evidence suggests that, in general, the choice of a certainty category is modestly related to the probability of a correct answer, which supports the argument that certainty is “a separate psychological and cognitive construct” as was suggested by Trow (1923), Johnson (1939) and most other researchers. The magnitude of the correlation found is in agreement with the previous findings (Dunning, 2005; Sinkavich, 1995) that suggest the perception of a skill being only moderately correlated with actual performance.

The rest of the section discusses a number of potentially confounding factors that would impact response time and certainty measurements in different testing situations and how these factors might have influenced the results in the present study. This part was included as an amendment and aims to clarify the focus of the study on the constructs (RT and Certainty) that are averaged across a group of students.
Certainty as an individual trait

The evaluation of the impact of individual differences on the choice of certainty categories has been a long standing issue of certainty measurements. Understanding of the role of the individual self-confidence factor grew from the initial suggestion of Trow (1923) that when accuracy is controlled there was “no evidence to support the notion that certain individuals were generally more or less confident across tasks” (as cited by Blais, Thompson & Baranski, 2005) to an opposite view (Swineford, 1938; Ahlgren, 1969; Kleitman & Stankov, 2007) that confidence is a broad personality trait.

Most of the recent researchers, for example Kleitman and Stankov (2007), support the notion of a general self-confidence factor as a personality trait. Kleitman and Stankov confirmed previous findings that confidence is related to, yet is independent of, cognitive abilities. It needs to be noted that the results of the present study are in an agreement with the latter.

Effect of prior practice in self-evaluation on appropriateness of the confidence

Sieber (1979) reported that secondary school students who were trained by a teacher to express warranted uncertainty and extensively used a confidence estimation procedure performed better on two tests of spelling, and improved more from Test 1 to Test 2, than did participants who merely used or did not use the confidence estimation procedure.

The recommendation by prior research (Sieber, 1979; Strater et al. 2004) to arrange practice before summative exams to help students calibrate their confidence was utilised in the present study. For 2 weeks before the test the students had unlimited access to a practice test that included certainty indications. Almost all students attempted the practice test at least once. It was expected that practice tests would also help to reduce test anxiety and computer anxiety, and assist the students with familiarising themselves with the testing interface. Feedback comprising correct answers and a total score was automatically generated immediately following the completion of a practice test. The students were then able to compare their score with their certainty indications to adjust their judgement.

Effects of gender and ethnicity

Examination of subgroup differences was considered crucial for applying certainty in testing by a number of researchers. Ahlgren (1969) compared test scores
with a certainty component to conventional test scores for subgroup bias by gender, average confidence, appropriateness of confidence and general test anxiety, and found no significant differences. The study by Kruger and Dunning (1999) investigated the ability of 65 tertiary students to recognise their own competence and that of their peers. Specifically, the authors asked the question whether those who did poorly recognised the low quality of their performance. Their findings suggest that the participants tended to generally overestimate their abilities, with a strong inverse relationship between their knowledge in the particular domain and the ability to self-evaluate their performance. Across all four experiments of the study a gender factor was found insignificant.

Gardner-Medwin and Gahan’s (2003) examination of the data from over 3 million answers recorded on campus computers and in exams did not find significant gender differences in the use of certainty/confidence scale by tertiary students in both formative and summative tests. Summative tests (exams) were characterised by the increase in the percentage of correct responses in each certainty category. The change equally affected males and females which suggests that the change in the testing situation did not introduce a gender bias in self-evaluation either.

However, gender bias in expression of confidence has been reported by some other studies, for example by Stankov, Lee, Paek (2009) who found significant differences in the realism of confidence judgements with respect to gender and race/ethnicity among college students: females demonstrated lower levels of overconfidence than males and Americans of European descent demonstrated smaller levels of overconfidence than Americans of African or Hispanic origin. The authors suggest that significant individual differences in the understanding of subjective probabilities are likely to be an underlying factor of the observed sub-group differences.

As the participants in the present study were extremely diverse from the perspective of ethnic origin, which is characteristic of the contemporary Australian population in general, grouping by the ethnicity factor was considered likely to be misleading. Due to the prevalence of females (over 90%) in the population of education students the study did not analyse the results for the gender factor.

**Effects of risk-aversion or over-confidence, motivation and the type of testing**

Swineford (1938) suggested that the confidence measure is directly associated with the tendency to gamble. In his experiment 160 students were given a 75-item true-false test that used a confidence marking scheme. The author noted that a few excellent
students were timid about using maximum confidence indication, while some of the poorer students used it too often. Swineford concluded that those poor students were trying to claim additional credits, gambling on the items which they did not know. The author warned that “if this tendency to gamble is a personality trait which is not intimately associated with the ability measured, then it should not be permitted to affect the achievement score” (Swineford, 1938, p. 298). It seems that Swineford’s argument should be considered in currently used certainty-based marking to provide distinction between certainty as a measure of knowledge quality and as a measure of individual tendency to take a risk.

Sieber (1979) found that the expression of certainty depends partly on whether it is costly or rewarded to do so. This finding should be taken as a warning about the potential threat to the reliability of certainty measurements. As with any measure based on self-assessment, certainty may be vulnerable to the bias introduced by student motivation. As Ahlgren (1969, p. 5) pointed out, the complex relationship between confidence and risk-taking can potentially introduce bias in results, which can become a major problem if the purpose of testing is not only to inform students but also to judge students or an instructional procedure. Additionally, Gardner-Medwin and Gahan (2003) demonstrated that students use a certainty/confidence scale differently under different conditions of risk-reward. The percentage of correct answers in each certainty category increased by 8-20% under exam conditions compared to practice. This result could be attributed to a higher motivation on a summative test which stimulates accuracy.

It seems unlikely that certainty data in the present study are affected by a risk-taking bias as the participants were informed that their certainty indications had no impact on their scores. This information was reinforced by a teacher during lectures and it was mentioned in a plain language statement given to each participant.

**Effect of overall achievement on appropriateness of certainty rating**

The present investigation also examined whether there was a bias introduced by a level of achievement. According to prior research (Sieber, 1979), high achieving students demonstrate a higher accuracy of confidence estimation and slightly tend to be under-confident. In the Sieber’s experiment, secondary school students exhibited consistent subgroup differences in confidence estimation on spelling, mathematics and reasoning tests. Subgroup differences were also reported by Newman (1984) who examined the relationship between basic numerical skill and confidence on a task of
quantitative estimation. The author found that a counting skill was related not only to accuracy but also to the appropriateness of certainty ratings, with the relationship being stronger for skilful test takers. Kruger and Dunning (1999) examined the links between competence, metacognitive ability and inflated self-assessment and found that the participants who were in the bottom quartile by overall test score not only overestimated their own performance in terms of a number of correct answers but persistently marked their ability as above-average. Those participants were also found less successful than top-quartile participants in the metacognitive tasks of discerning what they had answered incorrectly even after controlling for objective performance.

In section 4.2.3 the present study addresses the potentially confounding issue of certainty bias due to the achievement factor. A detailed analysis of the data from a summative test was undertaken to examine the appropriateness of the confidence, that is, whether there was any difference between how precisely the correctness of a response was evaluated by top, medium and lower third achievers. Our data were in agreement with the previous finding that people misjudge more in evaluating their accuracy when their skills in the domain are limited: when indicating the highest certainty (“Most likely correct”) the responses of low achievers were correct only in 80% of cases while the responses of high achievers were correct in 97%. However, the study also observed another phenomenon: high achievers demonstrated gross under-estimation when assigning category “Not sure” to a response which in 94% of cases turned out to be correct. Lower achievers showed a more realistic view of their lack of knowledge while using categories “ Probably wrong” and “Not sure”.

Investigation of the potential impact of achievement bias on the relationship between response time and certainty/confidence was beyond the scope of the present study but can be recommended for future research.

**Relationship between certainty and accuracy**

As confidence scores reflect the self-monitoring processes of metacognition (Stankov, 1999), it is reasonable to expect some positive correlation between achievement scores reflecting cognitive skills and confidence scores. In the discussion of the construct validity of certainty measurements earlier in this section it is reported that there was a significant association between correct responses and certainty indications on most questions. One of the surprising results was a high proportion of correct responses marked as “Probably wrong” on a 5-category scale (47%) and marked
as “Most likely wrong” on a 3-category scale (61%). The percentages of correct responses at each confidence level (also see Table 20) are presented below:

- “Most likely wrong” = 61%
- “Not sure” = 77%
- “Most likely right” = 91%

It is not clear what makes a student mark a correct response with the label that shows minimum confidence. An argument that students had insufficient motivation to express true certainty as they were not penalised for a false indication by a scoring scheme is not supported by the data from other studies that included the appropriateness of certainty indication into their scoring procedure. For example, in Gardner-Medwin and Gahan’s study (2003) the data about the use of a certainty/confidence scale by tertiary students for summative tests demonstrate the following proportions of correct responses:

- low confidence (C1) = 62%;
- intermediate confidence (C2) = 79%;
- high confidence (C3) = 95%.

The close match between the proportions of correct responses at each confidence level suggests that the motivation factor was unlikely to be of significant influence on the data. Further investigation into the evident underestimation of someone’s knowledge when assigning the lowest category of certainty is required.

**Effect of time pressure**

The time pressure factor was not given consideration because all tests were power tests, that is there was no time limit (on formative tests) or the time allocated was generous (on summative tests) and allowed the completion of the test without experiencing any time pressure. Still, individual difference influencing RT and certainty are acknowledged in section 2.1.1 (dual nature of individual differences), section 2.3 and earlier in this section.

**Effect of question format**

As was demonstrated by the present study the question format is a significant factor in certainty indications. Certainty was significantly lower on questions in the free/open response format than on the true/false or multiple choice format (MC). This can be explained by the possibility that the students can recognise the right answer
under conditions of partial knowledge (defined as knowledge sufficient to reject most wrong options) though they would not be able to produce it.

Quality of distractors could be another factor in certainty indications, as sometimes students can be certain by rejecting alternatives on a MC format question and narrowing the choice to one response. Again, in this case partial knowledge would serve as a determinant for the response and thus MC format questions are likely to be marked with a higher certainty than questions in the free/open response format.

Specifics of the research focus of the present study

Specific research questions and requirements for the feasibility of the study in the context of Australian education settings defined the scope of the study and its limitations. Though the study collected individual RTs and certainty indications, and it recognises that individual differences in performance speed and general confidence are very important for understanding individual behaviours, the individual differences were not considered necessarily relevant to the group test results in a standard testing situation. The variables of interest, such as MQRT or RT distribution parameters, were aggregations of individual RTs across a group. Correlation between RT and certainty were mainly examined on a group/subgroup level. Thus, the discussion about interactions between individual reaction times and the average group reaction time and/or between many other factors is included in the thesis when it is directly relevant to the research questions.
4.2.3. Exploring subgroup bias in self-evaluation.

As the observed high proportion of correct answers in the certainty category “Probably wrong” (section 4.2.2, Table 19, Table 20) stood out from a general trend for people to overestimate their ability that has been previously demonstrated across many cognitive and social domains (for reviews, see Alicke & Govorun, 2005), the study engaged in further exploration of the phenomena by analysing subgroup bias in the perception of the correctness of an answer, which is discussed below.

Two summative tests (S2005 and S2006) were analysed for subgroup difference in the relationship between certainty and accuracy. A detailed analysis was undertaken to examine whether there is any difference between how precisely the correctness of a response is evaluated by top, medium and lower third achievers (as defined by test scores). This section reports the findings from the S2006 test which used a three-category scale. Report on corresponding findings from S2005 which used a five-category scale can be found in Gvozdenko (2007).

Figure 18 demonstrates the proportions of correct/incorrect responses in each certainty category for each of the three subgroups of test takers: low, medium and high achievers. Absolute values can also be found in Table 21 which contains the data for the charts.

The study explored if there is a difference in the use of certainty categories between the subgroups. An analysis of variance showed that the effect of group on certainty was significant, $F_{2,177} = 21.19, p < .001$. Post hoc analyses using the Scheffé post hoc criterion for significance indicated that the average certainty was significantly lower in the subgroup of low achievers ($M = 2.50, SD = .37$) than in the subgroup of medium achievers ($M = 2.79, SD = .21, p < .001$) or a subgroup of high achievers ($M = 2.81, SD = .25, p < .001$). The difference between medium achievers and high achievers was not statistically significant ($p = .88$).

As can be seen in Figure 18, for the low achievement subgroup the total number of incorrect responses in all categories of certainty was higher than for any other group. This could be expected because low achievers were selected because they made more mistakes, so the study focused on proportions of correct/incorrect responses in each certainty category separately. The proportion of incorrect answers given by this subgroup was the largest in category “Probably wrong” (60 %), which suggests that in most cases when the students thought that their response was incorrect, it was, indeed
incorrect. The proportion of incorrect answers in category “Most likely correct” was the smallest (20%).

In the medium achievement subgroup the number of incorrect responses marked as “Probably wrong” was small (5) and in the high achievement subgroup there were no such responses at all. It seems that those two subgroups tended to use category “Not sure” to indicate their uncertainty. As could be expected, the proportion of incorrect answers in both categories “Not sure” and “Most likely correct” in the medium achievement subgroup (Table 21) was larger than in the high achievement group.

The high achievement subgroup is characterised by using mainly two categories: “Not sure” and “Most likely correct” with a very low proportion of incorrect responses (6% and 3% respectively). Considering the high average accuracy in this subgroup, it was not surprising that the category “Probably wrong” was almost abandoned. What surprises is the very low proportion of responses marked as “Not sure” which were incorrect (6%).

Prior research on subgroup difference in self-evaluation and the confidence with which people make their decisions (Kruger & Dunning, 1999; Sinkavich, 1995) found that people are more miscalibrated in evaluating their accuracy when they face difficult
tasks or they do not possess the skills required to solve the problem. Our data support this point in regard to the precision when indicating the highest certainty: responses of low achievers were correct only in 80% when marked “Most likely correct”, while responses of high achievers were correct in 97% of cases. It seems that low achievers do not know precisely where their knowledge ends and thus they overestimate their ability.

Table 21.
Number of correct and incorrect responses by certainty category by achievement from test S2006.

<table>
<thead>
<tr>
<th>Achievement</th>
<th>Certainty</th>
<th>Responses by accuracy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Incorrect</td>
<td>Correct</td>
</tr>
<tr>
<td>Low</td>
<td>Probably wrong</td>
<td>25</td>
<td>17 (40%)</td>
</tr>
<tr>
<td></td>
<td>Not sure</td>
<td>106</td>
<td>214 (67%)</td>
</tr>
<tr>
<td></td>
<td>Most likely correct</td>
<td>98</td>
<td>397 (80%)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>229</td>
<td>628 (73%)</td>
</tr>
<tr>
<td>Medium</td>
<td>Probably wrong</td>
<td>5</td>
<td>18 (78%)</td>
</tr>
<tr>
<td></td>
<td>Not sure</td>
<td>34</td>
<td>136 (80%)</td>
</tr>
<tr>
<td></td>
<td>Most likely correct</td>
<td>56</td>
<td>696 (93%)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>95</td>
<td>850 (90%)</td>
</tr>
<tr>
<td>High</td>
<td>Probably wrong</td>
<td>0</td>
<td>11 (100%)</td>
</tr>
<tr>
<td></td>
<td>Not sure</td>
<td>10</td>
<td>161 (94%)</td>
</tr>
<tr>
<td></td>
<td>Most likely correct</td>
<td>23</td>
<td>724 (97%)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>33</td>
<td>896 (96%)</td>
</tr>
</tbody>
</table>

As could be expected from the definition, high achievers produced a higher rate of correct responses (96%). However, high achievers demonstrated gross underestimation when assigning category “Not sure” to a response, which in 94% of cases turned out to be correct. Lower achievers showed a more realistic view of their lack of knowledge while using categories “Probably wrong” and “Not sure”. Thus, the findings of this study extend prior work by analysing the difference between high and low achievers in situations of different certainty. Overall, the results of the S2006 testing round confirmed the findings of the previous testing round (S2005) and thus validated the statistic of self-evaluated certainty and the approach to the investigation of this research question.
4.2.4. **Discerning the meaning of certainty data for teaching.**

The study aims to provide input in the academic discourse (Trow, 1923; Johnson, 1939; Jacobs, 1971; Kleitman and Stankov, 2007) about the reliability of confidence measurements, the relationship between decision time, accuracy and confidence of individuals, and conditions where people exhibit underconfidence or overconfidence. Certainty data are used to distinguish between well-informed, poorly-informed, and misinformed answers as was suggested by Gardner-Medwin (2006). The study investigates whether the information has added value for practitioners and could be used as one of the available tools for adjusting lesson plans and subject curriculum to tailor it to the needs of the particular cohort.

One of the specific aims of the study was to develop techniques for processing and presenting certainty data to make it useful for teachers and curriculum coordinators. First, the study defined cognitive states that can be determined from combinations of accuracy and certainty data. Then the meaning of those states was synthesized into four categories that were further used to inform teaching and learning.

Table 22 presents the meaning of six accuracy-certainty combinations that emerged from interviews and discussions with a teaching expert (Dr Vicki Steinle). After trialling several chart types for presenting data output the teacher-participants expressed their preference for a Clustered Column chart sub-type (Figure 19) for representing the distribution of responses between accuracy-certainty categories on a single question and a Stacked Column chart sub-type (Figure 19) for visual representation of responses on multiple questions.

Table 22. *Meaning of six combinations of certainty and accuracy.*

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Certainty indication</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>Most likely correct</td>
<td>Mastered state</td>
</tr>
<tr>
<td></td>
<td>Not sure</td>
<td>Consolidating state: ready for practice and revision</td>
</tr>
<tr>
<td></td>
<td>Probably wrong</td>
<td>Receptive state: ready for new learning</td>
</tr>
<tr>
<td>incorrect</td>
<td>Hazardous state: Not ready for new learning (Ambiguous question or misconception)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Receptive state: ready for new learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Receptive state: ready for new learning</td>
<td></td>
</tr>
</tbody>
</table>
Figure 19 illustrates the perception by a teacher-participant of the meaning that is conveyed by certainty-accuracy data output. The correct responses that were marked “Most likely right” were perceived as a confirmation of successfully learned material. The proportion of the correct responses that were marked “Not sure” was viewed as an indication of the number of students that need additional practice or consolidation of prior learning. The most alarming combination was the responses that were incorrect but marked as “Most likely right” as those responses were likely to be caused by some misconceptions as was also demonstrated by Stacey and del Beato (1992).

Identifying the students with probable misconceptions was suggested by teacher-participants as the most important for the purpose of informing lesson planning and curriculum design which coincides with the findings of Ball, Stacey and Pierce (2001). There is a fundamental difference between incorrect responses that are indicated by the student as most likely wrong and those that are indicated as most likely right. In the former case the response reflects a gap in the knowledge, that is the student is aware of it and ready to learn. In the latter case, it is likely that the student is misinformed, that is, the student holds either a false or insufficient knowledge accompanied with a strong but misleading belief in the correctness. This cognitive state requires re-teaching that would address underlying misconceptions, gained in prior learning. Pedagogical approaches and time allocation to those situations are different and it is crucial for a teacher to be informed of the cognitive needs of students in the classroom.

![Figure 19. Interpretation of accuracy across certainty categories on a test item 70 (WOT4_2007).](image)
The four major meanings of certainty/accuracy combinations could be summarised as following:

1. Mastered state is a desirable combination of being correct and certain.
2. Consolidating state is a combination of being correct but not certain yet.
3. Receptive state of being incorrect and aware of it represents readiness for new learning.
4. Hazardous state of being incorrect and unaware of it should be of a special pedagogical concern.

Though the suggested definitions are slightly different from the ones proposed by Bruttomesso et al. (2002) who used “mastered (certainty >90%, correctness >90%), hazardous (certainty >90%, correctness <50%), uncertain (certainty < 50%, correctness >90%) and residual, the main concern about either misinformed or holding misconceptions students is present in both classifications.

Accepting those four meanings as definitions of a cognitive state provides a platform for reporting a summary of individual cognitive states in a group and for analysing the dynamics of the distribution of cognitive states in a group across a learning period. These are the focus of the following two sections.

4.2.5. Analysing responses within a test.

The study examined the utility of certainty measurements to inform a teacher about a group cognitive state on the date of testing. Of special attention were test questions with high proportions of incorrect responses marked as “Most likely correct”. One of ten formative tests (WOT4_F2006) that were administered in the semester is offered below as an example of the processed data.

Figure 20 illustrates the break down of incorrect responses by certainty for the cohort of students (n = 175) that took the subject in 2006 and attempted all questions in test WOT4_F2006 (see Table 23 for content). It can be seen in Figure 20 that the test questions differed in the proportion of incorrect responses (\(M = 25\%, SD = 21\%\)) and in the composition of the incorrect responses in terms of indicated certainty. For example, questions 20 and 70 had a high proportion of incorrect responses that were marked as “Most likely correct” (15% and 31% of the total number of responses on the question, respectively), which should be treated as a special pedagogical concern (see discussion in section 4.2.4). Some questions had a high number of responses that were marked by students as “Probably wrong” and were actually incorrect, for example question 30
(15% of total number of responses on the question), which means that those students realised their lack of knowledge and thus were ready for new learning.

As can be seen from Figure 20, some questions are characterised by both a high proportion of the incorrect responses that were marked as “Most likely correct” and a high proportion of the incorrect responses that were marked as “Probably wrong”. For example, 16% of all responses to question 20 were incorrect and marked as “Probably wrong” and 15% of all responses to the question were incorrect but marked as “Most likely correct”. Both of those proportions are unusually high and so the information was considered very important by a teacher.

The test included two tasks (see Table 23 for content) from the same cognitive domain of conversion of numbers from base 10 to base 6 (question 20) and from base 6 to base 10 (question 30). As it can be seen in Table 23 and Figure 20, question 30 had a higher question facility (55%), with the total number of incorrect answers reduced by the third, compared to question 20. Correct responses on both questions were distributed between certainty categories in a similar pattern: about 45% of correct responses were marked as “Most likely correct”, about 35% were marked as “Not sure”, about 10% were marked as “Probably wrong”, and the remaining 10% did not indicate certainty. However, the breakdown of incorrect answers by certainty revealed that while the number of students who realised that they were wrong remained nearly the same for both questions, the number of those students who were incorrect but sure that their responses were correct approximately halved in question 30.

Question 20 required students to change the base 10 number “45” to base 6 number and the students were expected to follow the solution path:

$$45 \text{ (base 10)} = 7 \times 6 + 3 = (1 \times 6 + 1) \times 6 + 3 = 1 \times 6^2 + 1 \times 6 + 3 = 113 \text{ (base 6)}.$$  

Further examination of actual responses allowed distinguishing a most common incorrect response “73” that accounted for 35% of all incorrect responses, which is a very substantial proportion considering that the question had an open response format. About 30% of those responses were marked as “Most likely correct”, and 45% were marked “Not sure”. Only 11% of incorrect responses were marked as “Probably wrong”, which confirms that most students relied on prior knowledge and believed it to be correct. The misconception behind the answer “73” arises from incomplete knowledge of the concept of bases for number systems. In this case when converting from base 10 to base 6 the students were likely to have divided “45” (base 10) by 6 to get “3” as a remainder which they put after “7” (base 6) which they got by dividing 45.
by 6. They missed the point that every place in a number follows the rule of base-6 and “7” should have been represented as “11”.

Question 30 required students to change the base 6 number “205” to base 10 number and the students were expected to follow the solution path:

$$205 \text{(base 6)} = 2 \times 6^2 + 0 \times 6 + 5 = 2 \times 36 + 5 \text{(base 6)} = 77 \text{(base 10)}.$$  

The most common incorrect response on this question was “125” which accounted for 12% of all incorrect responses on this question. The misconception behind the answer is similar to the one in the previous question: when converting “205” from base 6 to base 10 the students did not appreciate the place value. Instead of using an expression “$2 \times 6^2 + 0 \times 6 + 5$” or “$2 \times 6 \times 6 + 5$”, the students mistakenly multiplied “20” by 6 and added 5.

Interestingly, the composition of incorrect answers by certainty was approximately the same as in the previous question (28% of those responses were marked as “Most likely correct”, and 44% were marked “Not sure), however the facility of this question was higher, than that of question 20. The data for questions 20 and 30 were interpreted by a teacher in lines that conversion from base 10 to base 6 is not only more difficult in general, but in that particular cohort a significant number of students were misinformed (or had incomplete knowledge) in their prior learning. A different pedagogical approach to such students would have to be taken by a teacher once the misconceptions were defined and the students were identified.

Further examination of Figure 20 flags out question Q70 as having an unusually high proportion of incorrect but certain answers. The question asked students to select the smallest fraction from 3/100, 34/1000, 2/10 and 1/10. An error of mistakenly choosing 34/1000 (46% of all incorrect responses) as the smallest number is not a surprise. This misconception was described in prior research (Steinle, 2004) as thinking that the denominator alone determines the size of the fraction. These students are likely to focus only on the numerator or the denominator of the fraction but not both together. The proportion of the students who felt certain that they were correct (nearly sixty percent of all incorrect answers) was perceived by a teacher as a definite signal to re-design the teaching plan to address this error as resulting from a misconception and requiring a re-learning/re-teaching approach.
Another question that was flagged by a high proportion of incorrect but certain responses was question 120 which asked students to select from four fractions (“14/16”; “7/8”; “52/64”, correct; “56/64”) the one that is not another name for the fraction “28/32”. This question generated more incorrect but certain responses than question 30, though the total number of incorrect responses on this question was only half of that of question 30. Incorrect responses were distributed evenly (by 6-7% of all responses on the question) between all three detractors, however the breakdown by certainty revealed that the incorrect responses “7/8” and “14/16” were marked as “Most likely correct” by more students (6 and 7 students correspondingly) than the incorrect response 56/64 (4 students). There may be several possible interpretations of the difference, such as the impact of the position in the list of options, or misreading the task and choosing the first option that is equal (instead of not equal) to the fraction in the task. However, small numbers in this case did not allow the study to make any conclusion.

To confirm the results, the same test was administered to a different group of students of a similar cohort the next year (2007). Figure 21 illustrates the breakdown of incorrect responses by certainty for the cohort of students (n=160) that took the subject in 2007 and attempted all questions in test WOT4_F2007 (see Table 23 for content).

As is clear from visual examination of Figure 20 and Figure 21, those two charts look nearly identical, with a slight increase in a total number of incorrect responses to question 30. The proportions of certainty categories were very similar in 2006 and 2007, which confirms the validity of the findings.

This second test also helped to differentiate between incorrect responses to question 120. In WOT4_F2007 incorrect response “7/8” attracted 11% of all responses to the question and two-thirds of those were marked as “Most likely correct”. This is twice that of the other options. It can be suggested that some students preferred this option because the fraction looks different from other fractions (and to the task fraction) due to single digit numbers. Thus, the analysis of certainty data indicated the area that needs to be addressed before new learning proceeds.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

Figure 20. The percentage of all responses that are incorrect by certainty in test WOT4_F2006.
Note: n = 175

Figure 21. The percentage of all responses that are incorrect by certainty in test WOT4_F2007 (the same content as in WOT4_F2006).
Note: n = 160

Overall, this format of accuracy/certainty data presentation with a breakdown of incorrect answers by certainty categories was found useful to a teacher for dynamic adjustment of a pedagogical approach to the needs of a particular student cohort.
Table 23. *Questions of formative test WOT4_F2006 and WOT4_F2007.*

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Question task</th>
<th>Choices</th>
<th>Correct answer</th>
<th>Hazardous responses*</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>An astronaut from Earth visiting Mars found that the Martians used a base 6 number system as they have 3 fingers on each hand. The Earthling counted the number of Martians in a room and got 45 (in base 10). How would the Martians write this number in their base 6 system?</td>
<td>113</td>
<td>15.4%</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>An astronaut from Earth visiting Mars found that the Martians used a base 6 number system as they have 3 fingers on each hand. The Martian leader wrote a note that said “The Earthlings will be visiting us for 205 days”. What would this number be in base 10?</td>
<td>77</td>
<td>6.9%</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Which of the following numbers is equal to 15.5?</td>
<td>75/5, 31/2, 15/5, 31/5</td>
<td>31/2</td>
<td>1.7%</td>
</tr>
<tr>
<td>50</td>
<td>Which is the smallest fraction:</td>
<td>3/4, 12/15, 4/5, 5/8?</td>
<td>5/8</td>
<td>5.1%</td>
</tr>
<tr>
<td>60</td>
<td>On a number line, 5/9 is closest to:</td>
<td>2, 0, 0.5, 1.4</td>
<td>0.5</td>
<td>2.3%</td>
</tr>
<tr>
<td>70</td>
<td>Which is the smallest number?</td>
<td>34/1000, 3/100, 2/10, 1/10</td>
<td>31.4%</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>Counting on in the following pattern (1 fifth, 3 fifths, 5 fifths, …) the next number is equal to:</td>
<td>5/6, 1.2, 1 2/5, 1 1/5</td>
<td>1.2/5</td>
<td>4.6%</td>
</tr>
<tr>
<td>90</td>
<td>Steve wrote on a test that 4/5 is bigger than 8/10 because “you only need one more to make a whole, instead of two more”. Which of the following numbers would Steve be likely to select as equal to 7/8?</td>
<td>7/9, 6/8, 14/16, 5/6</td>
<td>5/6</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

(table continues)
Table 23. (continued)

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Question task</th>
<th>Choices</th>
<th>Correct answer</th>
<th>Hazardous responses*</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Jill wrote on a test that 2/7 is bigger than 3/4 because “7 is bigger than 4”. Which of the following numbers would Jill be likely to select as bigger than 3/5?</td>
<td>1/9, 2/3, ½, 3/4</td>
<td>1/9</td>
<td>1.1%</td>
</tr>
</tbody>
</table>

| 110         | The letters “A” and “T” make up 2/3 of which of the words listed below?        | ATTENTION TATAMI TAUGHT ATOM | TATAMI         | 5.7%                |

| 120         | Which of the fractions below is not another name for the fraction 28/32?       | 14/16, 7/8, 52/64, 56/64      | 52/64          | 9.7%                |

Note. *Hazardous responses are expressed by the proportion of incorrect but marked as “Most likely correct” responses out of the total number of responses on the question.

As was suggested by Steinle, Stacey and Chambers (2006), students can be easily helped to expertise if subjected to ‘targeted’ teaching that addresses prior misconceptions. Identification of the learning areas that are characterised by a high proportion of incorrect responses followed by identification of the students who do not recognise a lack of knowledge can be especially valuable for targeted teaching.

This section has demonstrated that certainty data from a test can provide insights into otherwise hidden cognitive states of a student group by revealing learning areas that are characterised by high proportions of “hazardous” responses. This can allow differentiating test questions not only by question facility, but also by student readiness for new learning, and thus serve as a guide to teachers for adjusting lesson plans or selecting appropriate methods and activities.
4.2.6. Identifying “hazardous” content areas across curriculum.

In order to determine the criterion for flagging a test question cognitive area as “hazardous” the study examined the relationship between the total number of incorrect responses and the number of incorrect responses that were marked as “Most likely correct”. This would allow establishing the conditions when the question responses should be considered abnormal and thus “alerting”. Two approaches and corresponding visual representations were developed: (a) a linear regression model aiming at predicting a normal range for the incorrect responses, marked as “Most likely correct” from the total number of incorrect responses; and (b) a regression with a fitted power curve aiming at detecting the questions with unusually high odds of certain responses being incorrect. The data from 10 formative tests (F2006) were used to investigate the relationships.

The scatter gram in Figure 22 demonstrates the results of a linear regression analysis that was applied to the data to determine on which items the number of incorrect responses marked as “Most likely correct” should be considered outstanding. It can be seen that there is a reasonably linear relationship between the number of those students who made a wrong answer but were sure that they were correct and the total number of incorrect responses on a test question. The figure displays the fitted regression line with 95% prediction intervals for new observations.

Figure 22. Number of incorrect responses marked as “Most likely correct” vs. total number of incorrect responses in WOT1-WOT10_2006 test questions.
The second approach explored the relationship between the percentage of incorrect responses on a test question and the odds of a certain response being incorrect. The scatter gram and fitted curve in Figure 23 demonstrates that the relationship is quite strong using power function $y = 0.5611x^{0.5227}$. In this case the power curve provided the best fit of all the functions considered. It has to be noted that any question with a high percentage of incorrect but certain responses should be a pedagogical concern, however for the purpose of prioritising between the questions with an equal percentage of incorrect responses the regression line can provide an objective measure to distinguish the “most hazardous” questions. It could be suggested that the points clearly to the right-hand side of the line represent the questions with an abnormally high ratio of incorrect responses marked as “Most likely correct”. The study identified those questions (Table 24) for further investigation which is presented in section 4.3.

Figure 23. Odds of a certain response being incorrect vs. percentage of incorrect responses in WOT1-WOT10_2006 test questions.

Most of the questions that were categorised as “hazardous” by identifying the points above the upper prediction line in Figure 22 were also categorised as “hazardous” in Figure 23 and thus both representation formats deliver useful information.

In summary, the study suggests that the examination of certainty indications on test questions aggregated across a learning period can deliver additional information about a cognitive state of the student cohort. As was demonstrated in section 4.2.5 (Figure 20 and Figure 21) a certainty characteristic of a test question remains stable in
similar cohorts as it is likely to be associated with prior learning experience and, on a group level, the distribution of certainty reflects the strengths of a current curriculum. Identifying a substantially higher than normal proportion of certain but incorrect responses would allow targeting specific cognitive domains in curriculum coordination, so that those domains could be given more teaching emphasis in future curriculum.

4.3. Relationship Between Response Time and Certainty.

After examining different meanings that are conveyed by response times and certainty statistics and directions of the data analysis, the study moved on to investigate the relationship between the two variables. One of the aims of the investigation was to establish a potential of response time to deliver, at least partially, the information that was extracted from certainty indications in the situation when a high proportion of incorrect responses was associated with high certainty in the correctness of a response. The value of this would be to replace a measure of certainty which has a subjective nature and requires an effort and willingness to express it, by an automatically generated measure of the test-taker’s interactions with a test interface.

The study suggested two hypotheses concerning the relationship between RT and certainty: (a) longer RTs are associated with lower certainty; (b) odds of a fast response being incorrect are associated with odds of a certain response being incorrect. The hypotheses were tested on the data from both summative and formative tests and the results demonstrated consistency across the experiment.

The hypothesis about the association between longer RTs and lower certainty was first tested on the data from S2006 test (n > 200). Following the advice of a consultant from the University of Melbourne Statistical Consulting Centre, the study employed a generalized linear mixed model to examine the relationship. As described by McCulloch (2003, p. 29), such models incorporate random effects into the linear predictor portion of a generalized linear model. This allows accommodating correlation in the context of a broad class of models for non-normally distributed data and thus overcomes the need to normalise data through transformation. For the purpose of this statistical analysis, certainty indications were converted into binary scale: high (“Most likely correct”) and low (“Not sure” or “Probably wrong”).

119
The analysis found that there is a statistically significant relationship between response time and certainty. Probability of high certainty was found to decrease with an increase in response time according to the following equation:

$$\log_e\left(\frac{p}{1-p}\right) = 1.259 - 0.01197\text{RT}$$

where \( p \) = probability of high certainty (“Most likely correct”);

$$\frac{p}{1-p}$$ = odds of probability of high certainty;

RT = response time

From the formula it can be said that the odds of high certainty decrease for each additional second of RT by the following factor:

$$e^{-0.01197} = 0.988$$ which is a decrease in the odds by 1.1%

Then, the study examined whether a longer RT is associated with a lower certainty on a formative test. For each of 104 test questions from a series of 10 formative tests (F2006) mean response time was calculated separately for the responses that were marked as “Not sure” and for the responses that were marked as “Most likely correct”. Those two certainty categories were chosen due to the reason that they clearly denote difference in certainty and contain a mixture of correct and incorrect responses. A t-test for paired samples was performed to establish if there was a statistically significant effect of certainty. The average MQRT for the responses that were marked as “Not sure” was larger \((M = 34, SD = 13, n = 104)\) than average MQRT for the responses that were marked as “Most likely correct” \((M = 30, SD = 12, n = 104)\), \(t(104) = 8.0, p < .001\). The effect size (Cohen's \(d = 0.32\)) was significant in education terms.

Thus, analyses of the data from both summative and formative tests established that, in general, the responses marked as “Not sure” are more likely to be slower than the responses marked “Most likely correct”.

Building on this finding, the study hypothesised that fast incorrect responses are associated with high certainty. To test this hypothesis correlation analysis was performed on the data from ten F2006 formative tests. A statistically significant positive correlation between the number of fast incorrect responses (defined as the responses that are faster than mean RT for correct responses on the question) and the number of incorrect responses marked as “Most likely correct” was found \((r_{104} = .85, p < .001)\). As
an effect size, such a correlation is considered to be large based on Cohen (1992) criteria.

As both variables were also strongly correlated with the total number of incorrect responses, multiple regression was also performed to determine whether that variable had a moderating effect on the relationship between the number of fast incorrect and the number incorrect with high certainty (marked as “Most likely correct”). In order to remove the influence of the total number of incorrect responses, this variable was entered in the first step of a multiple regression and explained 69% of the variance in the variable “number of incorrect responses marked as “Most likely correct”, $F_{1,106} = 233, p < .001$. The partial regression coefficient was statistically significant ($B = 0.42, t_{107} = 15.2, p < .001$). The variable “number of incorrect responses with RT that is shorter than average RT for correct responses” was entered second and explained a further 5 percent ($F_{2,106} = 144, p < .001$) of the variance. The partial regression coefficient was statistically significant ($B = 0.46, t_{107} = 4.2, p < .001$).

The second hypothesis about the association between odds of a fast response being incorrect and odds of a certain response being incorrect was based on the findings described above. The hypothesis was tested on the data from 10 formative tests (WOT1_F2006 – WOT10_F2006) collected during the first week when the test was offered to students ($n > 200$). Regression analysis found a strong linear relationship ($r^2 = .904, F_{1,102} = 955.42, p < 0.001$) between the odds of fast responses being incorrect and the odds of the responses marked as “Most likely correct” being incorrect. Figure 24 demonstrates the association between the variables.
This finding also contributes to the discussion of the difference in mean response times (MQRT) for correct and incorrect responses (Hornke, 2000; Chang, Plake & Ferdous, 2005; González-Espada & Bullock, 2007) by suggesting that some variance in MQRT can be attributed to the variance in distribution of certainty. As more certain responses tend to be faster, there a higher proportion of certain responses in all incorrect responses, and the shorter is MQRT for the incorrect responses (MQRT_{incorrect}). Additionally, a substantial number of correct responses produced under low certainty conditions can increase MQRT for correct responses (MQRT_{correct}). A shift in the balance of certainty on correct and incorrect responses is likely to be a factor in the cases when MQRT_{incorrect} is shorter than MQRT_{correct}, however the dual nature of this shift does not allow a meaningful interpretation of the difference.

Odds of a fast response being incorrect, on the other hand, provide a good indication of the odds of certain responses being incorrect, and thus can be used to discriminate the test questions that generate an irregularly high number of incorrect but certain responses. Those questions were referred to as “hazardous” from a teaching perspective (see section 4.2.6) as they are likely to be associated with misconceptions and require a special pedagogical approach. A method of identification of “hazardous”
learning areas from certainty statistics was suggested in section 4.2.6 which discussed how a regression line can be used to obtain an indication of normal behaviour.

On the basis of the established (by the study) relationship between the odds of fast responses being incorrect and the odds of certain responses being incorrect, the study developed a method for detecting “hazardous” test questions from the data generated by logging response times.

The scatter gram in Figure 25 demonstrates that the relationship between the percentage of incorrect responses on a test question and the odds of a fast response being incorrect can be expressed as a power function, with the fitted function being $y = 0.518x^{0.6186}$. The power function provided the best fit of the curves considered. From the relationship between fast and certain responses it can be expected that the responses with higher odds of a fast response being incorrect would be associated with higher odds of a certain response being incorrect. Similar to the approach taken for identifying the “hazardous” questions with an abnormally high ratio of incorrect responses marked as “Most likely correct” (Figure 23) the study suggests that the points clearly to the right-hand side of the regression line in Figure 25 also represent the “hazardous” questions with an abnormally high ratio of fast incorrect responses.

*Figure 25. Percentage of incorrect responses vs. odds of a fast response being incorrect.*
Close examination of the “hazardous” questions that had been identified with certainty (Figure 23) and response time (Figure 25) demonstrated a high degree of agreement (8 out of 10 questions) between the selections. Table 24 presents the list of the top 10 questions that were flagged by an irregularly high percentage of certain but incorrect responses and their corresponding response time data.

Table 24. 
*Questions-outliers identified with regression analysis of certainty and response time.*

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Certainty</th>
<th>Response time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds of a certain response being incorrect</td>
<td>Identified as “hazardous”</td>
</tr>
<tr>
<td>100_W1_F2006</td>
<td>2.88</td>
<td>Yes</td>
</tr>
<tr>
<td>90_W1_F2006</td>
<td>2.64</td>
<td>Yes</td>
</tr>
<tr>
<td>80_W2_F2006</td>
<td>1.76</td>
<td>Yes</td>
</tr>
<tr>
<td>70_W2_F2006</td>
<td>1.37</td>
<td>Yes</td>
</tr>
<tr>
<td>100_W2_F2006</td>
<td>1.10</td>
<td>Yes</td>
</tr>
<tr>
<td>70_W4_F2006</td>
<td>1.00</td>
<td>Yes</td>
</tr>
<tr>
<td>60_W10_F2006</td>
<td>1.21</td>
<td>Yes</td>
</tr>
<tr>
<td>40_W3_F2006</td>
<td>0.89</td>
<td>Yes</td>
</tr>
<tr>
<td>90_W6_F2006</td>
<td>0.85</td>
<td>Yes</td>
</tr>
<tr>
<td>60_W7_F2006</td>
<td>0.74</td>
<td>Yes</td>
</tr>
</tbody>
</table>

From the distributions of data in Figure 23 and Figure 25 it can be noted that the discrimination of the questions by certainty or response time becomes more powerful with the increase of the percentage and total number of incorrect responses. This fits the needs of educational testing as the identification of “hazardous” learning areas is also more crucial with a larger number of students.

Overall, the data allows a conclusion that the proportion of incorrect among fast responses on a test question can be used as a valid indicator of the proportion of incorrect responses among certain responses. This finding provides a theoretical foundation for utilising the full potential of unobtrusive RT measurements to obtain some educationally important information that is offered by certainty measurements.
CHAPTER 5. CONCLUSIONS

This study started with an enthusiastic but neutral theoretical stance aiming at examining whether a current technologically rich academic environment can offer new quantifiable measures that can assist all stakeholders, but especially teaching practitioners in evaluating and directing teaching and learning. Through monitoring the dynamics of response time in its relationship with certainty, the study was able to establish a number of quantitative regularities in test takers behaviour. Qualitative insights from interviews with teaching academics allowed establishing the baseline of the applicability of RT measurements to inform teaching about individual and group cognitive states.

For both response time and certainty measurements, the study confirmed the reliability and validity of the statistics and discussed the requirements to data preparation. The specific types of data and the forms of the data presentation that were perceived useful by teaching practitioners comprise the highlight and manifest the significance of the study.

The following three sections of the chapter present conclusions by research area, namely: meaning of response time, meaning of certainty and relationship between response time and certainty.

5.1. Response Time

The study established that relatively simple statistical procedures, such as calculating descriptive statistics (MQRT, SD), creating a histogram of distribution or running analysis of variance can provide very rich information about test item performance for test design and teaching.

The study demonstrated that a visual examination of MQRTs of parallel test items can provide initial clues to detecting difference in item functioning. Further investigation of the difference was performed with ANOVA and allowed evaluation of the difference in terms of its significance and magnitude. Additionally, examination of shapes of response time distributions in histograms was found helpful for detection of the difference. Detection of difference in response times enabled targeted investigation of the causes of the difference, such as the difference in question complexity, in a set of required cognitive skills or in solution strategies used by test takers.
Practical implementation of this approach empowered the investigation of a cognitive difference between parallel test items which required conversion from square metres to hectares. As a result of the investigation, the difference was attributed to the additional skill of handling a decimal point in a division operation.

Examination of the potential and the limits of response time data to predict the range of solution strategies on correct answers extended the investigation into the difference between parallel items. The study demonstrated that a total time required to complete a question could be further segmented by steps for each operation involved. Exploring different requirements of each step is shown to provide interesting insights into students’ cognitive behaviours. However, the detection facility of RT measurements was found to be effective only when the solution strategies differ significantly in time requirements, which constitutes operational limits for RT measurements.

The study examined utility of response time for detection of rapid-guessing and cheating behaviours. Accuracy statistics for extremely fast responses on a question allowed evaluating the amount of such behaviours in a series of formative tests. The levels of both rapid guessing and fast cheating were surprisingly low for an unsupervised test in online formative testing, though the number slightly increased by the end of the semester. Monitoring of the behaviours related to low motivation was perceived as a useful tool by a teacher.

The study investigated the change in MQRT on a group and a subgroup level over a semester. It found a general trend for group MQRT to reduce over the semester which was observed on almost all test questions. This suggests that the change is associated with the impact of teaching and learning which would include both the feedback from the first test administration at the beginning of the term and the instruction which was partially guided by the statistics provided to a teacher after the first test.

The magnitude of the change differed in different learning areas and was not associated with a question format. Investigation of the particular factors that determined the scale of MQRT reduction was beyond the scope of the present study but can be highly recommended for future research.

Investigation of subgroup differences in response time between student-experts and novices revealed that at the end of the semester on most questions an average RT of a novice on a correct answer was faster than an average RT of a student-expert at the beginning of the semester but longer than an average RT of a student-expert at the end.
of the semester. This can be suggested as a proof that learning happened even for those students who already had some prior knowledge of the subject. This approach to analysing subgroup differences in the dynamics of RT across a learning period can be suggested for future studies into the potential of RT to deliver insights in the group cognitive states and learning progress.

5.2. Certainty

Utility of certainty measurements was another area of the investigation. Firstly, the study determined the specification for a scale of certainty, confirmed reliability and construct validity of the statistic. A trial of two different scales of certainty determined that a three category scale was preferable for the purpose of the study. Repeatability of the measurements for the same cohort and sensitivity to a change in students’ background and educational level were tested to confirm the overall validity of the measurements.

Then, the study proceeded to investigating research questions. It explored subgroup bias in the usage of the scale and found that the precision of self-evaluation by indicating certainty was achievement-related. The higher achievers demonstrated a higher precision in indicating correct answers, than low achievers, however they tend to underestimate their ability under the condition of low certainty. Low achievers were demonstrated to tend to overestimate their ability under conditions of high certainty but were found to be more realistic when they indicated the lack of knowledge.

Through a number of interviews with teaching experts the study discerned the meaning of six combinations of certainty (three levels) and accuracy (correct/incorrect) for teaching and subject coordination. The combinations were found to carry one of four distinctively different meanings about student cognitive state: (a) mastered state, no learning required; (b) consolidating state, more practice required; (c) receptive state, ready for new learning, and (d) hazardous state, need to address possible misconception.

Analysis of certainty responses within a test was shown to provide a useful guidance to a cognitive state of the student group. The break down of the number of incorrect responses by certainty assisted teachers in selecting appropriate pedagogy to address student learning needs, for example the high ratios of incorrect responses that were marked as “Probably wrong” pinpointed learning areas that require new learning, and the high ratios of incorrect responses that were marked as “Most likely correct” indicated the possibility of a hazardous cognitive state. In the latter case the teacher was
able to adjust future lesson plans to address probable misconceptions that stood behind certain but incorrect answers.

Analysis of certainty data aggregated throughout a semester allowed constructing a prediction model for normal distribution of certainty. The relationship between the total number of incorrect responses on the question and the number of incorrect but certain responses was established, which provided a baseline for detecting hazardous cognitive states that are characterised by abnormally high odds of a certain response being incorrect. Identification of those questions would allow targeting specific cognitive domains in curriculum coordination, so that those domain could be given more teaching emphasis in future curriculum.

The study suggests that certainty indication is a valuable element of educational testing and has the potential to guide ongoing adjustment of lesson planning and curriculum by targeting specific content areas and allowing informed and group-tailored choice of teaching techniques and approaches.

5.3. Relationship between Response Time and Certainty

To further synthesize the findings, the study examined the relationship between response time, certainty and accuracy with the purpose of determining the potential of response time measurements to deliver, at least partially, the information that was extracted from certainty measurements.

Analysis of data from both summative and formative tests found that that there is a statistically significant relationship between response time and certainty, with longer RT being associated with lower certainty. Further, the study found that the ratio of the probability of an incorrect response to the probability of a correct response among fast responses is strongly correlated with the ratio of the probability of an incorrect response to the probability of a correct response among certain responses. By other words, there is a strong positive linear relationship between the odds of a fast response being incorrect and certain response being incorrect. On the basis of the established (by the study) relationship, the study developed a method for detecting “hazardous” test questions from the data generated by logging response times. A trial identification of “hazardous” test questions on the basis of response time demonstrated high agreement with such identification on the basis of certainty, which confirms the validity of the method.

As calculation algorithms can be easily automated and built into testing software, making response time statistics available to teachers is a practical
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

recommendation from the research to software producers. Assessment For Learning is likely to benefit most from introduction of response time measurements into formative testing as this will provide valuable insights into latent cognitive states and processes.

5.4. Limitations and recommendations for future research

The strength of the study is in being conducted in a real classroom situation where the data collected bears the impact of the full range of variables, characteristic of a contemporary educational environment. However, there are inherent limitations of this setting, starting with the design type that it is not purely experimental, and thus it does not operate experimenter-manipulated independent variables and so the causal relations cannot be definitively concluded from the results.

The scope of the study is limited with the area of testing (mathematics) and the education level and specialisation of the population (tertiary education students and teachers). Because most questions of formative tests were on basic mathematics skills relevant to the elementary school curriculum, it was possible for students to be experts at the beginning of the course, but to be further exposed to the relevant ideas throughout the subject on teaching mathematics. Students were also expected to undertake private practice of elementary school mathematics during the semester.

Future research is needed to examine the consistency of response time measurements in other cognitive domains. The behaviour of students of pre-tertiary level on formative tests may differ. The teaching experts whose perception of usability of response time data guided the search for the most appropriate format of results output were experts in teaching mathematics and competent users of the charts of distribution frequencies. It can be expected that not all teachers will be able to easily read data in the format that was found the most useful by our participants.

Some cautions should be mentioned with respect to gender differences in response times in different cognitive domains. The participants of the current study were mostly female and the data were not analysed in regards to gender bias. It can be expected that mean response time would differ for males and females in some areas, such as spatial manipulations, rotations, reflections and others found to be gender biased by prior research.

Investigation into the relationship between response time and certainty is a challenging task. There are many other variables that could potentially have a confounding effect on the relationship, such as ethnicity and cultural background, general and mathematical ability, level of computer literacy and some other indicators.
of individual differences that were not accounted for in this study due to practicality reasons. Further experimental research in a more controlled environment might be able to adjust the results for those unaccounted variables.

This study focused on the potential of RT and certainty data to inform teachers, test designers and curriculum coordinators. Investigation of how individual learners can benefit from RT data was beyond the scope of this study, though the demand in this area is rapidly growing and further research would be highly recommended. For example, the following questions could be explored: “what would be the impact of RT information on a student’s motivation?”; “will the information about one’s position among peers in terms of RT stimulate greater student application to work?”.

For the purpose of the current study individual data points were aggregated by question and by group/class. The impacts of individual differences, such as personality traits, motivation, computer literacy, computer anxiety, or other confounding factors that may affect the RT and certainty of a test taker were thus not within the scope of the study. The study utilised standard testing procedures and assumed that due to the random distribution of individual differences in the population, the variables of interest, such as MQRT or distribution parameters, would reflect a combination of the impacting factors at a group level rather than be dependant on individual differences.

Prior research suggests that the application of the scoring rules of certainty-based assessment would motivate students to more precisely evaluate their certainty. However, as the test design was beyond the control of the current study, the study focused on investigation of the potential of RT and certainty to inform teaching in the context of standard and more common testing arrangements.

Overall, the potential of response time measurements to increase the precision of evaluation and inform teaching and learning seems to be getting even more research attention as computerised testing is becoming a major component of and an increasingly preferable method of assessment.
5.5. Outcomes

The study found that response time and certainty data deliver information about underlying cognitive phenomena that is not captured by traditional accuracy statistics. The study established that response time measurements can assist in the following education processes:

- designing parallel test items;
- evaluating differences in cognitive loads of test items;
- estimating prevalence of mental over written solution strategies;
- monitoring the impact of teaching on specific cohort subgroups across a semester.

The study found a relationship between response time and certainty which indicates that response time data, unobtrusively collected, can be used instead of certainty. A method for practical implementation has been designed and validated, which enables a teacher to investigate cognitive characteristics of a test item or cohort readiness for new learning in a specific learning area, using procedures developed by the study.

The findings of the study are especially valuable for enhancing the pedagogical practice of Assessment For Learning which stands in the centre of the current education assessment reform. Practical implementation of the findings in educational testing software can facilitate a change in teaching practice by providing the means for the ongoing monitoring of the cognitive dynamics of a student cohort to further enhance the teaching and learning process.
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective
LIST OF PUBLICATIONS PRODUCED OR PRESENTATIONS MADE DURING CANDIDATURE


Gvozdenko, E., & Chambers, D. P. (2007). *Beyond Accuracy: How certainty and timing data from computerised testing can inform teaching practices and test design*. Research and development seminars, Faculty of Medicine, Dentistry and Health Sciences, Biomedical Multimedia Unit, Melbourne University, March 22.


REFERENCES


Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective


Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective


APPENDICES

Appendix 1. Definition of Key Terms

Accuracy is correctness of an answer(s) given by a student(s) on a test.

Automated measurements are the data generated and recorded by a computer in the course of human-computer interactions. It includes taker’s inputs supplemented with a timestamp.

CAA is a term for Computer Assisted Assessment.

CAT is Computerized adaptive testing, that is a testing model where each answer affects selection of the next question, and the test is stopped when a test taker is unable to solve a more difficult question.

CBT is Computer-Based Testing. This is the most accepted term for computer-aided testing.

Entry is a unit of measurement for a discrete action of a test taker. It includes each input of information by a test taker.

Equal questions are parallel questions testing the same set of skills and having a similar cognitive load and time demands.

Experts in the context of the thesis are persons who are highly qualified in Mathematics teaching.

Facility of a question is the average percentage of students who gave correct answers on a question.

MQRT stands for Mean Question response time which is an average response time for all test takers on a question.

Non-adaptive testing is characterised by a fixed list of test questions where selection of questions is not dependant on test taker’s answers.

Performance speed is the construct measured by response time.

RT stands for response time, that is the time it takes a test taker to answer a test question.

SNRT is Standardised and Normalised response time. It is 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: \( \frac{(NRT_{ind} - NRT_{avrg})}{SD} \), where NRT_{ind} is the
Meaning and Potential of Test Response Time and Certainty Data: Teaching Perspective

logarithm of time spent by a test taker on the question, NRTavrg is the logarithm of average time spent by all test takers on the same question, and SD is a standard deviation for logarithms of RTs on the question.

Temporal behaviours describe test taker’s activities in terms of the time spent on each activity. The current study focuses on response time.

Test item is a test design unit which comprises a question stem (instructions) and inputs. An item can represent a group of equal/parallel questions from which a question is selected for presentation.

Test question is an instance of a test item to be presented to a test taker.
### Appendix 2. Content of test S2006v1

Table 25. Content of test S2006v1.

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>Complete the following statement: 23.4 cm = _____ mm</td>
</tr>
<tr>
<td>50</td>
<td>Complete the following statement: 48 mL = _____ L</td>
</tr>
<tr>
<td>60</td>
<td>Complete the following statement: 0.09 t = _____ kg</td>
</tr>
<tr>
<td>70</td>
<td>Add the following quantities, giving the answer in the indicated unit: 750 mm + 225 cm + 1.7 m = _____ m</td>
</tr>
<tr>
<td>80</td>
<td>Add the following quantities, giving the answer in the indicated unit: 5.1 kg + 360 g + 40000 mg = _____ kg</td>
</tr>
<tr>
<td>90</td>
<td>Complete the following statement: 0.7 km² (or km²) = _____ m² (or m²)</td>
</tr>
<tr>
<td>100</td>
<td>Complete the following statement: 27500 m² (or m²) = _____ ha</td>
</tr>
<tr>
<td>110</td>
<td>Complete the following statement: 740 cm³ (or cm³) = _____ mm³ (or mm³)</td>
</tr>
<tr>
<td>120</td>
<td>Complete the following statement: 4700000 m³ (or m³) = _____ km³ (or km³)</td>
</tr>
<tr>
<td>130</td>
<td>Determine the PERIMETER of the given figure. Please use pi=3 and give the answer in the specified unit: (m)</td>
</tr>
</tbody>
</table>

![Diagram]

151
140  Izabella is washing windows. It takes her 5 min to wash and polish 1 square metre area of glass. The room she is cleaning has 8 rectangular windows, 50 cm wide and 1.8 m high. How long will it take her (please, round your answer in minutes)?

150  How many cubic boxes of side length 40 cm can be stored in a shed where the floor measures 2.2 m by 3.5 m and the available height for storage is 2.4 m?

160  How many minutes are there between 11:15am and 4:35pm?

170  Find the volume of the given solid, if D = [D] cm and H = [H] m. Use pi = 3 for your calculation. Please, give the answer in the specified unit: (m3)

180  Susie walks at [n] km/hr. She is going on a [m] km walk to the top of a mountain. How long (in minutes) does it take Susie to walk up the mountain?*

190  Choose which of the following angles is likely to have a measure of 70 degrees.
(table continued)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 200 | ![Diagram of a solid]

What is the correct name of the given solid?

<table>
<thead>
<tr>
<th>210</th>
<th>Please tick all the statements that are true, for the following diagram.</th>
</tr>
</thead>
</table>
| A | ![Diagram of shapes A, B, C, D, F, G, H]

Give the number of lines of reflection in the following shape.

220

Give the number of lines of reflection in the following shape.

(table continues)
<table>
<thead>
<tr>
<th>Number</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>230</td>
<td><img src="image1" alt="Square" /></td>
</tr>
<tr>
<td></td>
<td>Give the number of lines of reflection in the given shape.</td>
</tr>
<tr>
<td>240</td>
<td><img src="image2" alt="Triangle" /></td>
</tr>
<tr>
<td></td>
<td>For the given shape, give the order of rotational symmetry.</td>
</tr>
<tr>
<td>250</td>
<td><img src="image3" alt="Square" /></td>
</tr>
<tr>
<td></td>
<td>For the given shape, give the order of rotational symmetry.</td>
</tr>
<tr>
<td>260</td>
<td><img src="image4" alt="Complex Shape" /></td>
</tr>
<tr>
<td></td>
<td>Please, choose if the following statement for the given shapes is true or false: Figure 1 is a reflection of figure 4</td>
</tr>
</tbody>
</table>

(table continues)
Please, choose if the following statement for the given shapes is true or false:
Figure 3 is NOT a rotation of figure 1

Please, choose if the following statement for the given shapes is true or false:
Figure 2 is a translation of figure 7
(table continued)

<table>
<thead>
<tr>
<th>290</th>
<th><img src="#" alt="Diagram 1" /></th>
<th><img src="#" alt="Diagram 2" /></th>
<th><img src="#" alt="Diagram 3" /></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="#" alt="Diagram 4" /></td>
<td><img src="#" alt="Diagram 5" /></td>
<td><img src="#" alt="Diagram 6" /></td>
</tr>
</tbody>
</table>

Match the following statement with the appropriate diagram:
A'B'C' is the image of ABC after reflecting about the green line L2.

<table>
<thead>
<tr>
<th>300</th>
<th><img src="#" alt="Diagram 1" /></th>
<th><img src="#" alt="Diagram 2" /></th>
<th><img src="#" alt="Diagram 3" /></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="#" alt="Diagram 4" /></td>
<td><img src="#" alt="Diagram 5" /></td>
<td><img src="#" alt="Diagram 6" /></td>
</tr>
</tbody>
</table>

Match the following statement with the appropriate diagram:
A'B'C'D' is the image of ABCD after rotating 90 degrees clockwise about the axis O.
Appendix 3. Investigation of solution strategies on question 110_F2006

Please, select the algorithm that you have used to solve a previous addition problem:

1) *Mental pad:*
   You carried out a pen and paper vertical algorithm
   
   \[
   \begin{array}{c}
   36 \\
   + 45 \\
   \end{array}
   \]
   
   saying to yourself: “Six plus five is eleven. Put down the one. Carry the one. Three and four is seven, and one more is eight.”

2) *Chunking (decomposition):*
   a) You decomposed both numbers into 30 + 40 + 6 + 5 to make 70 + 11.
   b) You chunked 36 and 40 to make 76 and then added on the 5 to make 81.
   c) You took the 6 from the first number and added to the second number.
      Then you got 30 + 51 = 81
   d) You took the 5 from the second number and added to the first number.
      Then you got 41 + 40 = 81

Please, fill in the index for your algorithm or describe your algorithm if it is different from the above.

Table 26. 
Results of ANOVA for three parallel versions of test S2006.

| Question ID | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 | 120 | 130 | 140 | 150 | 160 | 170 | 180 | 190 | 200 | 210 | 220 | 230 | 240 | 250 | 260 | 270 | 280 | 290 | 300 |
|-------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Count v.1   | 55 | 53 | 59 | 58 | 61 | 59 | 59 | 60 | 59 | 56 | 58 | 56 | 59 | 60 | 57 | 53 | 56 | 54 | 51 | 53 | 58 | 59 | 55 | 57 | 55 | 51 | 44 |
| Count v.2   | 58 | 55 | 56 | 59 | 57 | 58 | 58 | 62 | 65 | 66 | 64 | 66 | 65 | 64 | 62 | 65 | 64 | 66 | 62 | 64 | 62 | 63 | 63 | 66 | 59 | 56 |
| Count v.3   | 56 | 59 | 54 | 55 | 60 | 60 | 61 | 59 | 69 | 70 | 64 | 69 | 67 | 68 | 67 | 68 | 64 | 66 | 67 | 69 | 64 | 70 | 70 | 66 | 65 | 64 | 56 |
| Average RT v.1 (in sec) | 34 | 40 | 33 | 70 | 77 | 70 | 48 | 63 | 68 | 193 | 119 | 148 | 70 | 104 | 130 | 29 | 31 | 59 | 49 | 24 | 18 | 24 | 35 | 27 | 20 | 52 | 108 |
| Average RT v.2 (in sec) | 34 | 36 | 39 | 70 | 68 | 58 | 48 | 56 | 73 | 200 | 171 | 163 | 67 | 109 | 105 | 37 | 35 | 58 | 37 | 25 | 29 | 21 | 32 | 29 | 17 | 57 | 101 |
| Average RT v.3 (in sec) | 33 | 31 | 37 | 73 | 77 | 63 | 39 | 53 | 75 | 199 | 151 | 142 | 65 | 120 | 111 | 37 | 38 | 55 | 35 | 28 | 20 | 18 | 27 | 31 | 19 | 47 | 108 |
| df within Groups | 166 | 164 | 166 | 169 | 175 | 174 | 175 | 178 | 190 | 189 | 185 | 186 | 189 | 190 | 185 | 188 | 182 | 181 | 181 | 183 | 188 | 185 | 183 | 183 | 171 | 153 |
| F            | 0.13 | 3.35 | 1.40 | 0.19 | 1.90 | 2.82 | 3.03 | 2.59 | 1.15 | 0.11 | 9.81 | 1.60 | 0.38 | 1.72 | 4.40 | 3.60 | 2.36 | 0.63 | 5.22 | 0.96 | 8.63 | 2.61 | 4.60 | 1.00 | 1.65 | 2.41 | 0.29 |
| P-value      | 0.88 | 0.04 | 0.25 | 0.83 | 0.15 | 0.06 | 0.05 | 0.08 | 0.32 | 0.89 | 0.00 | 0.20 | 0.69 | 0.18 | 0.01 | 0.03 | 0.10 | 0.53 | 0.01 | 0.38 | 0.00 | 0.08 | 0.01 | 0.37 | 0.20 | 0.09 | 0.75 |
| F crit       | 3.05 | 3.05 | 3.05 | 3.05 | 3.05 | 3.05 | 3.05 | 3.05 | 3.04 | 3.04 | 3.04 | 3.04 | 3.04 | 3.04 | 3.04 | 3.05 | 3.05 | 3.05 | 3.05 | 3.05 | 3.04 | 3.05 | 3.05 | 3.05 | 3.05 | 3.05 | 3.06 |