A Semi-Automatic Computer-Aided Assessment Framework for Primary Mathematics

Adewale Oluwafemi Adesina

A Doctoral Thesis
Submitted in partial fulfilment of the requirements for the award of

Doctor of Philosophy Degree of Loughborough University

January 2016

To my wife Kikelomo and lovely children Feranmi,Juwonlo and Jomiloju . You all came during my research journey and have been a blessing.

To my loving parents and siblings.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Adewale Oluwafemi Adesina
January 2016

Acknowledgements

Above all, I say a big thank you to the Almighty God for life, strength and His visible hand throughout the journey. Thanks again for the unspeakable gift.

Abstract

Assessment and feedback processes shape students' behaviour, learning and skill development. Computer-aided assessments are increasingly being used to support problem-solving, marking and feedback activities. However, many computer-aided assessment environments only replicate traditional pencil-and-paper tasks. Attention is on grading and providing feedback on the final product of assessment tasks rather than the processes of problem solving. Focusing on steps and problem-solving processes can help teachers to diagnose strengths and weaknesses, discover problem-solving strategies, and to provide appropriate feedback to students.

This thesis presents a semi-automatic framework for capturing and marking students' solution steps in the context of elementary school mathematics. The first focus is on providing an interactive touch-based tool called MuTAT to facilitate interactive problem solving for students. The second focus is on providing a marking tool named Marking Assistant which utilises the case-based reasoning artificial intelligence methodology to carry out marking and feedback activities more efficiently and consistently.

Results from studies carried out with students showed that the MuTAT prototype tool was usable, and performance scores on it were comparable to those obtained when paper-and-pencil was used. More importantly, the MuTAT provided more explicit information on the problem-solving process, showing the students thinking. The captured data allowed for the detection of arithmetic strategies used by the students. Exploratory studies conducted using the Marking Assistant prototype showed that 26% savings in marking time can be achieved compared to traditional paper-and-pencil marking and feedback.

The broad feedback capabilities the research tools provided can enable teachers to evaluate whether intended learning outcomes are being achieved and so decide on required pedagogical interventions. The implications of these results are that innovative CAA environments can enable more direct and engaging assessments which can reduce staff workloads while improving the quality of assessment and feedback for students.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

CAA : Computer-aided assessment or Computer-assisted assessment

CBR : Case-Based Reasoning

MuTAT : Multi-Touch Arithmetic Tool

Key terms

As-presented strategy: Strategy used to solve arithmetic problems that pairs numbers the way they occur in a word problem

Case base: The collection of previous problem solving episodes stored in a database.

Detailed feedback: Feedback which supports learners to solve problems on their own instead of just giving a correct or wrong judgment

Number bonds: Useful pairs of numbers, such as 1 and 9 or 3 and 7, both pairs of which add up to 10

Place value: Determined by the position of a digit within a number, for instance in 6135, the value of the 3 is three tens, and the 6 is six thousands

Place-value strategy: Strategy used to solve arithmetic problems that reflects understanding of number-bonds

Random: Strategy used to solve arithmetic problems that does not utilized place-value or as-presented strategies

Semi-Automatic Assessment and Feedback: A combination of automatic and manual assessment and feedback

Solution/Problem-solving steps: All correct and wrong steps that lead to an intermediate or final solution to a problem

Strategy: The meaningful steps students take to solve problems.

Chapter 1

Introduction

Assessment is central to the education process and is one of the most powerful influences on learning and achievement (Black and William, 1998; Hattie, 2008). Assessment has been defined as the systematic collection of information about student learning, using the time, knowledge, expertise, and resources available, to inform a decision about how to improve learning (Walvoord, 2010). How students are assessed affects decisions about grades, placement, advancement, instructional needs, curriculum, and, in many cases, funding (Davies, 2010). Through assessment, teachers gather data about their teaching and their students' learning (Hanna and Dettmer, 2004).

Globalisation and technological innovation changes in the past years have created a new global economy powered by technology driven by information and knowledge (Dahlman, 2007). Today in schools, students and teachers are keen to "go digital", as computer numbers have more than doubled since 2006 and most schools are now "connected". Laptops, tablets and netbooks are replacing desktop computers (Wastiau et al., 2013). Integration of mobile technologies within the education process is on the increase (O'bannon and Thomas, 2014). Information and communication technologies (ICTs) is an umbrella term which includes radio and television, as well as the newer digital technologies such as computers, mobile devices, audio-visual systems, applications and the Internet which enable users to access, store, transmit, and manipulate information (Zuppo, 2012). ICTs have been touted as potentially powerful enabling tools for educational change and reform, offering opportunities for the provision of varied assessment formats which comprehensively assess students' knowledge, skills, competencies and thinking.

Computer-Aided Assessments (CAA) refers to the use of computer technologies to assess students (Bull and Mckenna, 2000).In general, however, computer-aided assessment tend to

replicate traditional assessment formats, which focus on knowledge rather than skills and attitudes, and are not usually employed as a means of supporting more personalised, engaging, collaborative or authentic tasks (Hosseini et al., 2014; Jeong, 2014; Rodríguez-Gómez and Ibarra-Sáiz, 2015). As a result, there have been renewed calls for assessment practices to keep up with theoretical, cultural and technological developments affecting teaching and learning (Broadfoot, 2007; Rodríguez-Gómez and Ibarra-Sáiz, 2015). This study investigates how modern and emerging technologies can provide support systems to help teachers adopt and implement technology to improve assessments within the classroom.

1.1   Motivation of study

Teaching and learning with technology have been on a steady increase from the mechanical teaching machines of the 1950s to the computer-based instruction applications of the early 1980s and more recently from the internet and world wide web to the currently emerging mobile and wearable technologies. The current technological and societal changes have lead to the emergence of new kinds of students variously describes as "Millenials" (Oblinger et al., 2005), "Net Geners" (Barnes et al., 2007) and, notably, "digital natives" (Prensky, 2001). The student is perceived to exhibit the characteristics of skilled use of technology, active learning rather than passive receiving of knowledge, love of authentic learning experiences rather than contrived tasks, construction of artefacts rather than only instruction and just in time learning (Elliott, 2008). Educational technology can play an important role in supporting teaching and learning, often enhancing it in ways that would be difficult or impossible to achieve in the traditional classroom environment.

Advances in technology are creating new possibilities and promises for assessment and development of environments capable of capturing observations and studying performances. (Bennett et al., 2007; García-Peñalvo, 2008). Technology offers many new opportunities for innovation through rich new assessment tasks, reporting and real-time feedback (Abedi, 2014; Scalise and Gifford, 2006). Additionally, it allows for anytime, anyplace marking, synchronous and asynchronous moderation, and self and peer assessment (Campbell, 2005). Interactive computer-aided assessment tasks are known to be engaging through the immediate appeal of their graphics and the sustained appeal of their interactivity (Richardson et al., 2002). Other benefits of using computer technologies in assessments include dynamic results reporting (Debuse and Lawley, 2016), tracking students' behaviours (Bennett et al., 2007), valid and reliable measures over traditional competencies (Wirth, 2008) and financial advantages by requiring fewer human resources and paperwork (Chua, 2012; Delen, 2015).

Despite the reported potentials and benefits of CAA, current implementations and practices have been criticised for failing to meet up with advances in ICTs. CAA tasks appear to be restricted, testing only surface learning which do not sufficiently engage students in their designs (Baker and Mayer, 1999; Redecker et al., 2013). There is therefore the need for a new generation of computational tools to engage students in assessment tasks while allowing for the extraction of usable assessment data.

Furthermore, more can be done with emerging ICTs to support teachers in classrooms. Hattie's synthesis of over 800 meta-analyses relating to students achievement revealed that the teacher is the single most important factor in student success (Hattie and Timperley, 2007). Teachers are expected to design and execute daily lesson plans for many hours of the school day, orchestrate student learning activities, administer and grade student assessments, develop and implement efficient and effective classroom procedures, and differentiate their approaches for diverse student needs, all while managing the daily range of student behaviours (Arnett, 2013). With all of these responsibilities, few teachers have the time, stamina, or cognitive and emotional capacity to do each job well. The United Nations Educational, Scientific and Cultural Organization estimates that more than 27 million teachers will be needed to achieve universal primary education by 2030 (UNESCO, 2014). Without concerted efforts to support teachers, these chronic shortages will continue to deny the fundamental right to primary education for millions of children for decades to come. One of the reasons advanced for low teacher recruitment, burnout and attrition is that they are often overwhelmed with assessment workloads (Darling-Hammond and Sykes, 2003; Denton, 2003). The advancements in computer technology has huge implications for teachers. Approaches regarding the use of computers for assessments can be classified into three: manual, automatic and semi-automatic computer-aided assessments (Li et al., 2015).

The first approach "manual assessment" also referred to as online systems replicate traditional marking and feedback techniques. All the scoring is done by hand by teachers but using the computer. Proponents of manual systems argue that high-quality teachers guide their students through activities and projects that stretch them to analyse, synthesise and apply what they have learned across academic subjects and into the real world. Also, they provide personalised, qualitative feedback to help students develop their critical and creative thinking (Arnett, 2013). Besides, they create a classroom culture that motivates students by honouring their hard work and making academic achievement socially relevant (Johannesen and Eide, 2000). These human aspects of good education cannot be replaced by computer technology anytime soon. However, manual systems do not scale well as it is impossible to provide high-quality teachers to handle large student populations. Furthermore, manual systems can

be plagued by errors, inconsistencies and this quickly results in teacher fatigue (Denton, 2003).

The second approach is the fully "automatic" computer-aided assessment. Fully automatic systems provide scores and feedback instantaneously to students hiding non-trivial pre-processing efforts. Some advocates claim that computer-based learning and assessments will soon replace teachers (Mitra, 2003). They argue that the accessibility of information through the internet and the advances in online instruction and adaptive learning point to the decreasing relevance of teachers. Although computers are becoming better at providing customised direct instruction and assessment, implementations have been limited to assessment of foundational knowledge and skills. Typically the majority of CAA systems utilise questions and tasks that only test surface knowledge (Bennett, 2011; Hunt, 2012; Jordan, 2013).

The third approach is the emerging semi-automatic assessment. Semi-automatic assessment or human-computer collaboration has been used to describe the cooperation between human markers and the computer with respect to assessments (Bescherer et al., 2011; Herding and Schroeder, 2012; Sargent, J et al., 2004). The computer is used to take away much of the toil of the marking process by detecting similarities between students' responses. The teacher makes the important judgements on what scores and feedback to give while the computer is used to apply the judgement and marks to similar responses. This potentially provides more flexible assessment but not the instant marking and feedback advantages of automatic computer-aided assessments.

Teachers play and will most likely continue to play a major role in computer-aided assessments no matter how sophisticated computers become. Therefore the semi-automatic approach provides an important balance because the element of human interaction in the process cannot be easily replaced. However, the way the interactions take place will change. Today's education system is not likely to meet modern needs at scale until innovations beyond the traditional factory-model classroom and the restricted personalised automated classroom are provided. It is against this backdrop that the needs for of more authentic and alternative forms of assessment that provide more personalised and engaging experiences exploiting the rich possibilities of the prevailing digital landscape arises. This research focuses on a semi-automatic assessment approach that allows technology do what it does best so as to leverage teachers to do what they do best.

1.2   Statement of the problem

Broadly, CAA tasks and questions may be divided into three: selected response, constructed-response and intermediate constrained type tasks. Tasks that use selected response items often expect students to recall specific content information and make appropriate selection(s) from options offered. Typical examples include multiple choice, true/false or yes/no question tasks. Constructed response type tasks require short answers (text or numeric) or extended text responses as well as the manipulation of objects. Students are required to use cognitive skills involving application, evaluation, and synthesis, in addition to basic knowledge and comprehension (Conole and Warburton, 2005; Jordan and Mitchell, 2009). The third type is the so-called intermediate constraint question type (Scalise and Gifford, 2006). This type combines properties of selected and constructed response type questions and include questions such as ranking and sequencing, concept maps and matrix completion (Weinerth et al., 2014).

One key advantage of using selected response formats is that marks and specific feedback can be given automatically to the student for any of the options they have chosen. A disadvantage of this format is the restricted opportunities for students to justify or explain their choice of answers. Regarding constructed response question types, authors like Jordan (2013), Mills et al. (2002) and Pellegrino and Quellmalz (2010) have argued that constructed responses provide opportunities for students to justify or explain their reasoning and so are valuable in assessing higher-order skills. However, the responses tend to be time-consuming to score as they are usually scored manually by teachers rather than automatically by computers. In an analysis of human judgement scoring, Brown et al. (2014) suggested that significant variations tend to occur in the scoring of constructed responses thereby making consistency in marking harder for teachers. Intermediate constraint question types basically seek to reduce the restrictions of selected responses so as to be able to test higher-order skills as in constructed responses.

Findings by Hunt (2012) have indicated that over 90% of CAA systems implements mainly selected response type questions. The consensus is growing that enhanced methods for capturing and connecting student responses to reveal content knowledge, reasoning and inquiry skills are needed in implementations of CAA across different levels of the educational system.

1.3   Context of study

It was long suggested by Bijou et al. (1966) that foundations of educational systems should be hinged on the so- called three" Rs" - **r**eading, w**r**iting and a**r**ithmetic. These functional skills are not just about knowledge in subjects such as english and mathematics; they are also about knowing when and how to use the knowledge in real life situations. According to Lembke et al. (2012), the acquisition of literacy and numeracy skills is central to effective learning in every area of the curriculum and the students' social and community life outside school. The successful development of these essential skills during the primary school years is crucial for educational success in post-primary school and in enabling every individual to realise his or her social and vocational potential. A Gallup poll in 2013 concluded that respondents rated mathematics as the most valuable subject they took in school, ahead of English, science, and history. Specifically, 34 % of those polled viewed mathematics as most valuable followed by English (21%), Science (12%) and History (8%) (Gallup, 2013). This is not surprising as mathematics enables the student to develop an understanding of particular and important dimensions of the physical world and of social interactions. It provides individuals with the means of manipulating, recording and communicating concepts that involve magnitude, number, shape and space, and their relationships (NCCA, 2006).

The National Council of Teachers of Mathematics (NTCM) recommends that the goals of teaching and learning mathematics be expanded to include active student participation, in which reasoning and communication are stressed (NCTM, 2000). Szetela (1992) argued that students are prone to making calculations without explanations or clearly communicating what they have done or what they are thinking. Therefore, as mentioned in Section 1.2 it is not enough to check for right and wrong answers or the use multiple-choice formats for assessments. Consequently, methods for eliciting better communication of students' thinking are needed. With this, more effective assessment and feedback can be provided which measures the quality of work and thinking. This information may help teachers design and implement instructions to promote greater success in problem solving. Thus, an important direction for development and implementation of assessment systems is the design of specific and general technology-based tools that can assist students to communicate their thinking in problem-solving tasks and teachers managing the assessment process efficiently and effectively. This research focusses on CAA system for primary mathematics as a case study for investigating how computer technology can be used to enhance assessment and feedback practices. The system should provide students with interfaces to enter detailed (in-depth) responses which can possibly lead to the detection of strategies and reveal relationships among response items. Strategies are meaningful steps students take to solve problems (Imbo

and Vandierendonck, 2008). Students refine their strategies as they continue to develop mathematically to become more efficient and skilled at solving problems. To help develop effective strategies in primary mathematics, for example connections need to be made on place value, base-ten concepts and number properties (Yeap and Kaur, 2001).

1.4   Aim and scope

The aim of this study is to develop and evaluate a semi-automatic computer-aided assessment framework that can potentially extend and improve on current assessment practices by providing opportunities for detailed feedback. There are two main areas of focus: (1) providing an environment for students to input process and product responses in a problem-solving tasks and (2) supporting teachers in assessment and feedback activities so as to have good consistency and greater efficiency. This thesis focuses on the development and use of a semi-automatic CAA framework for arithmetic word problem-solving in elementary school mathematics. Arithmetic problem-solving is a well-researched area in mathematics education and is important in the development of numeracy and problem solving skills (Torbeyns et al., 2015).

To meet the aim of the research the following objectives will be fulfilled:

- Carry out a literature review to look into the current state of CAA, educational pedagogies, artificial intelligence and case-based reasoning

- Design a novel approach to capturing problem-solving steps for detailed feedback

- Develop an approach for reusing scoring and grading experiences

- Develop an end-to-end semi-automatic assessment framework

- Implement a demonstrator system and test it on the primary mathematics domain

- Apply case-based reasoning problem-solving methodology in assessments

- Deepen our understanding of semi-automatic assessment concepts and principles

The core research questions (referred to as RQ1, RQ2, and RQ3 in Figure 1.1) that guided this study can be broadly stated as follows:

1. How can a CAA assessment environment be designed to capture students' problem-solving processes?

2. Will the data captured lead to the detection of strategies and opportunities for detailed feedback on primary mathematics word problems?

3. How can assessors score and provide feedback on students' responses consistently and efficiently?

1.5   Significance of study

Many schools aim at providing an internet-connected device for each student. There are so-called 1:1 programs that assign a laptop, tablet, or similar device to each student so as to support differentiated instruction, project-based learning, flipped classrooms, enhanced collaboration, practice communication in multiple formats, and other instructional strategies (Green, 1999). Also, international organisations like UNESCO are actively involved in implementing strategies to reduce the digital divide globally (Norris, 2001; UNESCO, 2016). The digital divide has been defined "as the gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard both to their opportunities to access ICT and to their use of the Internet for a wide variety of activities" (OECD, 2001, p. 25).

The findings of this study will be of benefit to society considering the role of assessments in the education as the digital divide decreases. The greater demand for technology in education justifies the need for more effective approaches. Rather than merely layering technology on top of traditional classrooms, this study will provide insights on how to leverage technology to engage students in assessments, amplify the role of teachers, accelerate student learning, and magnify the impact of educators. Schools that adopt and apply the approach derived from this study will be able to assist students and teachers better. Administrators will be guided on what should be emphasised to obtain efficiency and consistencies in marking and feedback activities. For researchers, instructional designers and developers, the study will help uncover possibilities of emerging technologies in the educational process that many were not able to explore.

1.6   Overview of study

The following figure provides a graphical roadmap of this thesis. The first box lists the structure of thesis detailing the sequence of the chapters. The second box outlines the

contributions of the study (to be outlined in Section 1.8) the arrows show how the chapters map to the contributions and how they relate to the broad research questions stated in Section 1.4.
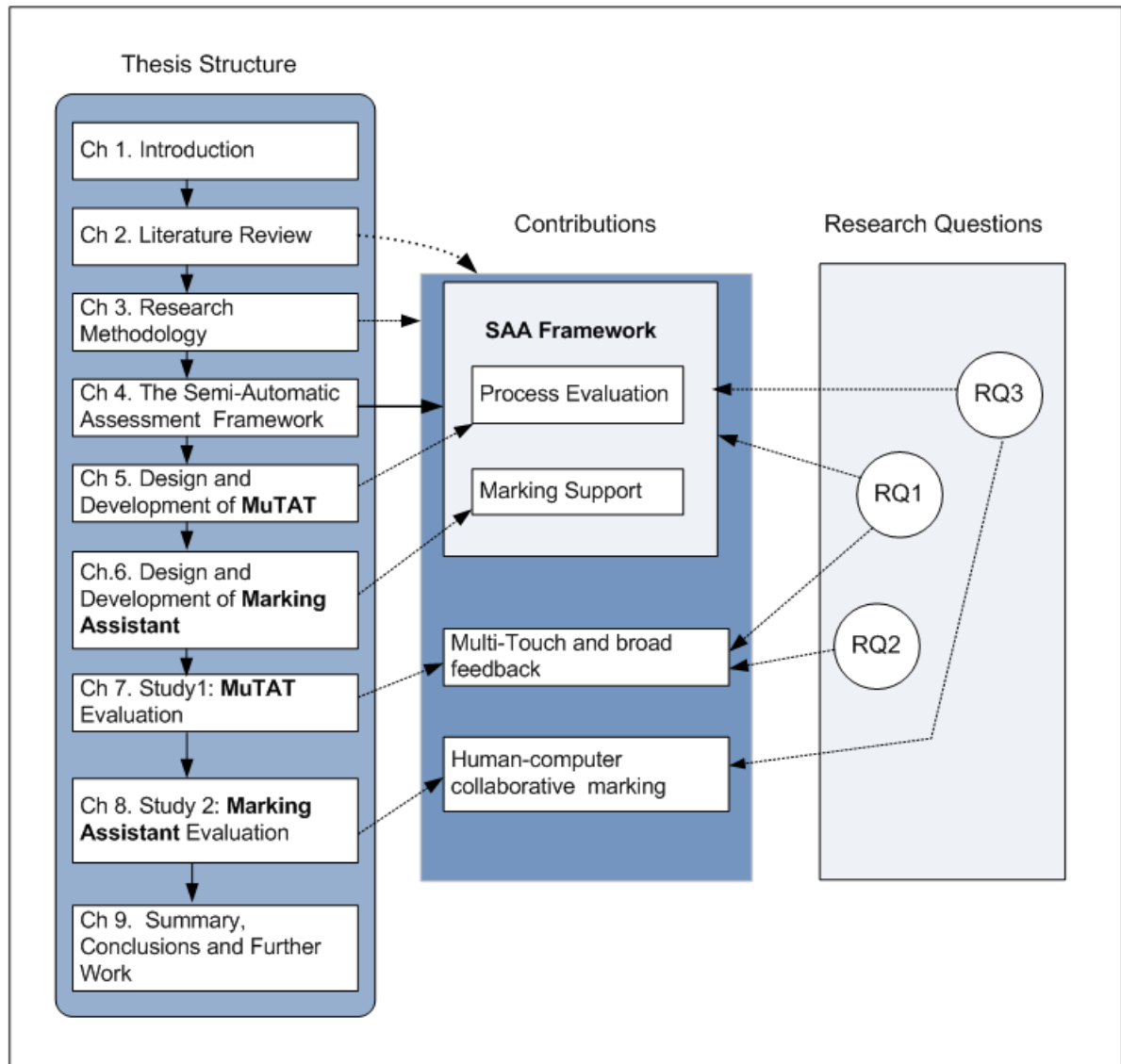
Fig. 1.1 Thesis roadmap

The main body of this thesis is organised as follows:

Chapter 2 situates the study in related literature. It discusses the theoretical underpinning of assessment and feedback practices to guide the research. It includes a critical review of computer-aided assessments, current practices and the role of technology in enhancing

the assessment process. The chapter also provides distinctions between paper-and-pencil and computer-aided assessments. It then looks at types of assessment responses expected from students. This is followed by a review of current practices on marking, grading and feedback production. The chapter concludes with discussions on school mathematics where it examines the use of technology in schools and assessment of conceptual understanding and arithmetic strategies

Chapter 3 discusses the research methodology. It first discusses the range of concepts and theories and methods that could be used in the study to answer the research questions. It then presents the choices made, and the justifications which underlie the methods adopted for the study. It addresses how data will be collected and analysed along with validity and reliability issues. Lastly, it outlines the ethical considerations and approaches, outlining why and how they are adopted and applied in conducting the research.

Chapter 4 describes the semi-automatic framework which forms the base for the remaining chapters. It presents theoretical and operational details of the assessment framework with a focus on school mathematics. The chapter outlines the core aspects of the framework which centre on interactive problem solving, use of analytic scoring rubrics and application of the artificial intelligence method called case-based reasoning for scoring and providing feedback comments.

Chapter 5 explains the design and development of the research prototype Multi-touch arithmetic tool - (MuTAT) with aids the capture of problem-solving steps. The chapter discusses the design and functional goals, the development of the design on a computer user-interface and its architecture. It uses a running example to demonstrate the problem-solving process.

Chapter 6 outlines the design and development of the second prototype tool (Marking Assistant). The Marking Assistant aims to support marking and feedback activities with a view to saving marking time while ensuring marking consistencies across students' responses. The design goals, user-interface and marking process innovations are described.

Chapter 7 presents evaluation studies carried out using the MuTAT prototype. It presents the studies investigating the feasibility of using the techniques in primary schools.

Chapter 8 presents an evaluation study carried out using the Marking Assistant. It presents the results, data analysis and key findings of a study conducted to quantify the savings in marking time by using the tool.

Chapter 9 contains the discussions, recommendations and conclusions drawn from the study. It summarises the contributions and limitations of the study. It also offers suggestions on areas for further research.

1.7 Publications arising from study

Table 1.1 describes the elements of this thesis which have been published in peer-reviewed conference proceedings and journal publication. Furthermore, it shows how the publications map to relevant chapters in the thesis.

Table 1.1 Publications arising from research

Sno.	Publication	Relevant chapter
1	Adesina, A., Stone, R., Batmaz, F., & Jones, I. (2015). A Semi-Automatic Computer-Aided Assessment Approach for Marking and Providing Feedback Comments, *In Proceedings of the 7th International Conference on Computer Supported Education, Lisbon, Portugal.*	4,6,8
2	Adesina, A., Stone, R., Batmaz, F., & Jones, I. (2014).Touch Arithmetic: A Process-Based Computer-Aided Assessment Approach for Capture of Problem Solving Steps in the Context of Elementary Mathematics. *Computers & Education* 78: 333–343 .	2, 4,5, 7
3	Adesina, A., Stone, R., Batmaz, F., & Jones, I. (2013).Use of Multi-touch Gestures for Capturing Solution Steps in Arithmetic Word Problems. *In Proceedings of the 5th International Conference on Computer Supported Education, Achen, Germany, pages* 210-215.	4,5,7

1.8 Contributions

The contributions of this work lie in the following areas.

First, the study will broaden understanding of how humans and computer technology can be used to improve teaching and learning by supporting assessment.

Second, the study describes and successfully tested a semi-automatic framework which allows the capture of responses from primary school students in elementary mathematics using novel interactive touch gestures in a new intermediate-constrained question type which can support the assessment of higher-order skills. The capture technique makes obvious and available important information and elements of the task useful for scoring and provision of feedback. The procedural information aids the detection of arithmetic strategies.

Third the study contributes to providing empirical support although exploratory, of the feasibility of capturing problem-solving process. It shows that students' performances were not affected by the new approach and the approach compared favourably with traditional paper-and-pencil test. This study represents an early step toward developing a methodology and analysis approach that can take advantage of the process data.

The fourth area in which this study contributes is in the novel use application of the artificial intelligence methodology known as case-based reasoning to educational assessments. Students' responses are broken down into hierarchical cases to facilitate their reuse. This method shows how efficiency and savings in marking time may be obtained by reducing repetitive activities.

Chapter 2

Fundamental Concepts and Review of Literature

This chapter presents the literature review for the study. It provides a background which establishes the developments of major concepts and influential studies on assessments and computer-aided assessments. It discusses relevant theories that help in the framing of the questions and phenomena being studied. The chapter also examines methods used in research and practice of using computer technology to improve assessments in education. The chapter starts with a general overview of assessments outlining principles and the different formats. It then looks closely at feedback and the various models as it relates to assessment. Following this, a review of computer-aided assessment and a comparison with paper-and-pencil based assessment is presented. A discussion of the different techniques of marking and grading as it relates to CAA is laid out. Discussions on Artificial Intelligence, Case-Based Reasoning and Human Computer Interactions are also provided in order to set out the research questions. The last section of the chapter examines mathematics problem-solving in elementary schools and the challenges to fully exploiting the potential of computer technology for assessments.

2.1 Assessment

Assessment is an integral component of the teaching process. It provides information that both enhances instruction and promotes learning (Brown et al., 1997). The American Psychological Association, and National Council on Measurement in Education (AERA, 1999) defined assessment as any systematic procedure for collecting information that can

be used to make inferences about the characteristics of people and objects. Messick (1987) described assessment as the process of appropriate interpretation and actions based on an appropriate collection of information about valuable content. These definitions highlight the important features of assessment such as collecting and interpreting information. Assessment therefore is the process of gathering, recording, interpreting, using, and reporting information about a student's progress and achievement in developing knowledge, skills and attitudes. It provides the teacher with information to make decisions about what and how the student is learning. This information, in turn, enables the teacher to identify the next steps in progressing the student's learning and adapt his/her teaching strategies and the learning activities as appropriate. Assessment affects decisions about grades, placement, advancement, instructional needs, curriculum, and, in many cases, funding (Airasian, 1997).

2.1.1 Purpose

Several reasons have been advanced for the purpose of assessment in education. The main purpose of assessment is to establish that all, or most of a class have "learned" what they were taught. Newton (2007) in an exhaustive analysis describes 17 purposes of assessments. Generally, assessment purposes may be grouped into three main categories. These are:

1. *Improvement*: Assessment provides useful and accurate information so that appropriate changes can be made to raise the quality of teaching and learning (Harry and John, 1998).

2. *School Accountability*: Assessment shows whether schools and teachers are making good use of the resources they have been given (Brown et al., 2014).

3. *Student Accountability*: Assessment shows whether students are achieving what is expected of them and judges the quality of their performance (Newton, 2007).

However, Shohamy (2014) argued that assessment is irrelevant,and could be bad and ignorable . He stated that assessment can be fundamentally flawed, inaccurate, and inappropriate. A low-quality assessment results in interpretations and actions that could have significant negative consequences on members of the society. Nevertheless, there is a general agreement among scholars that good quality assessment plays a major role in how students learn, their motivation to learn, and how teachers teach (Black and William, 1998; Conole and Warburton, 2005; Jordan, 2013).

The Assessment Reform Group (ARG) claims that using assessment to improve learning requires the following five elements (Firestone et al., 1999).

- The provision of effective feedback to students

- The adjustment of teaching to take into account the results of the assessments

- The recognition of the important influence assessment has on the motivation and self-esteem of students, both of which are crucial influences on learning

- The need for students to be able to assess themselves and understand how to improve their work

These positions serve as a challenge to thinking and practice about the broad aspects of assessment,

2.1.2   Principles

Principles clearly outlined contribute to achieving goals and purposes. Devising, administering, interpreting and responding to assessments should be based on well-researched principles. Figure 2.1 shows an assessment cycle advanced by Brown et al. (2014).
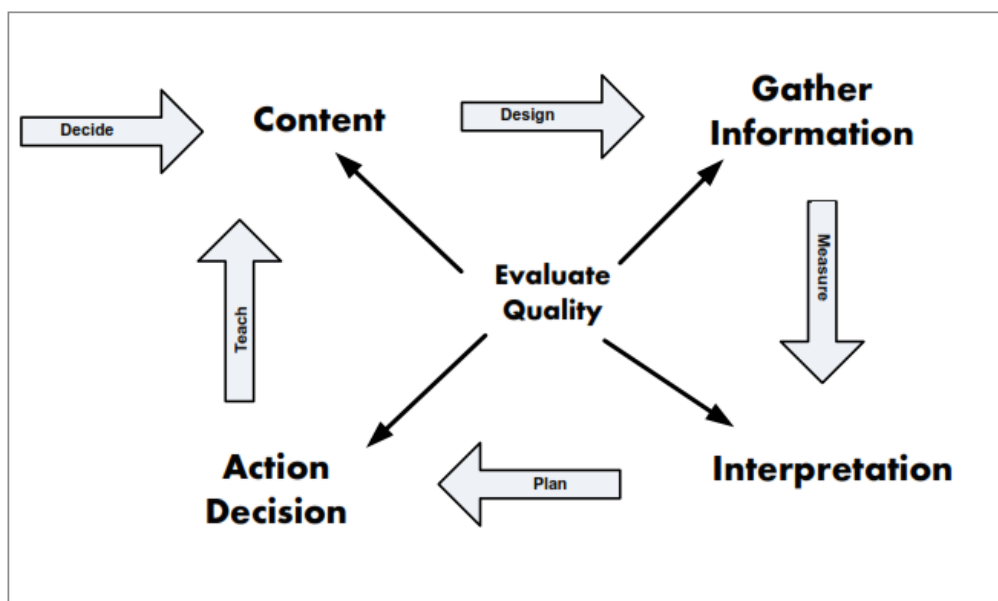
Fig. 2.1 Principles of assessment (Brown et al., 2014)

First, the content to be evaluated is decided in line with curriculum objectives, then different methods for gathering information are designed. After this, based on the information obtained, a measurement of how much or well a student has achieved is made. These measurements

aid teachers and students to interpret the work and make inferences on the quality of the performances. Once interpretation of the work has been done the next step is planning and making appropriate action based decisions to improve teaching. Throughout the process, evaluation is done on how well each decision point has been implemented.

2.1.3   Types of assessment

The main types of assessments (see Bull and McKenna, 2004) can be categorised as:

- Diagnostic – such assessments are usually given at the start of a course to ascertain the student's current knowledge and skills based also called *assessment before learning*.

- Formative – such assessments aim to reinforce and augment students' learning by regular testing and feedback. It is also referred to as *assessment for learning.*

- Summative – such assessments are used to grade students, often by end-of-term examinations (Ainsworth and Viegut, 2006). This is often referred to as *assessment of learning*.

Figure 2.2 illustrates the different perspectives of the major types of assessment in an "*assessment triangle*".
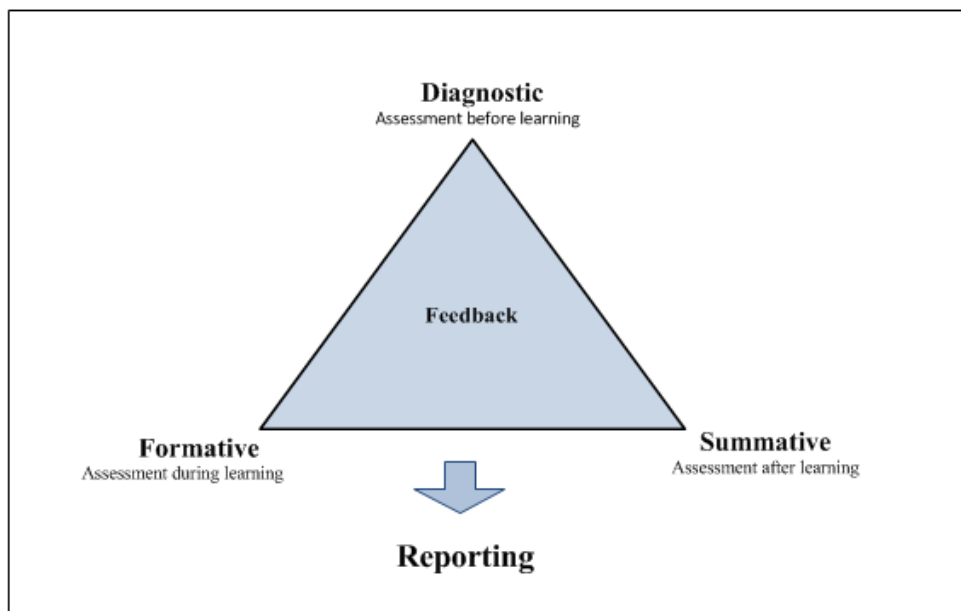
Fig. 2.2 Types of assessments

The figure shows the three most important assessment types with the descriptors. Regarding when the assessment exercise is conducted, diagnostic assessment is sometimes called assessment before learning. While formative and summative assessments are referred to as assessment during learning and assessment after learning respectively (Black and Wiliam, 2009). Feedback is key in all these assessment types. At the end of the exercise, a report detailing the tasks, results and comments can be produced.

2.1.4   Formats and methods of assessment

The literature describes three main methods of assessment. These are described below.

1. **Observational methods**: In observational methods the teacher uses naturally occurring behaviour of students to generate an interpretation of the students learning. Airasian (1997) described this as opinion formation as the teacher engages in on-the-fly interactions with the students. This approach may be time-consuming and sometimes requires levels of training in psychology.

2. **Oral or spoken methods**: Here what the student has to say for himself or herself provides insights into why a behaviour occurred and supplies information from which teachers can generate understandings of what has been learned (Kahneman, 2011). However, the behaviour or speech of another person may aid or detract a student in answering other than their ability or skill.

3. **Written methods:** In this method students provide evidence of what they know by what they write down. This method assumes that the student can read and write. No individual attention is required to observe or interview the student (Hattie and Brown, 2012). Written assessments may be by paper-and-pencil or by the use of computer technology.

According to Brown et al. (2014), written method is the dominant because of the efficiency and privacy benefits it provides. Another advantage is that every student can be potentially assessed in the same standardised way so that it is easier to make comparisons between students. On the other hand written test require quality instructions and questions in order to have valid interpretations and actions (Ward and Bennett, 2012). Also, written assessments may be affected by subjectivity often exercised in evaluating written responses (Brown et al., 2014).

After the consideration of assessment formats, it is helpful to consider the principal shareholders in the process so their needs can be identified and addressed.

2.1.5   Participants in the assessment process

A large number of individuals are involved in the different areas of the assessment process. According to the AERA (1999) these include test developers, test users, test takers, administrators and policy makers. The people involved can be classified into the following four stakeholder groups:

- Student– anyone engaged in learning

- Teacher– anyone engaged directly in facilitating learning: includes teaching assistants, associate lecturers

- Manager – anyone responsible for the organisation or administration of teachers: includes departmental-level and institutional-level management (e.g. managers, administrators, heads of department, Deans, executive officers (CEOs), presidents, principals, Headteachers and their deputies)

- Policymaker – anyone responsible for the setting of policy, whether at a local, regional, state, national, or transnational/intergovernmental level and including funders.

The student is the closest to the learning activity. Hattie's synthesis of over 800 meta-analyses relating to students achievement revealed that teachers have the most singular influence on students'learning and achievement (Hattie, 2008). There is therefore the need to have assessment systems designed to support teachers in assessing their students effectively, efficiently, and fairly.

2.2   Feedback

Learning depends on feedback (Hattie and Timperley, 2007). Whatever the type of assessment, feedback on students' performance represents an indispensable element in the learning process. Several investigators have argued that adequate feedback presented in a timely fashion may improve and accelerate student learning (Bescherer et al., 2011; Fyfe et al., 2012).The benefits of feedback have been described by many writers (Brown et al., 2014; Hattie and Timperley, 2007; Nicol and Macfarlane-Dick, 2006). They agree that feedback makes students aware of gaps that exist between their desired goal and their current knowledge, aids understanding (Sadler, 1989), enables deeper learning during problem solving (Mayer, 1998) and when provided in good time provides an optimal form of guidance (Bloom, 1956). On the other hand, it has been suggested that feedback can be counterproductive if it is

generalised or cheap praise (Walker 1931), grade only marking (Black and William, 1998), grade and unclear comment (Black and William, 1998). Most researchers, however, agree that feedback is beneficial if it is timely, positive, objective and specific. Feedback comes from different sources; peers, teachers and parents (Hattie and Timperley, 2007). This study will focus on teacher feedback to students about their learning.

2.2.1 Principles of effective feedback

Effective marking and feedback would address the seven principles of good feedback practice suggested by Nicol and colleagues (Nicol and Macfarlane-Dick, 2006; Nicol and Milligan, 2006b). According to them the seven principles of good feedback are:

- Helps clarify what good performance is (goals, criteria, standards)
- Facilitates the development of reflection and self-assessment in learning
- Delivers high-quality feedback that helps learners self-correct.
- Encourages interaction and dialogue around learning (peer and teacher-student)
- Encourages positive motivational beliefs and self-esteem
- Provides opportunities to close the gap between current and desired performance
- Provides information to teachers that can be used to help shape their teaching

These seven feedback principles have been recognised as important guides in the study of feedback in students' assessments. Gibbs and Simpson (2004) advocates that students should be provided with challenging learning with clear goals and standards which encourage learning. They further state that sufficient, high-quality feedback received on time will help focus learning and feed-forward to future studies.

2.2.2 Blooms taxonomy

The Bloom's Taxonomy proposed by Bloom (1984) describes a list of cognitive skills that is used to determine the level of thinking students have achieved. The taxonomy ranks the cognitive skills on a continuum from lower-order thinking to higher-order thinking. This is shown in Figure 2.3. This hierarchical taxonomy structure lists six levels of thinking and learning skills that range from basic learning objectives such as knowledge of content through higher-order learning such as synthesis, evaluation, and creativity.
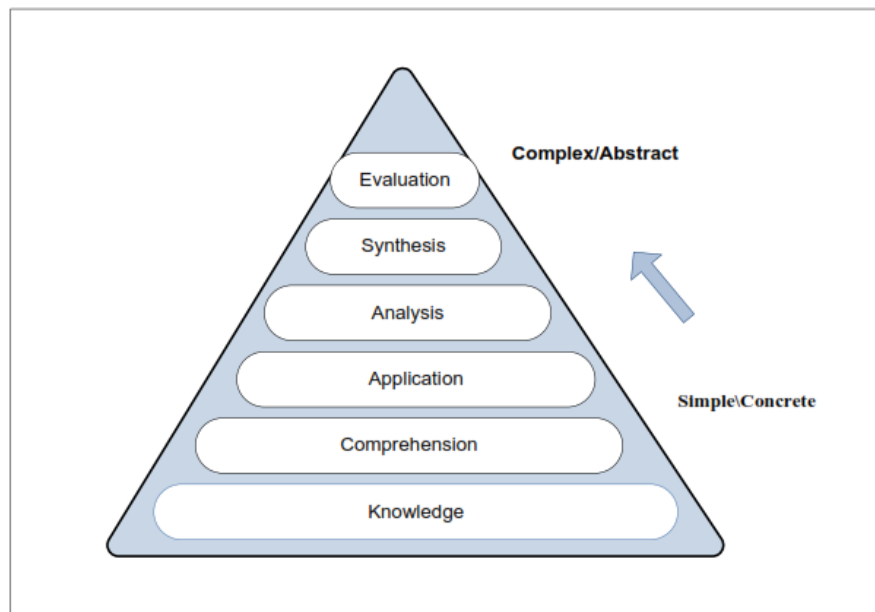
Fig. 2.3 The Blooms' taxonomy (Bloom, 1984)

Anderson and Krathwohl (2001), broke down the Bloom's taxonomy to further components to help provide an understanding of the base of knowledge. His modification identified four major classes of knowledge namely:

- Factual knowledge – Terminology, dates, elements

- Conceptual knowledge – Classifications, categories, principles, generalisations, theories, models and structures

- Procedural knowledge – Subject specific skills, algorithms, techniques, methods, criteria for use of specific techniques

- Metacognitive knowledge – Strategic, contextual, conditional knowledge, self-knowledge

These classes help educators distinguish more closely what they teach and, by implication what they should assess and provide feedback on. By providing the hierarchy of levels, this taxonomy can assist teachers in designing performance tasks, crafting questions for engaging with students, and providing feedback on student work.

2.2.3 Models of feedback

Hattie and Timperley (2007) provided a model for understanding the effectiveness of feedback. Their analysis of empirical studies on different kinds of feedback indicates that feedback may be focused on the *task*, the *processes* needed to do the task, and on *self-regulation* mechanisms. McAlpine (2002) earlier referred to the first two feedback types as product and process type assessments. The three levels of feedback as described by Hattie and brown are depicted in Figure 2.4.
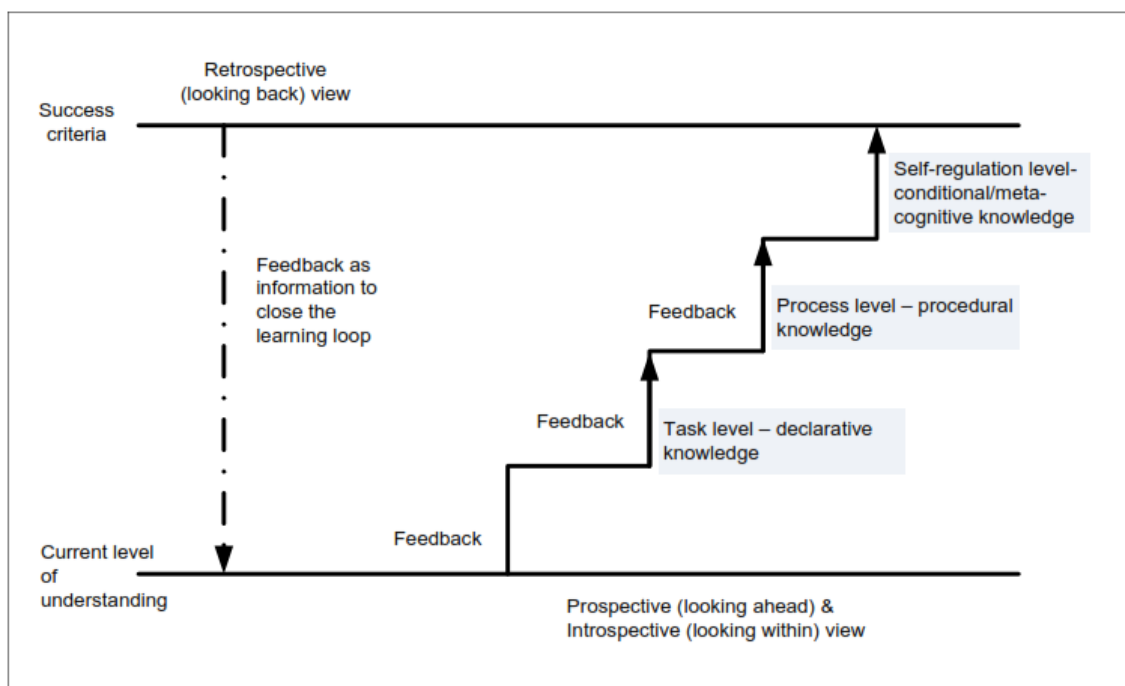
Fig. 2.4 Closing the feedback gap (Hattie and Brown, 2012)

The figure describes the main aim of feedback as closing the gap between the current level of understanding and desired success criteria. Essentially, effective feedback 'provides opportunities to close the gap between current and desired performance'(Nicol and Macfarlane-Dick, 2006, 205). The three feedback levels are steps to achieving the purpose. The main purpose of providing feedback is to give information to the students about where they achieved success in relation to the learning intentions and where they might improve.

I **Task or product level**

   This is the most common type of feedback we see in classrooms. It is more information-focused and aims to help students build their factual or surface knowledge. Examples

include telling a student when an answer is correct or incorrect or asking the student to provide more of or different information (Hattie and Timperley, 2007; McAlpine, 2002).

II **Process level**

This type of feedback is targeted at helping the student improve the process used to create the product. This feedback can help the student develop learning strategies, detect errors, or recognize relationships. When assessing a process, teachers may be interested in examining students' cognitive processes as well (Mory, 2004). Teachers can learn a great deal about a student's thinking by assigning a process task. Studies have shown that a deep understanding of learning involves the construction of meaning and relates more to the relationships, cognitive processes, and transference to other more difficult or untried tasks (Hattie and Timperley, 2007; McAlpine, 2002).

III **Self-regulation**

This is a self-directive process by which students transform their mental abilities into academic skills. It addresses the way students monitor, direct, and regulate actions toward the learning goal. Also, it implies autonomy, self-control, self-direction, and self-discipline. According to Zimmerman et al. (2000) self-regulation involves self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals. This can lead to seeking, accepting, and accommodating feedback information.

Table 2.1 presents questions that may be asked relating to the different levels of feedback. To answer these questions, teachers need varied assessment types to help provide them with evidence which will help to inform their future decisions about their teaching.

Table 2.1 Questions prompts to generate feedback (Hattie and Brown, 2012)

Feedback level	Question prompts
Task(product)-level	
	• Does the answer meet the success criteria?
	• Is this answer correct/incorrect?
	• How can the student elaborate on the answer?
	• What is the correct answer?
	• What other information is needed to meet the criteria?
Process-level	
	• What is wrong and why?
	• What strategies did the student use?
	• What is the explanation for the correct answer?
	• What are the relationships with other parts of the task?
	• What is the student's understanding of the concepts/knowledge related to the task?
Self-regulation level	
	• How can the student monitor his/her own work done?
	• How can the student carry out self-checking?
	• how can the student reflect on his/her own learning?
	• What further doubts does the student have regarding the task?
	• What learning goals was achieved?

According to Airasian (1997) about 90 % of teachers' questions in classrooms are aimed at product level feedback . Relating this to the Bloom's taxonomy, it may be argued that questions eliciting product-level feedback are at the base of the pyramid while process and self-regulation questions are higher up.

2.3   The spectrum of learning theories

Pedagogy is derived from the Greek word meaning "to lead the child" (Elliott, 2008). Its modern usage relates to the art and science of teaching and includes the theories of teaching

and learning (Pachler, 2005). The learning theories relevant to assessments and feedback and the use of technology will be explored in this section. A consideration of sound pedagogical principles guide when and how best to use computer technology.

2.3.1   Behaviourism

Behaviourism is one of the oldest teaching methods. This psychological theory asserts that learning manifests itself in behaviour either changed or reinforced behaviour (Cooper, 1993). It argues that learning is a process of forming associations between stimuli in the environment and corresponding responses of the individual. Reinforcement strengthens responses and increases the likelihood of another occurrence when the stimulus is present again (Elliott, 2008). It is typified by rote learning, drill-and-practice skill acquisition, and a punishment-and-reward system of learning. Current assessment practice, in all sectors, exhibits a behaviouralist approach – rewarding success (with a "pass") and punishing failure (by withholding certification) (Elliott, 2008). ICTs provides students with a Computer-Assisted Instruction that gives instant feedback and control over separate tasks (Pachler, 2005). In the behaviourist model, learning takes place in a highly controlled environment, through drill-and-practice techniques. It manifests itself through changed behaviours such as the acquisition of new practical or mental skills (Elliott, 2008).

2.3.2   Cognitivism

Cognitive learning theories view learning as a process of understanding and internalising facts and concepts about the world (Cooper, 1993). In the cognitivist model, knowledge and understanding are represented by discrete mental states; unique synaptic combinations that represent specific knowledge and understanding (Ertmer and Newby, 1993). According to Pachler (2005), cognitivism takes a data processing approach to learning, with the student being likened to the computer which inputs, processes and outputs information. Cognitivism relies on both teacher and student. The teacher provides content and leads learning (i.e. the creation of specific mental models); the student is responsible for internalising the material presented by the teacher. Cognitivism recognises the individual differences between students, each having his or her own pre-conceived ideas and preferred learning styles. But knowledge remains essentially pre-determined, with the role of the teacher being to facilitate learning through a series of learning activities (Pachler, 2005).

Piaget (1971) was among the first authors to study cognitive development in children. His theory has three components:

1. Schemas are units of understanding or linked mental representations which are stored and applied when needed

2. Adaptation is adjusting to the world by using existing schema to deal with new knowledge, also known as assimilation. Moreover, the process of accommodation happens when the existing schema needs to be changed because it does not work with the new knowledge

3. When the child's schema is able to accept new information through assimilation equilibrium is achieved. When the child's schema is not able to accepts new information disequilibrium occurs. When disequilibrium occurs, the child will adapt the existing schema to include the new information. When this occurs it is known as equilibration i.e. the child restores the balance

In cognitivism, ICTs provide computer as a tutor based on theories of artificial intelligence and theories of information processing (Pachler, 2005). The assumption was that it is possible to design software that emulates the thinking and problem solving of domain experts. The idea was that these "intelligent" technologies can also work as skilled teachers or tutors, providing every student with personal tutors that follow the progress of learning and provide feedback and support when needed.

2.3.3 Constructivism

Constructivism is based on the student being active in the learning process (Ertmer and Newby, 1993). Bruner (1966), considers learning as the process of building or constructing concepts based on current knowledge rather than the memorization of facts. He argued that the purpose of education was to build the child thinking and problem solving skills and that knowledge can be taught to any child at any level of development. His idea lead to the spiral curriculum where complex ideas are introduced in a simple manner and revisited at a later stage of learning (Glasersfeld, 1989).

Bruner (1966) and Vygotsky (1980) in their seminal works argued that the social environment help the child to learn. This lead to what is known as social constructivism where the social environment is believed to play a crucial role in learning (Vygotsky, 1980). For Bruner, this was a concept of scaffolding (achieve a specific goal through a structured environment) whereas for Vygostsky, this was the Zone of Proximal Development (where a person with

more knowledge can help the student achieve the next level of difficulty)(Bruner, 1966; Vygotsky, 1980). The theorists argue that scaffolding and the zone of proximal development can help the student achieve higher level thinking skills quicker. Vygotsky believed that social learning was the precursor to cognitive development. Other theorists such as Gangne and Dewey also contributed to this constructivist learning paradigm (Dewey, 2007; Gagne, 1980).

Learning approaches using contemporary ICTs provide many opportunities for constructivist learning through their student centered environments based on their context. Given, that knowledge is constantly advancing; the design and development principles need to be aligned with teacher and students emerging requirements. Adopting the theory of constructivism with ICTs in modern classes provide a contrast to traditional classrooms where teachers used a linear model and one-way communication, the modern learning is becoming more personalized, student-centric, non-linear and student-directed (Pachler, 2005).

2.3.4   Information Processing Approach and Cognitive Load theories

The information processing approach was developed in the 1960s when the computer first came on stream (Rogers et al., 1999). The computer allowed psychologist to compare the way the human mind processed information to that of the computer. Out of this approach came the principles of selective attention span and attentional capacity. In the selective attention span, humans are viewed to be selective when working on one activity and can ignore other stimulations. In attentional capacity, humans are viewed as being able to attend to a limited number of things at the same time before becoming overloaded.

Sweller (1994) developed the cognitive load theory in problem solving. According to him, to solve a problem a student needs a large amount of cognitive processing capacity. Humans however have limited working memory but an unlimited long-term memory. Because the working memory is limited and can only process a limited number of items at one time schemas are used. Schemas are considered one item even though the schema being recalled may be complex in nature. Information needs to be structured in such a way as to allow for the learner to be able to develop schemas quickly and automatically. These schemas once created, will be stored in the long-term memory and can be recalled when needed. This way the limited working memory is not overloaded.

Sweller (1994) identified three types of cognitive loads. The first is intrinsic cognitive load. This refers to the inherent complexity of the task being performed. It cannot be controlled

because some tasks are more difficult than others. Intrinsic cognitive load is managed by chunking content into smaller problems. The second is the extraneous cognitive load. This occurs when extra unnecessary information is provided with the task to be performed. It actually detracts from the learning process. When intrinsic cognitive load is high, extraneous cognitive load causes problems to the learning process. The third is germane cognitive load. It refers to the mental activity of task that is conducted during the learning process. This leads to the acquisition of schema and is considered the good cognitive load. When a task is high in intrinsic load but low in extraneous load, there will be high germane load (Sweller, 1994). This is when true leaning occurs.

2.3.5 Implications of the learning theories for development of assessment environments

Today, learning is bringing the shifts from linear to multimedia learning, from instruction to construction and discovery, from being largely teacher-centred to being student-centred education, from absorbing material to learning how to navigate and how to learn, from school to lifelong learning, from learning as torture to learning as fun and from the teacher as sole transmitter to the teacher as guide and facilitator. The implementation of current pedagogies is that students will be more active in their learning and assessments. Knowledge is embedded in interactions with the environment as the constructivist theory suggests. Meaningful learning occurs when learners are actively involved and have the opportunity to take control of their own learning. This means that current educational practices should emphasise the active engagement of students in the learning process and the use of teaching strategies. Assessment processes then, should enable students to demonstrate deep understanding of concepts rather than surface knowledge and recall of facts. Assessment should be able to reveal the quality of students' understanding and thinking as well as specific content or processes. Appropriate feedback throughout the learning process should lead students to modify and refine their thinking. ICTs can provide student centred environment where the learning processes can be measured and assessed.

Both the Cognitive Load Theory and Information Processing Approach can be applied in instructional and tool design. The idea is that instructional design should not overload a student's working memory and should allow the student to complete one task before attending to the next. Also learning cannot take place when the working memory and attention span is overloaded. These ideas provide the basis for developing effective learning and assessment environments that benefits students and other stakeholders.

2.4   Computer-Aided Assessment (CAA)

Computer-aided assessment (or computer-assisted assessment) is a term that covers all forms of assessment, whether summative or formative, delivered with the help of computers (Conole and Warburton, 2005). This covers both assessments delivered on the computer, either online or on a local network, and those marked with the aid of computers, such as those using Optical Mark Reading. The terms E-assessment, Computer-Mediated Assessment, online assessment, and Computer Based Assessment have also been used to describe various aspects of the use of computer technology in assessment and are often used interchangeably. In this thesis computer-aided assessment is used generically to cover the whole process of assessment done by the use of computer technology and ICTs. The attributes of CAA are further described in the following sub-sections.

2.4.1   Anatomy and properties of a CAA system

Generally, CAA systems provide a technology driven environment for presenting questions to students to solve and teachers to mark. According to Rasila et al. (2007), a CAA system consists of the following parts:

1. Access and identity control for identifying students and teachers

2. User-interface for the student for taking exercise tasks

3. User-interface for the teacher for marking students responses

4. Database for storage of the assessments

5. A general purpose computation system

6. A grade book where the results are stored

Although not captured in their list, CAA may also to be used to deliver results and feedback to the students. Regarding Step 5, there some general purpose systems that mark and present feedback to students' submitted work items before return. These include Computer Algebra System CAS (e.g. Maxima, Maple and Mathematica) for mathematics (Maple, 2016; Sourceforge.net, 2013) and automated essay scorers (e.g. Intelligent Essay Assessor, IntelliMetric and eRater) for essay grading (Burstein, 2003; Vantage Learning, 2016).

Computer-aided assessment has many benefits which include:

- Availability - students can take the assessments whenever they can assess test resources on a network (Conole and Dyke, 2004).

- Immediate feedback – Students can obtain feedback on their work as soon as a task is completed, it is completely computer assessed.

- Repeated practice – If there is a question bank of similar question that has been authored students can take tests on the same topic repeatedly to develop skills (Conole and Dyke, 2004).

- Anonymity: Some students are more confident in providing answers to machines than humans (Pellegrino and Quellmalz, 2010).

- Potential for self-regulated learning – Computers can support more personalised responses to work and progress and facilitate self-evaluative and self-regulated learning.

- Authenticity – Can present challenging problems and ways to assess complex skills like problem-solving, decision making, testing hypotheses, which are argued to be more authentic to future work experiences(Davies, 2010).

- A wide range of measurement – With the ability to create and visualize complex data sets and models, digital technologies can elicit and measure multi-faceted skills, sets of knowledge and cognitive processes that are difficult to assess using conventional methods (Vendlinksi and Stevens, 2002).

- Increased efficiency and reduction of teachers' workloads – Computers can potentially improve efficiency of data management such as marking, moderating and storing, offering more environmentally friendly administration of assessment (Davies, 2010).

- Improved student performance - Evaluations show that electronic feedback can improve student performance and demonstrates other benefits, such as better student engagement (Whitelock and Watt, 2008; Angus and Watson, 2009).

- Randomised variables in questions - Some questions can be constructed using randomised variables, so each student gets a similar, but not a completely identical problem. This can also be useful if students are using an assessment formatively where they can practice repeatedly, but the detail of the variables used in the questions change each time.

- Improved assessment validity and reliability – Computers can track assessment validity (if the activity is a fair measure of skill and understanding) through the use of rich

media. Improvements in reliability of scoring and robust data sets can be obtained for deeper analysis (Jordan, 2013).

2.4.2 Historical developments of technology in assessments

Computer-Aided Assessment has a history dating back over 50 years. The first known attempt at using computers to assist in assessment was recorded in 1960 when the PLATO (Programmed Logic for Automatic Teaching Operations) project was started at the University of Illinois (Hollingsworth, 1960).. Time-Shared Interactive, Computer-Controlled, Information Television initiated in 1967 was a large-scale project which adopted the use of computers of education (Hayes, 1999). An early attempt at using computers to automate the process of assessing students programming assignments was the "Automatic Grader" (Hollingsworth, 1960).

During the 1970s, the huge growth of multiple-choice testing enhanced attractions of automatic marking in a self-reinforcing cycle. In the 1980s, a large number of educational programs were developed to support learning but with less emphasis on assessment. The 1990s were impacted by the World Wide Web (WWW). After this, assessment systems started to be web-based with systems like Blackboard, QUIZIT, WebCT, ASSYST and PILOT gaining acceptance in many institutions (Al-Smadi and Gütl, 2008).

In the 80s Bunderson et al. (1988) forecasted four generations of computer-aided assessment measurement, namely:

- Generation 1: Computerised testing by administering conventional tests by computer

- Generation 2: Computerised adaptive testing by tailoring the difficulty or contents or an aspect of the timing on the basis of students' responses

- Generation 3: Continuous measurement by using calibrated measures to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory

- Generation 4: Intelligent measurement by producing intelligent scoring, interpretation of individual profiles, and advice to students and teachers by means of knowledge bases and inferencing procedures

This forecast is depicted in Figure 2.5. It shows a progression from efficient testing to personalised learning.
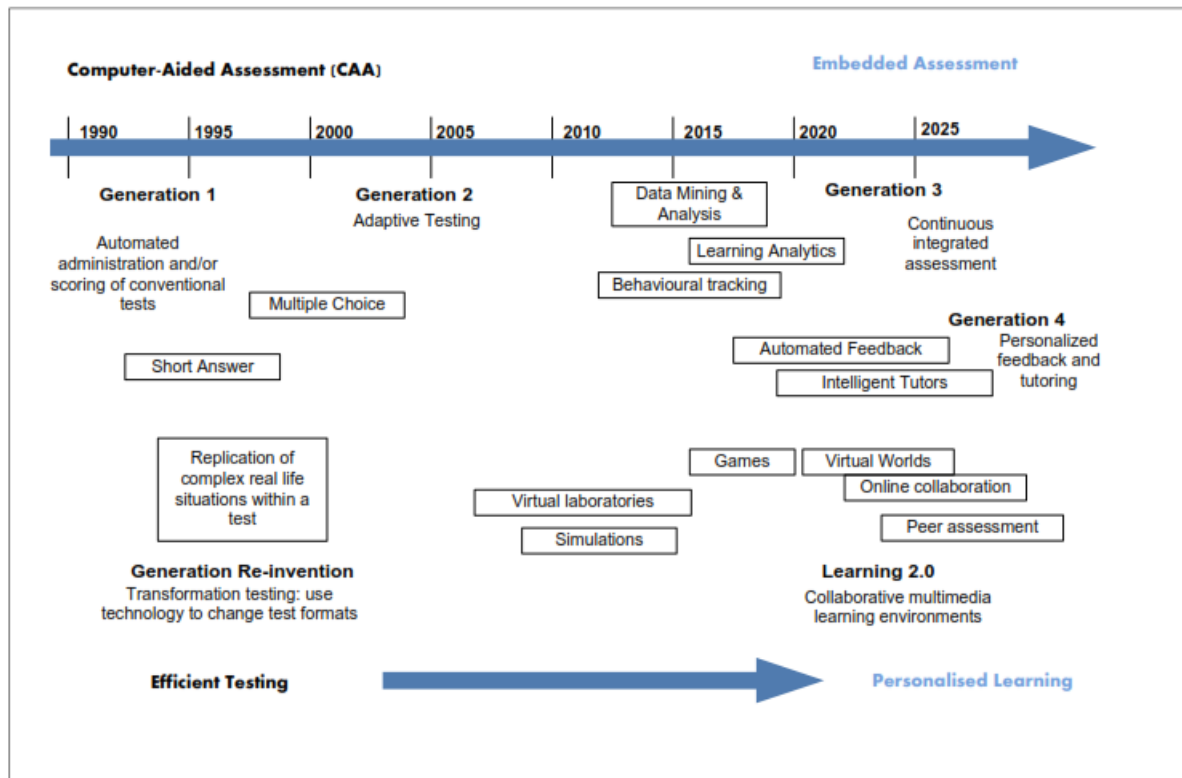
Fig. 2.5 Current and future Computer-Aided Assessment Strategies (Bunderson et al., 1988)

These predictions were not too far off the mark as at 2015. At present CAA can be said to be in the era of data mining and analysis, learning analytics and behavioural tracking. The general direction is towards embedded (unobtrusive) assessment aimed at providing personalised learning and feedback.

2.4.3 From paper-and-pencil assessment to computer-aided assessment

Currently, the world of education is being influenced by emerging technologies. Some studies (e.g. McDonald, 2002; Ripley, 2009) have looked into comparing paper-and-pencil with computer-based mediums.

I **Characteristics of paper-and-pencil assessment**
Ripley (2009) submits that paper questions are confined to a mixture of printed words and diagrams and are inert and inflexible. There is restriction for the teacher in that there is no control over how students read the information on a sheet of paper. Explanations of context have to be accommodated in a limited number of carefully chosen words.

However, responses from students can be very free-form and easily accommodates constructed response question types. Paper-and-pencil task focuses more on retrieving student's answers than on revealing their cognitive processes (Van den Heuvel-Panhuizen et al., 2011).

II **Characteristics of the computer-aided assessment**
ICT environments make it possible to present tasks to students in a way that cannot be done using a conventional paper-and-pencil format (Van den Heuvel-Panhuizen et al., 2011). For instance, a computer can be flexible in the presentation of a question. An animation can communicate what might be inexpressible through text (Ripley, 2009). Also screen dynamics can offer interactions that are impossible on paper – such as rotating a polyhedron, stretching a polygon or doing a hidden calculation on a number. A degree of control can be exercised over the presentation of text setting the time when and the way it should appear. Clements (2000) pointed out that the dynamic and interactive character of ICT environments allows the development of more sophisticated skills such as mathematical problem solving. Studies by Van den Heuvel-Panhuizen et al. (2011) revealed that primary school students can be supported in solving non-routine mathematics problems in an ICT environment with appropriate tasks and a dynamic interactive computer.

These two method of assessment are architecturally different. Figure 2.6 illustrates the major drivers for moving from paper-and-pencil based assessment to computer-aided assessment as described by Scheuermann and Björnsson (2009). The diagram provides a representation of the contrast between these two drivers: the business efficiency gains against the educational transformation gains. It contains four quadrants. The lower-left quadrant represents traditional assessments, typically paper-and- pencil based and which do not exhibit any tendency to develop or innovate. The majority of classroom assessments reflect these characteristics. Moving from the lower-left to the lower right quadrant represents a migratory strategy. Paper-and-pencil based assessments are migrated to a computer-based environment, but the level qualitatively unchanged. The tests are merely replicated on the computer without deep innovation and reflect closely many of the characteristics of traditional paper-and-pencil based tests. At this level, the tasks remains the same, but the tool is substituted. However, business process improvements such as cost reduction, marking reliability and speed of reporting are achieved.
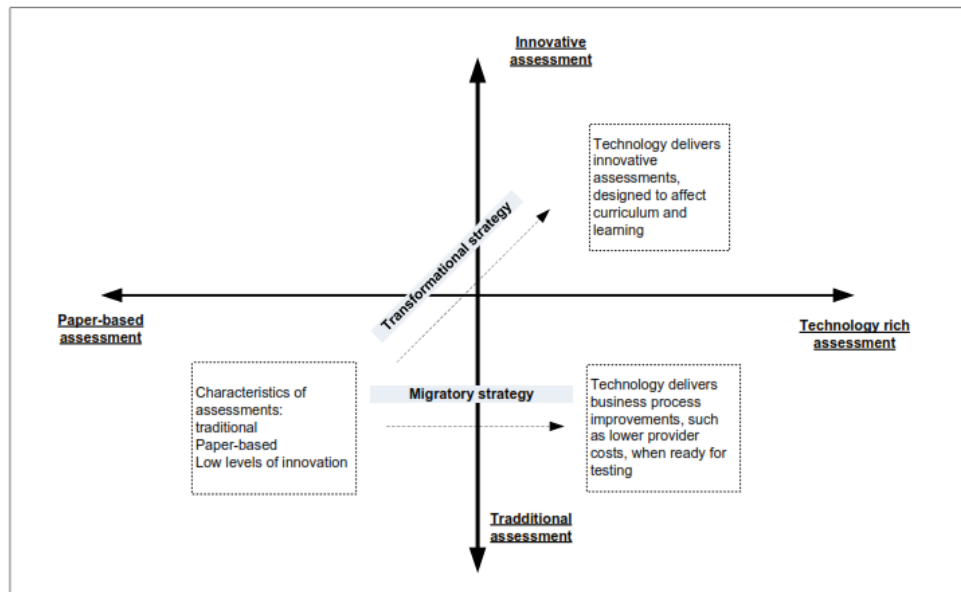
Fig. 2.6 Dimensions of computer-aided assessment innovation (Scheuermann and Björnsson, 2009)

By contrast, the upper-right quadrant represents a transformational strategy underpinning the use of computer technology in assessment. The defining characteristic of innovative assessment is that it is designed to influence (or minimally to reflect) innovation in curriculum design and learning. For example, a computer-aided assessment of problem solving – which uses computers to innovate and redesign the nature of the problem-solving domain and seeks to provide an assessment of skills and abilities not normally assessed through paper-and-pencil based tests. Usually, in this quadrant, the question items and tasks are transformed.

The SAMR model proposed by Puentedura, R (2012) also provides a lens for viewing the migration from traditional assessments to CAA. The model as shown in Figure 2.7 shows a progression of the use of technology from Enhancement (Substitution and Augmentation) to Transformation (Modification and Redefinition).
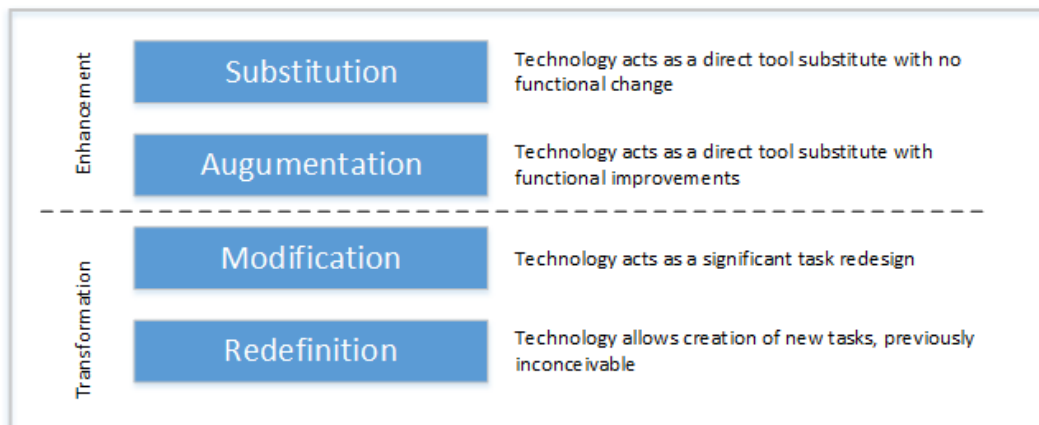
Fig. 2.7 The SAMR Model (Puentedura, R, 2012)

For instance, in a paper-and-pencil assessment task can be changed to CAA task in the following ways (example taken from Puentedura, R (2014)):

- **Substitution**: A Word Processor replaces the paper-and-pencil in the Writing Assignment

- **Augmentation**: A Word Processor and text-to-speech function are used to improve the writing process

- **Modification**: The document created using the Word Processor and text-to-speech function is shared on a blog where feedback can be received and incorporated to help improve the quality of writing

- **Redefinition**: Instead of a written assignment, students convey analytic thought using multimedia tools

The above discussions show that ICTs not only can provide rich enhancements to traditional paper-and-pencil assessments but can transform the whole assessment process.

2.4.4   Partial credit in assessments

One particular feature that distinguishes paper-and-pencil assessments from computer-aided assessment is the allocation of method marks or partial credit (Lawson, 2012). It is generally accepted in mathematics, for instance, that incorrect answers can still demonstrate the achievement of some learning outcome (e.g. the answer is incorrect due to a minor flaw in an algebraic calculation but the student planned the solution steps well) . Such work may be

awarded some of the marks available. Lawson (2012), for example argues that method marks are commonly awarded to questions requiring a multi-step solution process. He states that simple problem such as 'what is the value of -8 + 3?' may not require method marks. When more steps are required to solve problems, method marks would be awarded to a student who demonstrated knowledge of at least some of the steps required to solve the problem even if none of them were carried out correctly. Partial credit may be awarded to a student who demonstrates some knowledge of how to approach a problem and some of the intermediate results.

To incorporate partial credit to CAA, researchers such as Beevers and Paterson (2003) have suggested a design for administered questions. They proposed that the single large problem may be broken down into a series of smaller questions. Students can choose to answer this in stages rather than giving the complete answer initially. In their approach, they prompted students with a series of steps, each of which they must answer in order to determine the solution of the original problem. While this approach gives students opportunities to achieve partial credit, it alters the assessed outcomes when compared with this question delivered during paper-and-pencil tasks. However, they suggested that students derive great benefit from attempting questions and getting immediate feedback on their answers (not just right or wrong but also an indication of how to obtain the correct answer if they were wrong).

gThe assignment of partial credit in a CAA environment is therefore one of the features that needs to be explored in greater detail is the design of CAA assessment systems.

2.5   Types of assessment responses expected from students

Hunt (2012) identifies about thirty different question types available in the online assessment software Moodle. Bennett (1991) described an organisational scheme for questions in various domains of study. This is shown in Figure 2.8.
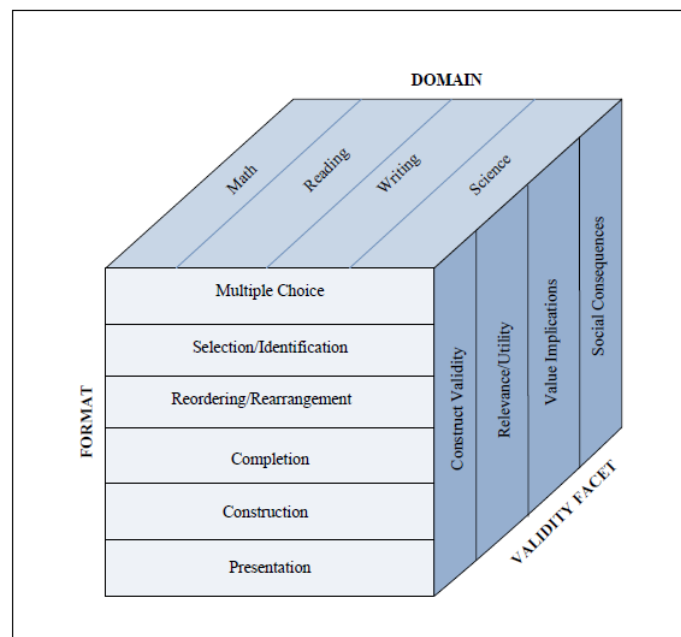
Fig. 2.8 Bennett's many-sided organization scheme (Bennett, 1991)

The figure shows six categories or formats of questions that may be applied to the domains such as mathematics, reading, writing and science. The questions range from simple Multiple Choice to the more advanced question type such as construction and presentation. Bennett also emphasised the importance of considering validity facets for the question types in the respective domains.

Developing on the propositions made by Bennett, Scalise and Gifford (2006) argued that CAA tasks can be categorised based on the degree of the constraint on the options for completing the assessment task. Assessment tasks may be divided broadly into three types, selected response, constructed response and intermediate constraint response types.

2.5.1   Selected response question format

In selected response items, the student must select a response from the options offered. They are often referred to as fully constrained tasks and usually focus on the product or the final outputs. Typical examples would be multiple-choice-questions, true/false or yes/no questions. Selected response items often expect students to recall specific content information. The advantages of this category are: they are easily scored by a variety of electronic means, and specific feedback can be given automatically to the student for any of the options selected. A

disadvantage of this format is the restricted opportunities for students to justify or explain their choice of responses. Even when a student gets the answer there is no way to assess how the student got the answer. Students may guess answers to selected-response questions, so their teacher may not be able to ascertain whether the student understands. Scholars such as Bennett et al. (2008) and Jordan (2013) have argued that selected response question only test the lower level skills in the Bloom's taxonomy.

2.5.2 Constructed response question format

Constructed response type questions are less constrained. They require short answers (text or numeric) or extended text responses, diagram drawing as well as the manipulation of objects (Jordan and Mitchell, 2009). Usually, in this question type students are required to use cognitive skills involving application, evaluation and synthesis, in addition to basic knowledge and comprehension. Constructed responses provide opportunities for learners to justify or explain their reasoning and permit responses to be evaluated according to both the processes used to arrive at a solution and the degree of correctness (Nicol and Milligan, 2006a).

According to Livingston (2009), one of the greatest problems in constructed response testing is the time and expense involved in the scoring as it often includes elaborate systems for monitoring the consistency and accuracy of the score. This is because significant variation may occur in the marking of constructed response of students work, therefore consistency in marking is more complex for teachers (Downing and Haladyna, 2006). Additionally, it requires considerable time to assess by hand and provide constructive feedback to students. In recent years, software are being designed to automatically mark constructed responses using algorithms that emulate the processes used by human experts. One method is to assign scores based on predicted scores a human scorer would assign (Al-Smadi and Gütl, 2008). Another method is to scan for written statements that have been specified as correct answers for other statements having the same meaning. In the field of mathematics engines that score the correctness of algebraic expressions, lines or curves on a graph and geometric figures are being tested (Livingston, 2009).

2.5.3 Intermediate constraint question format

Scalise and Gifford (2006) introduced a taxonomy categorising 28 question types that may be implemented in CAA. The question types are organised along the degree of constraint

on the options for answering or interacting with the assessment task. They described a set of question type termed intermediate constraint items which fall between fully constrained questions (MCQs) and fully constructed responses (e.g. essays). The 28 example types are shown in Table 2.2.

Table 2.2 Intermediate constraint taxonomy for CAA questions (Scalise and Gifford, 2006)

	Fully Selected	Intermediate Constraint Types					Fully Constructed
	1. Multiple Choice	**2. Selection/ Identification**	**3. Reordering/ Rearrangement**	**4. Substitution/ Correction**	**5. Completion**	**6. Construction**	**7. Presentation/ Portfolio**
	1A. True/False	2A. Multiple True/False	3A. Matching	4A. Interlinear	5A. Single Numerical Constructed	6A. Open-Ended Multiple Choice	7A. Project
	1B. Alternative Choice	2B. Yes/No with Explanation	3B. Categorizing	4B. Sore-Finger	5B. Short-Answer & Sentence Completion	6B. Figural Constructed Response	7B. Demonstration, Experiment, Performance
	1C. Conventional or Standard Multiple Choice	2C Multiple Answer	3C Ranking and Sequencing	4C Limited Figural Drawing	5C Cloze-Procedure	6C Concept Map	7C Discussion, Interview
	1D. Multiple Choice with New Media Distractors	2D. Complex Multiple Choice	3D. Assembling Proof	4D. Bug/Fault Correction	5D. Matrix Completion	6D. Essay	7D. Diagnosis, Teaching

Most Constrained → Least Constrained. Less Complex (top) → More Complex (bottom).

The table shows twenty-eight item examples organised into categories based on the task response. The most constrained item types, in column 1 uses fully-selected format. The most constrained item types in column 7 uses a fully-constructed response formats. Between these two types are "*Intermediate constraint items*" arranged in order of decreasing constraint from left to right. Within each column, the order of complexity increases from top to bottom. In the table, it can be seen that the ubiquitous multiple choice question type is a fully selected response type question and is less complex while presentation/portfolio is fully constructed format of assessment. Essay and diagnosis teaching represent the most complex forms of constructed response type format. The concept map assessment format (item 6C in Table 2.2) is an intermediate constraint question type that is fairly complex and close to constructed responses. This question type offers a balanced form of question that can be implemented in CAA for assessment of higher order thinking skills. The concept map question type is briefly explored in the next section.

2.5.4   Concept map intermediate constraint question type

Concept mapping (also known as semantic networking) is a process used to make spatial representations of ideas and the relationships between them (Schau et al., 2001). The concept maps are similar to graphs containing ideas and labelled lines which describe the relationships between them. The purpose of the maps is to help a student show what and how he/she thinks about an idea. While there are different kinds of concept maps, they all help a student to organise and represent his/her thinking. In this way, the maps are graphic organisers or picture summaries of a student's understanding of ideas and the relationships between the ideas.

Herl et al. (1996) have suggested that content understanding can be measured using knowledge maps, which are graphical representations consisting of terms (concepts) and links (interrelationships between concepts). Knowledge maps have been extensively used in primary schools, especially in the teaching of science (Schau et al., 2001). Various research studies on knowledge maps also show them to be effective not only for teaching and learning, but also for assessment purposes (Herl et al., 1996; Ruiz-Primo et al., 1997; Schau et al., 2001). In assessment tasks, students are instructed to construct a map of his or her understanding of how given concepts relate to each other. Students are free to configure their maps any way they choose, and they can add, delete, or move concepts links at will. The rationale for using concept maps in assessments is that they are constructed-response tasks that measure content understanding (Baker and Mayer, 1999; Chung and Baker, 2003; Jonassen, 1999). This idea will be utilised in subsequent chapters for the design of a CAA environment.

2.6   Human Computer Interactions

Advances in ICTs have brought about news ways of interacting with the computer. Human computer interaction incorporates concepts and methods from the fields of Human factors, computer science, and information systems to build interfaces that are easy to use, accessible and efficient (Jacko, 2012). The effectiveness of a computer application is determined by the quality of its interactions and interfaces provided to its users (Hutton, 2013). These have huge implications for the design of assessment environments and tools.

2.6.1 Early computer interactions

The first computer interactions were through toggle switches, blinking lights, paper tape punches, and cathode-ray tubes (Stair and Reynolds, 2012). The need for more efficient input/output devices brought about the use of card readers/punches and teletype printers. Up to 1950, however, interaction with the computer was the domain of information technology specialists. The use of computers in practical systems required more usable, standard human–computer interfaces. The Hollerith card then became a standard medium of input where program data were punched on standardised cards and output, printed on computer paper. Businesses hired computer operation staff to maintain and control access to the computer interfaces. End user computing was not common (Shneiderman, 2010).

As computer use grew in the 1970s, the demand for more direct interaction with applications grew. The move from batch computer architectures to online distributed architectures created the need for new terminal-based interfaces for the application users (Carroll, 2010). Computer terminals were designed to combine a typewriter input interface with a cathode-ray tube output interface. The terminals were connected to the mainframe computer via direct communication lines. HCI interfaces for online applications were based on scrolling lines of text or predefined bit-mapped forms with fields for text or data entry. Standard applications such as word processing, spreadsheets, computer-aided design, presentation and graphics were developed. Text-based command languages were the principal forms of HCI during the 1970s and 1980s for the major operating systems such as UNIX, IBM's MVS and CICS, and DEC's PDP-ll minicomputer (Asada et al., 2009; Mittal and Mittal, 2011). Users were required learn a lot of system commands and formats.

The years following saw an increase in research and development of computer graphical user interfaces (GUIs). These have resulted in today's WIMP (windows, icons, mouse, menus and pointers), HCI standards. The works by Licklider at ARPA, Englebart at Stanford, and the group at Xerox PARC led to the many innovative ideas found in the WIMP interface (Myers, 1998). The first commercial computer systems popularising WIMP features were the Xerox Star, the Apple Lisa, and the Apple Macintosh. The X Window system and the Microsoft Windows versions made the WIMP interface a standard for current computing systems (Stair and Reynolds, 2012). More than any other technology, the WIMP interface and its ease of use brought the personal computer into the home and made computing accessible to everyone.

2.6.2   Current trends in HCI

The advent of the internet in the 1980s brought may changes, opportunities and challenges to HCI (Hevner and Chatterjee, 2010). The World Wide Web which is based on the concept of hypertext, whereby documents are linked to related documents in efficient ways, became revolutionary. Web servers and several web browsers emerged to support the creation and sharing of content.

There are numerous important new directions in the field of HCI as identified by Hevner and Chatterjee (2010) and Poole et al. (1998) which are:

- **Gesture recognition**: The recognition of human gestures began with light pens and touch-sensitive screens, handwriting recording and recognition

- **Three-dimensional graphics**: Research and development of three-dimensional interfaces have been an active area, particularly in CAD-CAM systems

- **Virtual Reality**: Scientific and business applications of virtual reality are just now being explored. Head-mounted displays and data gloves may become commercially viable in the future for marketing demonstrations and virtual design walkthroughs

- **Voice recognition and speech**: Audio interfaces to computer systems have been available for a long time. However, the limited vocabulary and requirements for specific voice pattern recognition remain problems to overcome before widespread use

- **Mobile devices**: HCI and useful interfaces for mobile devices call for a full understanding of the challenges (e.g., battery power, small screen size) and opportunities (e.g., connectivity, computing power) found in handheld and wearable devices

Two HCI areas directly relevant to this study are gesture recognition and mobile devices. Other areas like virtual reality and three-dimensional graphics also offer opportunities. Further discussions of gesture recognition will be provided in Section 4.4.

2.6.3   Mobile devices

Mobile devices are small computing devices small enough to be held with the hands (Poslad, 2009). Typically mobile devices have a display screen, with a small alphanumeric physical keyboard or touch screen which enables a virtual keyboard. Most mobile devices have cell phone capabilities, enabling the user to make phone calls and send text messages. Many "smart" handheld devices have operating systems and equipped with Wi-Fi, Bluetooth, near

field communication (NFC) and Global Positioning System (GPS) capabilities. They are commonly connected to the Internet and other devices. Also, these devices can usually be used to take digital snapshots and play digital video or audio. There is a range of mobile devices designed for different users and target markets. The categories include Personal Digital Assistants (PDA), smartphones and tablet PCs.

Mobile devices offer many advantages for education in comparison to laptops or netbooks. Goundar (2011) and Warschauer (2011) identified six benefits of mobile technologies. First, their lighter weight and orientation flexibility makes them superior for digital reading or accessing of content. Second, their instant-on capability and fast switching among applications allow learning activities to proceed with less delay. Third, their touch screen interface permits a high degree of user interactivity. Fourth, they are much more mobile than laptops, as students can carry them inside or outside a room without having to close and reopen the screen and store them in the carry case and can also use them for mobile data collection or note taking. Fifth, since it is inexpensive to develop apps for mobile platforms, there is a rapidly growing amount of free or low-cost apps for mobile devices, many of which are suitable for education. And finally, mobile devices' long battery life makes them more appropriate for a school day (Warschauer, 2011).

However, mobile devices present computing problems commonly associated with small screens and reduced usability. Reading text due to their relatively small size can be hard. Also, the amount of visible information is limited, and it makes a reader to frequently scroll through a portion of text to read it. This disadvantage makes navigation sometimes tedious and limited (Ceobanu and Boncu, 2014). Some mobile phones tend to be too compact, which results in difficulties in interaction; data entry can be slow and awkward. Despite these limitations mobile devices are an exciting way to interact with technology, especially, when they are touch screen enabled. The intimacy and immediacy of the personal screen and the ease of use and intuitive design of the modern touch screen operating systems considerably reduces user fears and expedites user adoption (Goundar, 2011).

The limitations of current product based CAA systems such as the shortage of tools that capture both the product and processes of problem solving, restricted question types, obscured intermediate answers and the need to take advantage of advances in computer technologies signals the need to have new and innovative methods to capture students responses. These lead to the first research question addressed in this study.

Research Question 1: How can a computer-aided assessment environment be designed to capture students' problem- solving processes?

2.7 Problem solving and school mathematics case study

The revised primary school curriculum introduced into national schools in 1999 consists of 6 curriculum areas that are further divided into 11 subjects (NCCA, 2016). These are:

- English Language

- Mathematics

- Social, Environmental and Scientific Education

- Arts Education

- Physical Education

- Social Personal and Health Education: Social, Personal and Health Education

The aims of the curriculum are to ensure that students are provided with learning opportunities that recognise and celebrate their uniqueness as well as developing their full potential. It recommends focusing on the child as a learner and using a variety of teaching methodologies to facilitate learning. Mathematics has a prominent role in the curriculum as is one of the most popular subjects in schools (Gallup, 2013).

2.7.1 School mathematics

Students who develop proficiency in mathematical problem solving early are better prepared for advanced mathematics and other complex problem-solving tasks (McCloskey, 2007). The study of mathematics is important for two reasons. First, mathematics is integral to many important life skills (Lembke et al., 2012). Each day, people are presented with the need to have basic mathematics proficiency, such as when purchasing goods and services, performing household budgeting, and meeting technical work demands. Secondly, mathematics has become integral to student understanding in other subject areas, such as the sciences, economics, and computer literacy. Without a basic knowledge of mathematics, students may struggle to pass courses and standardised tests, leading to potential academic failure. The National Council of Teachers of Mathematics NCTM (2000) reported that assessment in mathematics should be more than a test to gauge learning at the end of instruction. Rather, assessment should become an integral part of the instruction that guides teachers and enhances students' learning. NCTM recommends that teachers continually gather information about student performance and make appropriate decisions about instruction, content, pacing, review, and enrichment or remediation for students who may be struggling.

2.7.2 Mathematical modelling and problem solving

Research in mathematics education emphasises extensive problem-solving skills and mathematical processes. Gagne (1980) argued that "the central point of education is to teach people to think, to use their rational powers, to become better problem solvers" . Schacter et al. (1999) defined problem-solving as consisting of three facets: content understanding, problem-solving strategies, and self-regulation. Baker et al. (1994) further analysed problem solving into components as suggested by literature – that is content understanding or domain knowledge, domain-specific or domain-interdependent problem-solving strategies and self-regulation. This is illustrated in Figure 2.9. Self-regulation is composed of metacognition (planning and self-checking) and motivation (effort and self-efficacy).
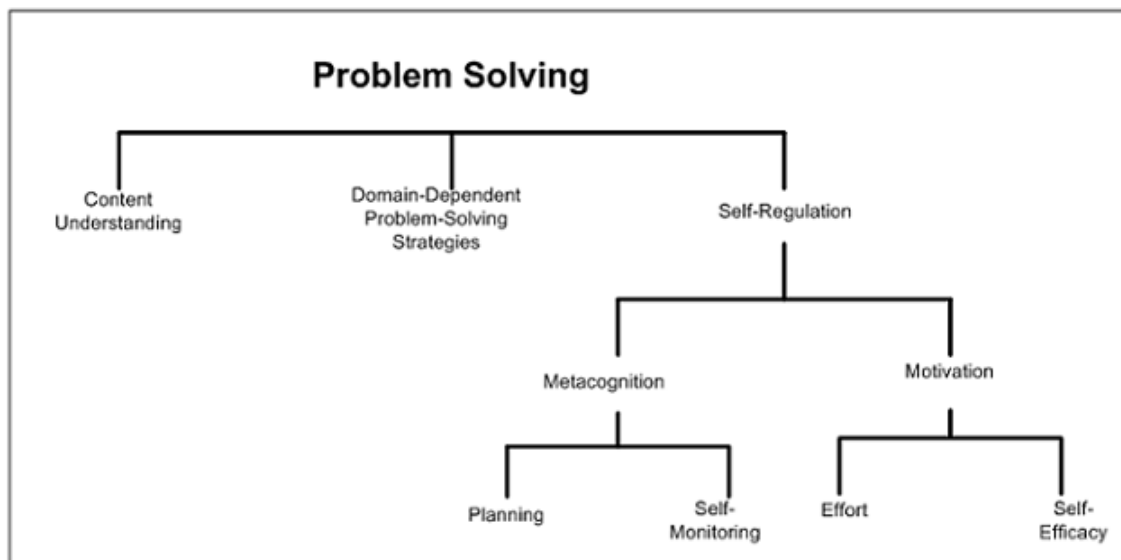
Fig. 2.9 Center for Research on Evaluation, Standards, and Student Testing (CRESST) model of problem-solving (Baker and Mayer, 1999)

In summary, to be a successful problem solver the student must know something (content knowledge), possess intellectual tricks (problem-solving strategies), be able to plan and monitor progress towards solving the problem (metacognition).

Studies in problem solving by Polya (1957), Lesh and Behr (1987) and Schoenfeld (1989) have showed that an important element of problem solving is choosing a model or representation for the problem situation. These researchers and theorists stressed the importance of natural language, concrete models, physical or mental visual images (including pictures, graphs, and diagrams), and symbols in representing mathematical ideas (Bruner, 1966; Lesh

and Behr, 1987; Silver, 1986). Researchers also noted that the symbolic manipulations that students carry out in school are often disconnected from reality and common sense (Baroody and Ginsburg, 1986; Hiebert, 1984; Van Lehn, 1986). As a result, students may produce wrong responses without realising it. Research also showed, however, that if symbolism is closely related to actions and referents that are familiar to young students, then they are able to deal effectively with it (Carpenter, 1981; Hiebert, 1984). The following subsection discusses the different categories of mathematics tasks.

2.7.3 Mathematics tasks

Bennett (2011) identified that in mathematics, tasks requiring complex responses fall into at least four categories:

1. Those calling for equations or expressions, where the problem has a single mathematically correct answer that can take many different surface forms

2. Those calling for one or more instances from a potentially open-ended set that meets given conditions

3. Those requiring the student to show the symbolic work leading to the final answer

4. Those asking for a short text response

The accuracy with which the responses to such tasks can be scored varies dramatically. Also highly variable across categories is the extent to which the computer affords an easy and familiar way for students to respond. The effort needed to create scoring models also differs significantly across these four task classes.

The third task class involves questions that require the student to show the work leading to some final answer (i.e. the student's solution process). Responses to such items will usually show a sequence of steps within one or more correct approaches. Identifying correct approaches, steps, and sequences, and then programming that analysis into a computer can be highly complex and labour-intensive process (Bennett, 2011). Mathematics proficiency requires students to move from a basic level of understanding, often referred to as concrete, toward an abstract understanding of the complex nature that is inherent in mathematics. Three research-based assessment practices, Concrete-Representational-Abstract assessment, error pattern analyses, and mathematics interviews, may be useful for teachers to explore their students' understanding of mathematical concepts (David Allsopp et al., 2007).

As mentioned in Section 2.1, most traditional assessment of problem solving relies on questionnaires, self-reporting, interviews, or naturalistic observations. However, those methods may not adequately capture the essence of the problem-solving process (Chung and Baker, 2003). Computer technology can make capturing problem-solving processes easier and less costly, especially for large-scale testing.

2.7.4 CAA systems of mathematics

Many CAA systems of mathematics have been designed for common use and research. These include:

I **DIAGNOSYS**

DIAGNOSYS is a knowledge based package to investigate mathematics skills (Appleby et al., 1997). In the system, the teacher identifies topic areas relevant to the investigation, and the student then sits a test, delivered and assessed by the computer. The test is adaptive, as not every student gets the same set of questions. As students answer correctly or incorrectly the computer selects a new question from its bank in order to check more fully on the particular skills of that student. Hence, students are not faced with a whole series of questions that they cannot attempt to answer, or a whole series that trivial are to them. The diagnostic report can be made available for either the student or the tutor. Individual or class profiles can be provided.

II **CALM**

CALM is a learning and formative assessment package that employs diagnostic, self-tests and monitoring (Beevers and Paterson, 2003). CALM was one of the first CAA systems to move away from multiple-choice questions and accept mathematical expressions as answers. It also adopted random generation of questions and supported items broken down into steps. Although CALM continues to be influential, it provides a high level of constraint in the question design.

III **STACK**

STACK, a System for Teaching and Assessment using a Computer algebra Kernel, is a computer-aided assessment system for mathematics (Sangwin, 2007). It uses the computer algebra system "Maxima" to generate random questions, establish the properties of answers and to provide mathematical feedback. STACK continues to be widely used and provides a question type for the Moodle quiz.

IV **OpenMark**

OpenMark is an assessment system developed by the UK Open University and has its foundations in computer-assisted learning (Butcher and Jordan, 2010). Open mark supports online delivery of diagnostic and summative assessments. It encourages the use of constructed response type mathematics questions.

V **Maple T.A.**

Maple T.A. is an online testing and assessment system designed for courses involving mathematics (Maple, 2016). It allows instructors to assess student understanding of math- based concepts in science, technology, engineering, and mathematics (STEM) courses. Maple T.A especially supports free response questions that require a numeric answer to fall within given margin of error.

Other products that have been used in the assessment of mathematics work includes Pass-IT Ashton et al. (2006), Mathwise, WebWorK, MathXpert (Sangwin, 2013). While many of these tools are sophisticated and seek to provide automatic scoring and feedback for complex mathematics task, few of them have explored process evaluation (Baker and Mayer, 1999).

2.7.5 Technology in schools

Technology use in schools has been steadily on the rise. In the UK and Europe, about 70% of primary and secondary schools now use tablet computers (Clarke et al., 2013). In the U.S. one-third of students in primary and secondary education have access to school-issued mobile devices. 89% of high school students and 73% of middle schools students use smartphones, with another 66% of middle and high schools students reporting access to laptops (Tomorrow, 2014). Gray et al. (2010) found the ratio of students in the classroom to computers (in or can be brought into the classroom) is 1.6 to 1 at the elementary and 1.7 to 1 at the secondary level. . Given these trends, it is not surprising that computers are fast becoming relevant in assessments and feedback activities.

2.7.6 A word problem case study

Word or story problems are commonly used in schools to train and test understanding of underlying concepts within a descriptive problem as well as to test student's capability to perform arithmetic manipulations (Hegarty et al., 1995). Figure 2.10 shows a two-step arithmetic word problem that involves summing double digits and single digit numbers i.e.

34, 18, and 6 on paper. The question is a two-step arithmetic problem which requires an intermediate result and a final answer. Four student solutions are shown in the diagram
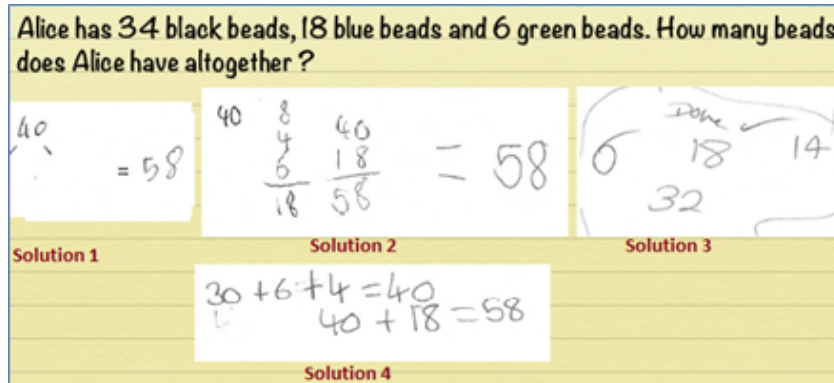
Fig. 2.10 Manual record of solution steps

Solution 1 shows a final answer of 58 without showing explicitly how the answer was arrived at. Solutions 3 and 4 show the workings using vertical and horizontal arrangement for the addition. Solution 3 is less clear; the number 14 is not in the problem text but appears in the solution thereby making the student's thinking less obvious. Too often, the opportunities to provide meaningful feedback are missed when student's responses to these types of assessment tasks are not detailed or clear. To elicit more explicit and detailed responses, better assessment techniques are required.

2.7.7 Conceptual understanding and arithmetic strategies

In many fields of study, students are taught important concepts and correct procedures. Conceptual knowledge has been defined as an explicit or implicit understanding of the principles that govern a domain and the interrelations between them while procedural knowledge is seen as the action sequences for solving problems (Rittle-Johnson and Alibali, 1999). Rittle-Johnson and Alibali (1999) argued that mathematical ability lies in students developing and connecting their knowledge of concepts with procedures. The association between conceptual knowledge and procedural knowledge has been identified in many studies in the mathematical domain for instance counting (Cowan and Renton, 1996), single-digit arithmetic (Baroody and Gannon, 1984; Cowan and Renton, 1996), fraction arithmetic (Byrnes and Wasik, 1991) and proportional reasoning (Dixon and Moore, 1996). Overall, the literature suggests that conceptual understanding plays a major role in strategy adoption and generation. Imbo and Vandierendonck (2008), described three classes of strategies that can be used to solve mental

arithmetic problems. These are direct memory retrieval, procedural strategies such as counting (*e.g.*, $8+5 = 9...10...11...12...13; 9 \times 6 = 9...18...27...36...45...54$) and transformational strategies (e.g., $8+5 = 8+2+3; 9 \times 6 = 10 \times 6 - 6$).

When students are given tasks to solve, it is often valuable to know what strategy was applied to solve the problem. For example, a concept usually taught in maths classes is number bonds. Number bonds refer to useful pairs or numbers that bond to 10 or multiples of 10, they help students see that numbers can be transformed or "broken" into pieces to make computation easier and to recognise relationships. Teachers may want to know if students understand and use this concept in the context of a word or story problem. To accurately assess the strategy so as to give meaningful feedback, they need to have access to the detailed steps and final answer. Performance on such problem-solving tasks is dependent on both strategy selection and strategy efficiency (Mayer and Wittrock, 1996). According to the definitions provided by Imbo and Vandierendonck (2008), strategy selection is the choice of a strategy among several alternatives while strategy efficiency refers to how fast and how accurately a strategy leads to the solution.

Consider the arithmetic elementary mathematics problem outlined above: Alice has 34 black beads, 18 blue beads and 6 green beads. How many beads does Alice have altogether? There are different ways of pairing the numbers. The two-step problem in Figure 2 can be completed in 3 ways (ignoring pairwise commutations, i.e. 34 + 18 + 6, 34 + 6 + 18 and 6 + 18 + 34). The possible groupings are illustrated in Figure 2.11.
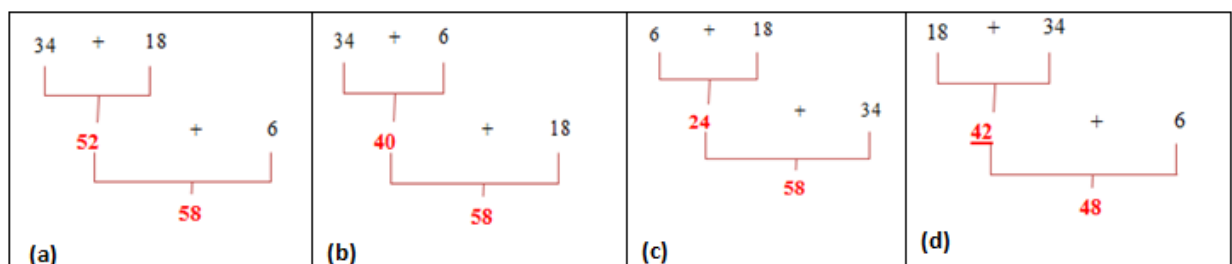
Fig. 2.11 Possible groupings (or cases) of correct solutions

Figure 2.11a shows the pairing of the number from left-to-right in the order the numbers appear in the problem text. Throughout this thesis, the term as-presented strategy will be used to refer to this type of pairing. The solution in Figure 2.11b shows a pairing that reflects a conceptual understanding of number bonds (34 and 6 easily bond to obtain 40). Number bonds refer to useful pairs or numbers that bond to 10 or multiples of 10, they help students

see that numbers can be transformed or "broken" into pieces to make computation easier and to recognise relationships between numbers. This type of grouping will be referred to as place-value strategy. The third grouping, c, will be called random strategy indicating that neither the as-presented nor place-value techniques were applied. Figure 2.11d is the commutative equivalent of Figure 2.11a but has the addition computation done incorrectly. This problem is immediately identified from the clear representation. The approach of reconstructing the solution steps from an audit trail of explicit problem-solving actions provides detail information which can be useful for assessment and feedback processes. The practical application requires a tool that can implement the features described above.

The assessment of processes and strategies is necessary for CAA systems. The first research question in the study investigates the design of CAA for the capture of problem solving steps. The second research question addresses the the need to be able to evaluate the higher-order thinking skills of the Blooms taxonomy using the captured data.

Research Question 2: Will the data captured lead to the detection of strategies and opportunities for detailed feedback on primary mathematics word problems?

2.8 Marking, grading and feedback production

After a student has solved a problem and submitted a response, the next step is marking and feedback production by a teacher or assessor. Teachers are required to assess students' performances based on responses. While essential, marking is time-consuming, potentially error-prone and difficult to complete consistently without bias (Venable et al., 2012). When there are large student numbers, the overall time spent in marking increases dramatically and the time for individual feedback decreases. To address these challenges marking and feedback support tools are employed (Venable et al., 2012).

The purpose of a marking and feedback support tool is to improve the efficiency and effectiveness of marking and feedback activities (Berger and Dreher, 2011; Burrows and Shortis, 2011).

1. Improved efficiency of marking and feedback: Reduced marking time – Reduction of the effort in marking and feedback is important to teachers so that they can reduce their workload or devote effort to other teaching activities. Also, it may reduce stress levels in teachers

2. Improved effectiveness of marking and feedback: Improved accuracy and consistency of marks – Correct and fair marks are important to teaching staff, students, and administrators. More 'objective' marks that more correctly assess work are important to all concerned.

Marking can be done automatically, semi-automatically and manually.

2.8.1 Manual (online) assessment

As mentioned in the previous section hand-scoring students responses can be time-consuming and takes a substantial amount of teachers time as it relies completely on human effort. However, an advantage of this mode of scoring is that it can produce a high level of detailed feedback and personalization (Denton, 2003). A teacher who is familiar with the strengths and weaknesses of a particular student will be able to tailor his or her feedback to meet the needs of the student by supplying individualised comments and advice. However, manual assessments are subject to human factors such as boredom and fatigue of the teacher which often leads to inconsistencies.

In manually marked CAA systems, the design is similar to the traditional paper-and-pencil tests. The computational tool is used to capture responses from students while human assessors assign marks and provide feedback (see, for example, (Bennett et al., 2008). Studies have shown that manual marking can be frustrating (Bailey and Garner, 2010), stressful (Hogan et al., 2002), time-consuming (Ferns, 2011) and subject to inconsistencies (Brown et al., 2004).

The challenges with manual marking done with paper-and-pencil or computer are summarised below:

- Teachers have difficulty being consistent with assessment—namely, awarding grades based on performance measured against assessment criteria. It can be difficult for a teacher to consistently calculate marks when multiple weighted assessment criteria need to be factored into determining a final grade.

- It is necessary to moderate grades that have been generated by a variety of assessors because differences some occur due to the subjective evaluations by different assessors.

- It takes a large proportion of the teacher's time to provide feedback and assessment, which often may be simply misunderstood or, at worst is ignored or never collected.

- Providing feedback can be repetitive in nature. For example, the same comment is often applicable to many students who have made the same error.

- Time constraints may mean that feedback is not returned to students quickly enough for it to 'feed-forward' into the next learning task.

- Administrative errors can occur when recording grades/marks.

2.8.2 Automatic assessment

A large number of studies have reported on implementing automatic assessments (see Conole and Warburton, 2005; Jordan, 2013). The most obvious motivation for fully automatic systems is the huge savings that could be realised in assessment costs. Automatic assessment also contributes to the objectivity and reliability of the grading system because the system is not affected by boredom and fatigue. However, as mentioned in Chapter 2, several of these studies do not consider the problem-solving process. Grading MCQs and product based assessment tasks are relatively easy as only the final product are compared with the standard or model answer. Grading and providing feedback on the problem-solving process is more challenging. Bescherer et al. (2011) argued that a fully automatic system requires the complete modelling of all students, considering all possible solution paths, errors and misconceptions they might have. Because of this limitation, fully automatic systems have limited opportunities for detailed and personalised feedback (Rebholz and Zimmermann, 2013). Others have argued that a computer does not have the same degree of background knowledge or level of inferential capability as a human marker.

2.8.3 Semi-automatic assessment

Semi-Automatic assessment aims to combine manual and computer efforts in the assessment process. It seeks to support the teacher in marking by providing automatic support where possible. It aims at providing the benefits of fully automatic CAA with the advantages of detailed assessment and feedback by human assessors. Several studies have explored this assessment approach. The principle behind semi-automatic systems is the reduction of repetitive tasks which overwhelms and bores human assessors. The computer is used to group identical components in students' responses and the human assessor only has to make one judgment (with feedback) on the whole group of identical components. This is repeated over the whole assignment yields a huge speed increase for the markers. In recent

times semi-automatic techniques are increasingly being used in different CAA research areas including computer programming (Ahoniemi and Reinikainen, 2006), mathematical induction (Rebholz and Zimmermann, 2013), Entity-Relationship (ER) diagrams (Batmaz et al., 2009; Tselonis et al., 2005) and free texts (Kakkonen and Sutinen, 2009; Sargent, J et al., 2004). These investigations have applied different frameworks to combine human and computer assessment.

Sargent, J et al. (2004) use the term Human-Computer Collaborative Assessment (HCCA) for free text constructed response type problems. In this approach, the marking was viewed as a process of active collaboration between the human marker and the computer. The computer does the routine work of sorting student responses into different groups while humans make the important judgments on marks. This approach was used by Tselonis et al. (2005) for diagram marking problems. However, the HCCA approach did not provide much intelligence and personalised feedback in the design. Bescherer et al. (2011) developed the model called Intelligent Assessment. This model combines computer-based assessment with human assessment. First, computer software automatically assesses problems in order to detect standard problems and mistakes as well as standard solutions. These set of standard mistakes and solutions are defined when the program is written. Based on these standard solutions and mistakes, the computer is used to provide feedback to the user. Unclassified mistakes, non-standard solutions or solutions which cannot be assessed by the software are automatically forwarded to a human teacher. The teacher then assesses the solution manually and gives the feedback needed by the student. As a consequence, tutors do not need to concentrate on all possible solutions and can spend more time providing feedback on unusual solutions or weak students. The teacher can give more individual and more intensive assessment and formative feedback.

Batmaz et al. in their submission, improved on the HCCA model by developing a framework that used partial marking to mark solution groups manually and the Case-Based Reasoning (CBR) method to mark some of the groups automatically. These studies suggest that semi-automatic assessment approach has the potential to significantly reduce the time and effort required to mark and provide feedback comments to student work for assessment tasks that require constructed responses.

In summary, these studies demonstrated the potential of the approach; however the studies have not clearly focused on supporting broad feedback and improving the efficiency of human assessor. Table 2.3 summarises the features as outlined by Kakkonen and Sutinen (2009), Sargent, J et al. (2004) and Bescherer et al. (2011) of the three major ways marking and feedback can be done.

Table 2.3 Comparison of manual, semi-automatic and fully automatic systems

Assessment	Manual	Semi-automatic	Automatic
Pedagogical background	Instructionist/Behaviorist	Constructivist/Instructionist	Behaviourists
Implementation	A teacher	A system and possibly pre-graded responses or other material and a teacher	System and pre-graded responses or other material
Support for learning	Some support	High support	Little support
On-line/ off-line	Usually offline	Online	Usually online
Speed	Low	High	Very high
Feedback	Grade and optionally some verbal comments	Grade, verbal and graphical comments	Mostly grade only
Personalization	Very high	High	Very low
Affected by errors caused by lack of consistency, fatigue or bias	Very high	Low	Very low

The next step is the consideration of techniques that may be implemented in computer-aided assessment systems.

2.9   Artificial Intelligence

Artificial Intelligence is concerned with making computers behave like humans (Russell et al., 2009). The possibilities for AI to make significant contributions in any field are tremendous (Stair and Reynolds, 2012). The educational domain should not be left behind. One area artificial intelligence can make significant contributions is in the field of marking and feedback. This section explores the AI concepts and identifies approaches which can aid assessment processes.

2.9.1 Artificial Intelligence in Education

Advances in the area of artificial intelligence (AI) have long been seen as having positive impact on educational applications (Jones, 1985). In artificial intelligence the computer system is viewed as having the ability to take on the characteristics of human intelligence (Russell et al., 2009). Stair and Reynolds (2012) described an intelligent behaviour as one having the characteristics of learning from experience, applying the knowledge to new experiences, handling complex situations with few pieces of information, thinking in a logical and rational way, understanding visual information and processing symbols. The broad field of AI includes robotics, vision systems, natural language processing, learning systems, neural networks and expert systems (Russell et al., 2009). In robotics, mechanical or computer devices are used to perform tasks that require a high degree of precision or are tedious or hazardous (Asada et al., 2009). Vision systems include hardware and software that enable computers to capture, store and manipulate images and pictures such as in face-recognition software (Szeliski, 2010). Natural language processing enables the computer system to understand and react to statements and commands made in a 'natural language' such as English. Learning systems allow a computer to change how it functions or reacts to a situation based on feedback received. Expert systems stores knowledge and makes inferences similar to those of a human expert. It consists of components such as a knowledge base, and inference engine, and explanation facility, a knowledge acquisition facility, and a user interface (Jackson, 1998). Other artificial intelligence applications are genetic algorithm and intelligent agent (Stair and Reynolds, 2012).

2.9.2 Authentic assessment and performance assessment

According to the analysis by Salem (2015) , research in the field of AI in Education consists of seven main areas, namely: Intelligent Educational Systems (IES), Teaching Aspects, Learning Aspects, Cognitive Science, Knowledge Structure, Intelligent Tools, Shells and Interfaces. Perhaps the most popular application of AI in education is in the development of intelligent tutoring systems. Intelligent tutoring systems have been used to support students during learning and problem-solving. They have also been utilised to prescribe learning activities at the level of difficulty and with the content most appropriate for the student (Azevedo and Hadwin, 2005; Shute, 2008). The two main components in developing efficient and intelligent tutoring and learning systems in any domain are the "knowledge base" and the "inference engine" (Salem, 2015). Relating to the knowledge base, there are many knowledge representation and management techniques. These includes lists, trees, semantic networks,

frames, scripts, production rules, cases, and ontologies. The key to the success of such systems is the selection of the appropriate technique that best fits the domain knowledge and the problem to be solved, (Mazza and Milani, 2005). Concerning the inference engine, there are many methodologies and approaches of reasoning, e.g. automated reasoning, case-based reasoning, commonsense reasoning, fuzzy reasoning, geometric reasoning, nonmonotonic reasoning, model-based reasoning, probabilistic reasoning, causal reasoning, qualitative reasoning, spatial reasoning and temporal reasoning. These methodologies are receiving increasing attention within the AI in education community (Salem, 2008, 2011).

2.9.3 Reasoning systems

CBR is an alternative to rule-based reasoning. A rule-based reasoner has a large library of rules of the form, "IF A THEN B." These are chained together in various combinations to solve problems. A case-based reasoner solves problems from a library of cases.

Case-based reasoning offers two main advantages over rule-based reasoning; First expertise is more like a library of past experience than a set of rules; for this reason it better supports knowledge transfer (communication of expertise from domain experts to system). Second, many real-world domains are complex, and it is difficult and impractical to specify fully all the rules involved. Cases on the other hand that provide solutions for similar problems can always be given.

2.10 Case-Based Reasoning (CBR)

Case-based reasoning CBR is an artificial intelligence methodology that capitalises on past experiences to solve current problems (Watson, 1999). According to Kolodner (1993) CBR may be viewed simultaneously as a research paradigm, as a perspective on human cognition and as a methodology for building intelligent systems (Shiu and Pal, 2004). CBR is based on a model of human cognition dealing with knowledge in form of concrete experienced examples. It came out of research in the area of Cognitive Science by Schank and Abelson on dynamic memory (Schank, 1999). During research into the human ability to solve problems, they realised that, in general, people develop solutions based on previous experience(s) with similar situations and that most people discuss problems and solutions regarding past experiences. Thus, it appears obvious that, complete solutions derived solely from first principles are fairly rare. Instead, most problem solvers approach new problems and their

associated solution(s) by relating both the problem and the solution to previous experiences. Hence, new solutions are built from information gained from previous experiences, coupled with some reasoning from first principles.

The core assumption of CBR is similar problems have similar solutions (Shiu and Pal, 2004). This assumption's basic idea is to solve a current problem by reusing solutions that have been applied to similar problems in the past. Therefore, the current problem has to be compared to problems described in cases. Solutions contained in cases that represent very similar problems are then considered to be candidates for solving the current problem, too.

2.10.1 Motivations for CBR

CBR improves problem solving efficiency because it stores already generated solutions to reapply them for problems occurring in the future compared to starting from scratch (Riesbeck and Schank, 2013). CBR enables handling of incomplete domain knowledge. Because many areas are poorly understood, and domain theory is not available, it is often easier to make decisions based on similar situations in the past rather than on generalised knowledge. CBR also simplifies the knowledge acquisition. Formalised knowledge representation requires rules as in traditional AI approaches This often makes the acquisition process sophisticated and time-consuming CBR, reduces the knowledge acquisition effort (Pal et al., 2012).

2.10.2 The Case-Based Reasoning Cycle

The process of CBR is commonly described using the CBR cycle proposed by Aamodt and Plaza (1994). The circle is shown in Figure 2.12. It comprises of four steps often referred to as the four "Re"s: **Retrieve:** The case or cases most similar to the query is retrieved from the case base. Similarity among cases depends on the domain and how cases are represented. To realise this retrieval task, CBR systems employ special similarity measures that allow the computation of the similarity between two problem descriptions. Because the interpretation of this similarity strongly depends on the particular domain, similarity measures are part of the general knowledge of the system.

Reuse: the information/experience stored in the solution descriptions of the retrieved case(s) to solve the presented problem. Solutions may in the simplest case be reused directly, and in more complicated scenarios be adapted to meet criteria specific to the new problem

Revise: the retrieved solution if it is necessary is revised to solve the presented problem in a satisfying way. Depending on the employed adaptation procedure, the correctness of the suggested solution often cannot be guaranteed immediately. Then it becomes necessary to revise the solved case. Usually, the focus of the revise phase lays on the detection of errors or inconsistencies in the suggested solution and the initiation of further problem-solving attempt

Retain: The query case with the newly created solution is stored with the rest of the cases in the case base. This way new experiences are kept with the old knowledge, and can contribute to future problem solving. The task of the CBR cycle's last step is to retain this new case knowledge for future usage. Therefore, the new case may be added to the case base.
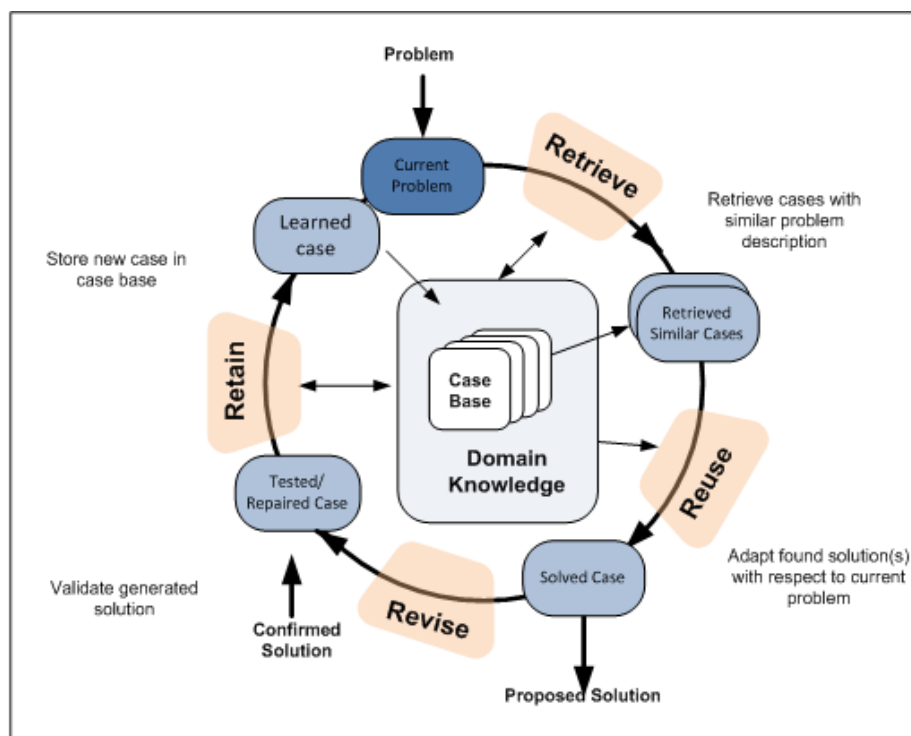
Fig. 2.12 The CBR Cycle (Aamodt and Plaza, 1994)

Earlier discussions addressed the need to support teachers in marking and feedback work. While manual and fully automatic systems have provided helpful solutions they come with identified limitations. The semi-automatic approach provides a way to exploit the strengths of both methods. The case-based reasoning artificial intelligence methodology provides opportunities for supporting teachers to achieve efficiency and consistency goals. This leads to the third research question addressed in the study.

Research Question 3: How can assessors score and provide feedback on students' responses consistently and efficiently?

2.11   Chapter summary

This chapter has reviewed the current state-of-the-art in related work contributing to assessments and computer-aided assessments in particular. The literature covered in this chapter sets a useful background for the thesis as a whole. The chapter explored the principles of feedback and showed that process level feedback is important for students' learning. It also explored the different types of responses that are produced by students. The constructed response type tasks provide more opportunities for process level feedback and assignment of partial credit where necessary. However, the question format involves significant time expense and often involves elaborate systems for monitoring consistency and accuracy of scores. Discussion on mathematical modelling and problem solving in elementary mathematics were also outlined with the important requirement of providing detailed feedback.

The current study, therefore, was guided by the following research questions:

1. **Research Question 1**: How can a CAA assessment environment be designed to capture student's problem-solving processes?

2. **Research Question 2**: Will the data captured lead to the detection of strategies and opportunities for detailed feedback on primary mathematics word problems?

3. **Research Question 3**: How can assessors score and provide feedback on students' responses consistently and efficiently?

Chapter 3

Research Methodology

Research is a well-coordinated activity aiming to contribute more knowledge to the existing body of knowledge (Bryman, 2012). The research methodology in a study describes the strategy of inquiry used to answer the research questions in the study. It includes a consideration of the concepts and philosophies which underlie the methods adopted in the study (Vaishnavi and Kuechler, 2015). The research methodology helps to explain how the study was conducted, and a consideration of different choices enhances the capacity of developing quality solutions (Vaishnavi and Kuechler, 2015). This chapter identifies the various approaches, methods, and techniques that could be adopted in the study to answer the research questions. It examines the various choices that may be made in the quest to finding an appropriate philosophical stance, research approach, method and strategy to meet the research aim and objectives as set out in Chapter 1. Additionally, it briefly looks into the advantages and disadvantages of the various methods discussed in order to clarify the path chosen for the study. The chapter also discusses validity and reliability issues taken into consideration in the study. Finally, it outlines ethical elements considered during the conduct of the research.

3.1 The research onion

Research studies are structured to address a problem by working out what data are needed and how to obtain these. The selection of technique(s) used to obtain data along with the procedures to analyse the data contributes to the overall research process. Two major frameworks commonly used to provide perspectives for the research methodology are the

nested method and the research onion (Bryman, 2012; Saunders et al., 2012; UK, Essays, 2015). These two models contain the steps required for effective research; however, the research onion contains more information compared to the nested model. The research onion, as advanced by Saunders et al. (2012), provides a detailed means of describing the stages through which researchers typically pass through when formulating an effective methodology. In the model, Saunders et al. (2012) used the metaphor of 'Research onion' to illustrate the paths and decisions that may be taken to execute a study successfully. Figure 3.1 shows the different layers of the research onion. The outer layers are considered first and represent the research philosophies and research approaches that may be adopted. The middle layers help to shape the research design; they include methodological choice, research strategy and time horizon. The preceding layers lead to the core layer which involves data collection and analysis.
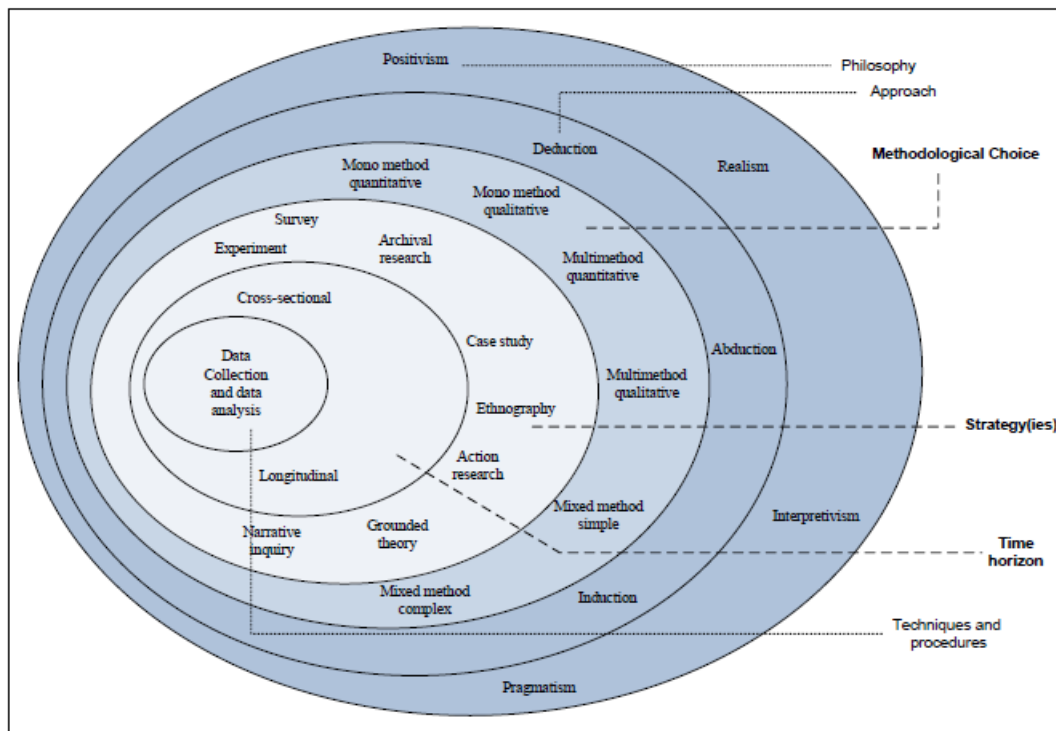
Fig. 3.1 Research onion. (Saunders et al., 2012)

The research onion model has been adopted for the present study because of the broad perspectives it provides as a descriptive model. The series of stages by which different methods of data collection can be understood enables the illustration of the steps by which a study can be described. The following sections describe the various stages in the research process and provide rationales for the choices made in the conduct of the present study.

3.2   Research philosophy

The research philosophy creates the starting point for an appropriate research approach. A research philosophy refers to the set of beliefs concerning the nature of the reality being investigated (Bryman, 2012). Crotty (1998) argues that the research philosophy adopted is based on the researcher's assumptions about the way he or she views the world and affects the way in which data about a phenomenon is gathered, analysed and used. The research philosophy used can help explain the assumptions inherent in the research process and how this fits the methodology adopted.

3.2.1   Epistemological perspectives

Epistemology is a branch of philosophy that relates to the nature and scope of knowledge that constitutes acceptable knowledge in a particular field of study (Bryman and Bell, 2012). According to research onion model (Figure 3.1), four main positions may be adopted namely, positivism, realism, interpretivism and pragmatism. These are briefly described.

I **Positivism**

Positivism is the epistemological position that acceptable knowledge consists of observable phenomena that can provide credible data and facts. It can focus on causality and law-like generalisations, thereby reducing phenomena to simple elements (Myers and Avison, 2002). Also, this view uses mathematical and logical means to generate and test hypotheses. Experiments (field and laboratory), case studies and surveys may be utilised for gathering data (Galliers, 1992; Mingers, 2003). As a philosophy positivism adheres to the epistemological view that only "factual" knowledge gained through observation (the senses), including measurement, is trustworthy. The role of the researcher is restricted to data collection and interpretation through the objective approach and the research findings are usually observable, quantifiable and lend themselves to statistical analysis(Saunders et al., 2012).

II **Realism**

Positivism is similar to positivism in its processes and argues that observable phenomena provide credible data and facts (Saunders et al., 2012). It holds that the researcher is independent of the study and so will not create biased results. However, where they differ is that realism thinks that scientific methods are not perfect. It believes that all theory can be revised and that the ability to know for certain what reality is may not exist without continually researching and leaving the mind open to using new methods of research

(Crotty, 1998; Vaishnavi and Kuechler, 2015). Forms of realism have been suggested, for instance, where there is insufficient data, it means inaccuracies in sensations may exist this is known as direct realism. Alternatively, when a phenomena create sensations which are open to misinterpretation, it is referred to as critical realism (Saunders et al., 2012).

III **Pragmatism**

Here either or both observable phenomena and subjective meanings can provide acceptable knowledge dependent upon the research question. The focus is on practical and applied research integrating different perspectives to help interpret data (Saunders et al., 2012). Pragmatism is concerned with action and change and the interplay between knowledge and action. This makes it appropriate as a basis for research approaches intervening into the world and not merely observing the world. This is usually the case if an intervention is introduced to create an organisational change as in action research or building artefacts in design research (Goldkuhl, 2012; Iivari and Venable, 2009).

IV **Interpretivism or Anti-positivism**

This is used to describe an alternative stance to positivism. It advocates that a researcher may interpret data and gain an understanding of a context based on social roles (Bryman, 2004; Walsham, 1993). Interpretive researchers start out with the assumption that access to reality (given or socially constructed) is only through social constructions such as language, consciousness and shared meanings (Saunders et al., 2012). Interpretive research does not predefine dependent and independent variables but focuses on the full complexity of human sense-making as the situation emerges (Kaplan et al., 1994).

3.2.2 Philosophical choice and rationale

According to Podsakoff et al. (2012), one philosophy is not inherently better than the other, although researchers may favour one over the other based on their beliefs. Research in the sciences usually takes the positivist approach because "factual" knowledge is gained through observation (the senses), and includes reliable measurements. Also, in positivism studies, the role of the researcher is limited to data collection and interpretation through the objective approach and the research findings are usually observable and quantifiable. The research questions in the present study centre on developing innovative tools for computer-aided assessments. A positivist approach enables approaching the study objectively and provides a useful basis for obtaining observable empirical data and testing of hypothesis.

3.3 Research approach

When conducting research, it is helpful to determine which approach is being implemented (Bryman and Bell, 2012). The research approach may be viewed as the type of reasoning adopted in a study. According to Saunders et al. (2012), there are three main research approaches: deduction, induction and abduction.

3.3.1 Research approaches

The three research approaches as outlined in the research onion model onion model are briefly summarised.

I **Deduction**
The deductive approach develops hypotheses based on a pre-existing theory and then prepares the research to test it (Silverman, 2013). The logic of deductive inference states that when the premises are true, the conclusion must also be true. Generalisation is from general to specific, and data collection is used to evaluate propositions or hypothesis related to an existing theory (Bryman, 2012). The deductive approach is considered particularly suited to the positivist philosophy because it permits the formulation of hypotheses and the statistical testing of expected results to an accepted level of probability (Snieder and Larner, 2009).

II **Induction**
The inductive approach is characterised as a move from the specific to the general (Bryman and Bell, 2012). In this approach, the observations are the starting point for the researcher, and patterns are looked for in the data (Beiske, 2007). In an inductive inference, known premises are used to generate untested conclusions. Unlike deductive inferences, generalisation is from specific to the general. Also, in induction data collection is used to explore a phenomenon, identify themes and patterns and create a conceptual framework (Myers and Avison, 2002).

III **Abduction**
The term abduction (also referred to as 'retroduction') involves creating a theory based both on real-world observations that are inductively observed as well as theoretical viewpoints, premises, and conceptual patterns that are deductively inferred (Gregory and Muntermann, 2011). Saunders et al. (2012), described an abductive approach as one which moves back and forth between deduction and abduction. Here, known premises

are used to generate testable conclusions. Generalisation is from the interactions between the specific and general. Data collection is used to explore a phenomenon, identify themes and patterns, locate these in a conceptual framework and test this through subsequent data collection and so forth (Saunders et al., 2012).

3.3.2 Choice of approach and rationale

The abductive approach will be adopted for the study. This is because abduction involves both inductive and deductive reasoning and their continuous revision order to execute a study. The flexibility provided between the two approaches and modifications that may be made as the study progresses enables the generation of better ideas (Gregory and Muntermann, 2011). The research questions in the present study involve the design and testing of functional tools. Using this approach, a conceptual model will first be developed that will form the basis for data collection and analysis. The development will utilise both the inductive and deductive approaches. The model will be used to establish a series of hypotheses which will be tested.

3.4 Methodological choice

The third layer relates to the strategies that may be adopted given the philosophical position and selection of deductive , inductive or abductive approaches. This layer of the research onion highlights a basic but essential choice researchers face when designing their research: whether to use a quantitative method or methods, a qualitative method or methods, or a mixture of both (Saunders et al., 2012).

3.4.1 Common research designs

The choices outlined in the research onion include the mono method, the mixed method, and the multi-method (Saunders et al., 2012). The mono-method involves using one research approach for the study (i.e. qualitative or quantitative) while the mixed-methods requires the use of two or more methods of research. In the multi-method, a wider selection of methods is used (Bryman and Bell, 2012; Vaishnavi and Kuechler, 2015).

I **Quantitative research design**
Quantitative research is usually associated with a deductive approach where the focus is on using data to test the theory. It may also incorporate an inductive approach where

data are used to develop theory. Quantitative studies examine relationships between variable, which are measured numerically and are analysed using a range of statistical techniques and controls to ensure the validity of data (Vaishnavi and Kuechler, 2015). Although this methodological choice is informed by a positivist philosophy, it can be used to investigate a broad range of social phenomena, including feelings and subjective viewpoints (Bryman and Bell, 2012). The quantitative approach can be most effectively used in situations where there are a large number of participants available, where the data can be effectively measured using quantitative techniques, and where statistical methods of analysis can be used (Saunders et al., 2012).

II **Qualitative research design**

Qualitative design is closely linked with the interpretive philosophy (Denzin and Lincoln, 2005). It measures subjective data, such as beliefs and opinions (Myers and Avison, 2002). This is because researchers need to make sense of the subjective and socially constructed meanings expressed about the phenomenon being studied. (Saunders et al., 2012), suggest that many varieties of qualitative research commence with an inductive approach where the research design is used to develop rich theoretical perspectives as found in the literature. However, in some cases deductive and abductive approaches may be incorporated.

III **Multiple/mixed methods research design**

Mingers (2003) has suggested that rather than conducting a purely quantitative or qualitative research work, researchers can combine one or more research methods in the study. This concept is often referred to as triangulation. Triangulation involves the careful collection and review of data collected through different methods in order to achieve a more accurate and valid estimate of a construct (Oliver-Hoyo and Allen, 2006). The main difference between the mixed and the multi-method is that the mixed-method involves a combined methodology that creates a single dataset (Saunders et al., 2012). The multi-method approach is where the research is divided into separate segments, with each producing a specific dataset; each is then analysed using techniques derived from quantitative or qualitative methodologies (Creswell, 2013). Saunders et al. (2012) subdivided the multiple -method into four categories namely; multimethod quantitative study, multimethod qualitative, mixed method simple and mixed method complex.

3.4.2   Rationale for research strategy

Since both quantitative and qualitative research can be used together to support each other's findings, the multimethod research design has been adopted for the study. This choice was made because the multimethod research takes advantage of using multiple ways to explore a research problem. Interpretation is continual and influences the stages in the research process. This potentially provides deeper insights and facilitates explanation of the constructs being studied.

3.5   Research Strategies

The fourth layer relates to the strategies that may be selected given the choices made in outer layers of the research onion. The research strategy is how the researcher intends to carry out the study (Saunders et al., 2012). For example, if a deductive approach is adopted the strategies may be more experimental. On the other hand, if the inductive approach is selected the study may move towards active, grounded research or ethnographic research. The following provides an overview of dominant research strategies.

I **Experimental research**

This refers to the strategy of creating a research process that examines the results of an experiment against expected results (Saunders et al., 2012). It can be used in all areas of research, and usually involves the consideration of a relatively limited number of factors (Vaishnavi and Kuechler, 2015). The relationship between the factors are examined, and judged against the expectation of the research outcomes.

II **Surveys**

A survey is a research tool that collects demographics: fact-based information and psychographics: opinion-based information. Surveys tend to be used in quantitative research projects, and involve sampling a representative proportion of the population (Bryman and Bell, 2012). The surveys produce quantitative data that can be analysed empirically. Surveys are most commonly used to examine causative variables between different types of data.

III **Archival research**

An archival research strategy is one where the research is conducted from existing materials (Flick, 2015). The form of the investigation may involve a systematic literature review, where patterns of existing research are examined and summed up in order to

establish the sum of knowledge in a particular study, or to examine the application of existing research to specific problems. Archival research may also refer to historical research, where a body of the source material is mined in order to establish results.

IV **Case study research**

Case study research is the assessment of a single unit with a view to establish its key features and draw generalisations (Dubé and Paré, 2003). It can offer deep insight into the specific nature of any example, and can establish the importance of culture and context in differences between cases (Silverman, 2013). This strategy seeks to answer 'why' 'how' and 'what' in relation to issues behind reality of research investigations (Calzadilla et al., 2012). Flick (2015) opined that the use of case studies is particularly useful in studies where the opinions of individuals and key stakeholders are being examined.

V **Ethnography**

Ethnography involves the close observation of people, examining their cultural interaction and their meaning (Bryman and Bell, 2012). It is used to understand the connection between human behaviour and culture. In this research process, the observer conducts the research from the perspective of the people being observed and aims to understand the differences of meaning and importance or behaviours from their perspective.

VI **Action research**

Action research is characterised as a practical approach to a specific research problem within a community of practice (Bryman and Bell, 2012). It involves examining practice to establish that it corresponds to the best approach. It tends to involve reflective practice, which is a systematic process by which the professional practice and experience of the practitioners can be assessed. This form of research is common in professions such as teaching and nursing, where the practitioner can assess ways in which they can improve their professional approach and understanding (Saunders et al., 2012).

VII **Grounded theory**

Grounded theory is a qualitative methodology that draws on an inductive approach whereby patterns are derived from the data as a precondition for the study (May, 2011). For example, interview data may be transcribed, coded and then grouped accordingly to the common factors exhibited between respondents. This means that the results of the research are derived fundamentally from the research that has been completed, rather than where the data is examined to establish whether it fits with pre-existing frameworks (Flick, 2015).

VIII **Tracking**

Eye tracking for instance is a method where equipment is utilised to capture and analyse where a person is looking (Duchowski, 2007). Also, computer technology has been deployed to capture and analyse the clicking and scrolling behaviours of people - also called click or scroll tracking (Moe, 2003). These are especially useful for live websites, software, and applications, but can also be used to carry out click tests of nonlive designs. The use of tracking technologies enables the understanding of what actions participants are taking without having to rely on their memory or ability to self-report. However, it cannot tell why the paricipants are behaving a certain way.

3.5.1   Choice of strategies

Three choices were made for the present study. These are case- study, tracking, and experiment. The study focuses on the case of problem solving in elementary mathematics to provide insights into problem-solving processes and to serve as an example whereby the proposed design may be implemented and evaluated. In addition to this tracking will be implemented on the tools developed to capture assessment data so that the actions of the participants may be recorded without having to rely on their memory or self-reporting abilities. Experimental studies will be conducted to empirically evaluate the tools.

3.5.2   Experiments and Sampling

As mentioned in Section 3.5.1, experimental strategies will be used in the study. This will aid the investigation of the usablilty of developed tools. Also, it will facilitate comparison with traditional assessment methods.

In conducting experimental studies, sampling methods are used. There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified (Devore, 2015; Mason et al., 2003).

- In random sampling, each element in the population has an equal chance of being selected. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained.

- In systematic sampling, the list of elements is "counted off". That is, every k^{th} element is taken for the study. It is easier to implement than random sampling.

- In convenience sampling, easily accessible data is used. The elements are selected because of their convenient accessibility and proximity to the researcher such as when students volunteer for a study.

- Cluster sampling is accomplished by dividing the population into groups – usually geographically called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

- Stratified sampling also divides the population into groups called strata. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

Because of the time constraints for completing the study, convenience sampling will be utilised in the evaluation of the tools developed. The absence of randomisation of samples makes the study quasi-experimental. Although convenience sampling is vulnerable to selection bias and influences beyond the control of the researcher, it facilitates data collection in a short period of time (Devore, 2015). According to Mason et al. (2003) for pilot studies, convenience sample is usually used because it allows the researcher to obtain basic data and trends regarding the study without the complications of using a randomised sample . It is also useful in exploratory studies to gain initial primary data and trends regarding specific issues such as usability, quality and perception (Devore, 2015). During the early stages of a study the use of convenience sampling helps in obtaining quick and useful results.

3.5.3   Sample sizes

The goal of the evaluation studies is to test the designs and solutions so as to obtain insights, to find and document problems, and to improve the design. Qualitative usability testing has traditionally been based around small sample sizes of 5-20 participants (Holzinger, 2005). The main argument for small tests is the Return on Investment: testing costs increase with each additional study participant, yet the number of findings quickly reaches the point of diminishing returns. For Quantitative studies aiming for statistically significant numbers, at least 20 participants are recommended for a reasonably tight confidence interval (Nielsen, 2012). Nielsen (2010) suggests that it is that it is better to distribute time and resources to representative participant tests across many small tests instead of investing all resources on a single, elaborate study. For this research, a series of small studies will be conducted with 8 – 40 participants.

3.6   Time Horizons

The Time Horizon is the time framework within which the study is intended for completion (Saunders et al., 2012). Two types of time horizons are specified within the research onion: the cross-sectional and the longitudinal (Bryman and Bell, 2012). The cross-sectional time horizon is one whereby the data must be collected within a short period. It is also called the 'snapshot' time collection, where the data is collected at a certain point (Flick, 2015). This is used when the investigation is concerned with the study of a particular phenomenon at a specific time. A longitudinal time horizon for data collection refers to the collection of data repeatedly over an extended period, and is used where an important factor for the research is examining changes over time (Goddard and Melville, 2004). Usually, the time horizon selected is not dependent on a specific research approach or methodology (Saunders et al., 2012). The present study adopts the cross-sectional time horizon so as to collect and evaluate data to aid the development of more efficient computer-aided assessment systems within the timeframe for the study.

3.7   Research Strategies

Data collection and analysis is the innermost layer of the research onion. It is dependent on the methodological approach used (Bryman, 2012). The process used at this stage of the research contributes significantly to the study's overall reliability and validity (Saunders et al., 2007). Regardless of the approach used in a study, the type of data collected can be separated into two types: primary and secondary.

I **Primary Data**

Primary data is that which is derived from first-hand sources. This can be first-hand historical sources or the data derived from the respondents in survey or interview data (Bryman, 2012). However, it is not necessarily data that has been produced by the research being undertaken. For example, data derived from statistical collections such as the census can constitute primary data. Likewise, data that is derived from other researchers may also be used as primary data, or it may be represented by a text being analysed (Flick, 2011). The primary data is therefore best understood as the data that is being analysed as itself, rather than through the prism of another's analysis.

II **Secondary Data**

Secondary data is that which is derived from the work or opinions of other researchers

(Newman, 1998). For example, the conclusions of a research article can constitute secondary data because it is information that has already been processed by another. Likewise, analyses conducted on statistical surveys can constitute secondary data (Kothari, 2004). However, there is an extent to which the data is defined by its use, rather than its inherent nature (Flick, 2015).

This present study aims at obtaining primary research data to answer the research questions. The study involves the design, development and analysis of innovative tools aimed at providing better understanding on how computer-aided assessment systems may be improved.

3.8 Validity and Reliability

As mentioned in Section 3.3, quantitative and qualitative data will be collected to answer the research questions. Measurements provide an important connection between empirical observation and mathematical expression of quantitative relationships (Devore, 2015). Bryman and Bell (2012) underscores the importance of taking steps to ensure the key requirements of validity and reliability in the study. Validity and Reliability are the psychometric properties of measurement variables, that describe the benchmarks against which the adequacy, accuracy and quality of quantitative or qualitative method procedures are evaluated (Devore, 2015). The following briefly discusses the concepts and steps taken to ensure valid and reliable research outcomes in the present study.

3.8.1 Validity

Validity describes whether the data collected really measure what the researcher set out to measure. It encompasses the entire experimental concept and establishes whether the results obtained meet all of the requirements of the scientific research method (Shuttleworth, 2009). Different validity types are described in literature (Recker, 2012; Shuttleworth, 2009; Vaishnavi and Kuechler, 2015). Face validity refers to whether an indicator seems to be a reasonable measure of its underlying construct ("on its face"). Content validity refers to how well a set of measurement items matches with the relevant content domain of a theoretical construct. The key question of content validity is whether the instrumentation (questionnaire items, for example) pulls in a representative manner all of the ways that could be used to measure the content of a given construct(Recker, 2012; Vaishnavi and Kuechler, 2015). Other

types of validity include construct validity, discriminant validity, statistical construct and external validity (Bryman and Bell, 2012; Saunders et al., 2012).

The following steps will be taken in the conduct of the study to ensure validity in constructs measured.

- The experimental designs will be structured to encompass the steps of the scientific research method

- Before carrying out the main studies, pilot test will be conducted to see how participants work with developed tools. This provides opportunities to develop consistent practices and enhances the data integrity

- Control groups will be employed to provide baseline data with which results are compared

3.8.2 Reliability

Reliability describes the extent to which a variable or set of variables is consistent in what it is intended to measure (Shuttleworth, 2009). It is such that if multiple measurements are taken, the reliable measurements will be consistent in their values. Reliable measurements approach a true, but unknown "score" of a construct. Sources of reliability problems often come from a reliance on subjective observations and data collections (Recker, 2012). All types of observations one can make as part of an empirical study inevitably carry subjective bias because researchers can only observe phenomena before the background of their own history, knowledge, presuppositions, and interpretations at that time(Mason et al., 2003). Other sources of reliability problems stem from poorly worded questions that are imprecise or ambiguous, or, simply, by asking respondents questions that they are either unqualified to answer, unfamiliar with, predisposed to a particular type of answer or uncomfortable to answer. In the study validity will be ensured by taking the following steps.

- The user interfaces of the tools will be designed for problem-solving suitable for primary school students

- In the conduct of evaluation studies, interview scripts will be developed and used for all the participants to ensure consistency

- In addition to observations made in carrying out the empirical measurements, questionnaires will be given to participants to obtain more factual data

- The standard arithmetic word problems for the study will be obtained from mathematics education literature

3.9 Overview of complete research design

The present study involves the design, development and evaluation of information system tools for computer-aided assessment. It includes the search for, construction and synthesis of methods to improve assessment and feedback practices. The research design is the description of how the research process will be completed (Nunamaker Jr et al., 1990). The above discussions provide a basis for arriving at a complete research design for the study. Peffers et al. (2007) proposed and developed a model for the production and presentation of research involving design and implementation of tools and artefacts. This model showed in Figure 3.2 provides steps for the research process which can guide design and development studies.
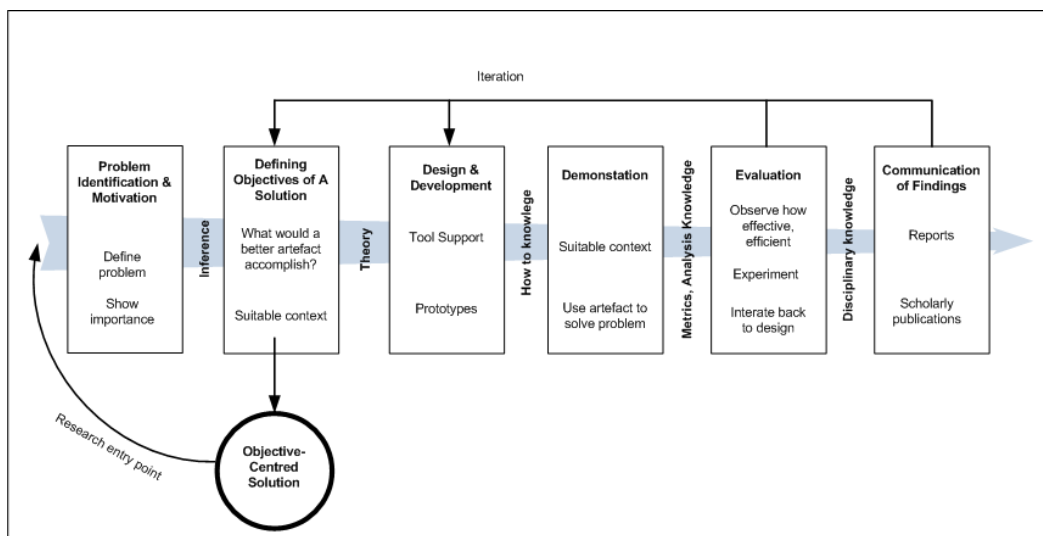
Fig. 3.2 Complete Research Design (Peffers et al., 2007)

The model is made up of six distinct steps. The first step involves identifying the problem(s) for which data will be collected. It provides the context and motivation for the study. Because identified problems do not necessarily translate directly into objectives for the tools, this step iteratively combines with the second step until clear objectives are arrived at. The second step involves the determination the objectives for a solution and is closely associated with the first phase. The step encompasses the the definition of objective centered solution

from the problem definition and knowledge of what is possible and feasible in attaining a desirable solution. An iteration between the first two steps leads to the drawing of an objective-centered solution. Chapters 1 and 2 of this thesis identified the research problems and outlined the objectives of the study. As outlined in Chapter 1, the present study aims to develop and evaluate a semi-automatic computer-aided assessment framework that can potentially extend and improve on current assessment practices by providing opportunities for detailed feedback.

The conceptual framework for the proposed solution is described in Chapter 4. The chapter sets the stage for the research and specifies the components of the semi-automatic framework. The third step involves the design and development of the tools arising from the components specified in the framework. This step produces the prototype tools in which the research contributions are embedded. The activity here includes determining the functionality of the tools, describing the architecture and selecting the appropriate platforms to implement the solutions. The fourth step encompasses demonstrations to prove that the idea works. The tool is used to to solve one or more instances of the problem. Chapters 5 and 6 sets out the design, development and demonstration of the prototype tools. Running examples from primary mathematics are used to demonstrate these functionalities.

The fifth step is the evaluation of the designed solutions. Evaluation is a key step in the research process and centers on observing and measuring how well the tools support a solution to the problem. This involves empirically comparing the objectives to the actual measured results. The evaluation in done in terms of functionality, completeness, consistency, accuracy, performance, reliability and usability. The empirical evaluation of the prototypes with real users is presented in Chapters 7 and 8. The chapters validates the solutions and informs future iterations.

The final step proposes the need for communication to diffuse the resulting knowledge. This involves communicating the problem and its importance, the tool and its utility, novelty, design and effectiveness to different audiences. This thesis, and the scholarly publications outlined in Section 1.7, are means to communicate the outputs of the study.

3.10   Ethical considerations

Access and ethics are critical aspects for the conduct of research studies. Research ethics refer to the standards of behaviour that researcher's follows in relation to the rights of those who become subject of the study or are affected by it (Thomas et al., 2008). Ethics define

the principles of right and wrong conduct in a community or profession and can be used to make choices to guide behavior (Cocks, 2006; Thomson and Schmoldt, 2001). Gaining access involves convincing people that the researcher has decided upon on who should be the participants that would provide information in conducting research (Feldman et al., 2004). This would require the researcher to talk to many people by developing rapport with them and to be in a position to learn from them (Feldman et al., 2004). According to the World Health Organization, it is important to adhere to ethical principles in order to protect the dignity, rights and welfare of research participants (World Health Organization, 2016). As such, all research involving human beings are reviewed by an ethics committee to ensure that the appropriate ethical standards are being upheld.

The evaluation of the prototype tools developed for the present study will involve the use of human participants i.e. children, students and teachers. Saunders et al. (2012) and Bryman and Bell (2012) summarized ethical issues that should be taken into consideration in research studies. This is presented in Figure 3.3.
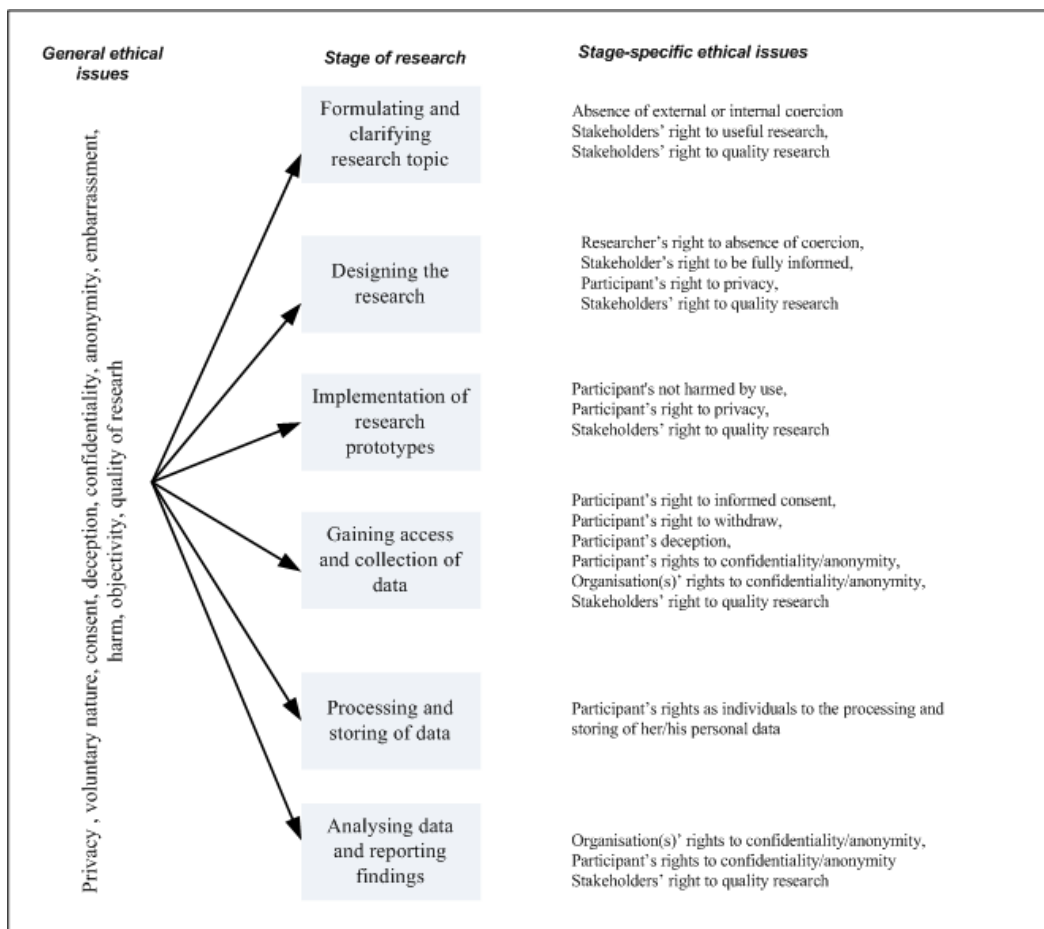
Fig. 3.3 Ethical issues at different stages of the research study (Saunders et al., 2012)

In line with these requirements the following steps will be taken to ensure that the study complies with high ethical standards.

1. Before the data collection commences, approval will be sought and obtained from the Research Ethics Committee of Loughborough University

2. The informed consent of participants will be obtained before involving them in the study.

3. Participants will not be subjected to harm in any way as they use the research tools. Details of the requirements of the study will be clearly communicated to all participants

4. Participants in the respective schools will not be subjected to coercion in any way. All participants will be informed of their freedom to choose whether or not to participate in the study without any consequence

5. The privacy of the research participants will be ensured so that no personal data will be collected from respondents. No individual will be identified based on the data collected. Pseudonyms will be used to refer to participants in discussions relating to the data collected.

6. The participants will be briefed about aims and objectives of the study before the primary data collection process. All the details about the study are summarized in the consent and information forms used for the study (see Appendix A).

7. All research data, including primary materials and raw data such as questionnaires, measurements, recordings and computer captured data will be stored in suitable and durable storage facilities.

These steps help to ensure that the study is conducted with high ethical standards. As mentioned earlier, the context of the study is within elementary mathematics. As a result, children participants will be sought during the evaluation. According to Flewitt* (2005) and Anderson et al. (2008), four additional provisos are specific to research involving children. These are:

1. Children's competencies, perceptions and frameworks of reference, differ according to factors such as age and environment

2. Children are potentially vulnerable to exploitation in interaction with adults

3. The differential power relationships between adult researcher and the child participant must be taken into consideration

4. The role of adult gatekeepers in mediating access to children, with concomitant ethical implications in relation to informed consent

To meet these requirements, additional steps need to be taken in addition to those outlined earlier.

- An initial consent letter will be obtained from the gatekeepers (i.e. parents and teachers or others with a duty of care for the child. This permission does not mean the child has consented to participate, but it allows consent from the child to be sought

- Individual consent will be sought from each child rather than as a group. This is because some children in the group could feel pressured to take part against their wishes (Cocks, 2006; Morrow, 2008)

- Question items used for the study will be carefully worded and made simple to facilitate easy understanding by the children

3.11   Chapter summary

This chapter has presented a detailed account of the research philosophy, approach, method, strategy and ethical considerations used for the study. A brief review of different alternatives and available choices were first outlined. After this appropriate choices and rationale for them were presented. A positivist and interpretive philosophical stance was adopted, abduction was selected as the research approach and mixed/multiple method methodological choice was made. Regarding the strategy case-study, experiment and tracking strategies were viewed as appropriate for the investigations. The overall research process was summarized by the complete research framework flow diagram. Building on these choices, the following chapters present the theoretical and practical details of the study.

Chapter 4

The Semi-Automatic Assessment Framework

This chapter presents the development of a new framework for semi-automatic assessments in education, with a specific focus on elementary mathematics. Chapter 2 provided the background and the need of having CAA systems that can measure assess the product and process of problem solving. It also explored intelligent techniques that can be used in marking and feedback production. A framework has been defined as a broad overview, outline, or skeleton of interlinked items that supports a particular approach to a specific objective (Framework, 2016). The semi-automatic CAA framework provides a combination of approaches to help address the needs. It aims at aiding students to respond compressively to an assessment task and supporting teachers to assess students' responses consistently and efficiently. This chapter presents in detail the three important techniques that would be used to build tools for the semi-automatic CAA system. These are multi- touch technologies, analytic rubrics and the case-based reasoning AI methodology. The framework will serve as a prelude to the design development and evaluation of the tools. The chapter also provides theoretical background and discussions from the literature regarding the technologies, design alternatives, design decisions and their justifications for the choices. In the next section, a general direction for the study is outlined. This is followed by a broad description of the conceptual framework. After this, discussions on interactive technologies for problem solving, rubrics and the case-based reasoning method for marking and feedback are presented.

4.1   Rationale for a semi-automatic assessment approach

Section 2.8 contrasted the three major assessment approaches i.e. fully automatic, semi-automatic and manual systems. Manual systems were described as allowing for a high degree of personalisation from the teacher but suffer from slow speeds and inconsistencies. Fully automatic systems, on the other hand, are largely consistent have the advantages of instant scoring. However, they have restricted question types and low degree of personalised feedback. Although fully automatic assessment systems for selected responses such as multiple-choice and true-false tasks are now widespread, the use of automatic assessment systems for constructed response question tasks like essays has not yet received widespread endorsement by educators. Critics have advanced several reasons suggesting that automated systems are no match for live teachers. For instance, Yang et al. (2002) argued that automated scoring of constructed responses suffer from an overreliance on the surface feature of responses, insensitivity to the content of responses, creativity and the vulnerability to new types of cheating and test-taking strategies. Dikli (2006) and Yang et al. (2002) articulated the concern that students' motivation will be reduced if they know that no human will read their responses. In 2013, Humanreaders.org (2013) launched an online petition, calling for schools, colleges and educational assessment programs to stop using automatic scoring for students' responses to essay tasks in high stake exams. They listed several research findings suggesting that students, parents, teachers, employers, administrators and legislators should not rely on automatic scoring of essays and that automatic assessment does not promote the authentic arts of writing (Humanreaders.org, 2016).

These arguments advanced suggest that human judgement is still considered critical for assessment of complex tasks. The semi-automatic assessment approach which allows teachers to make judgments while the computer supports the teacher in storing and reusing his or her judgments is highly desirable. This creates a system that leverages on the advantages of both the automatic and manual assessment approaches.

4.2   General direction

In Chapter 2, a semi-automatic assessment system was described as a computer-aided assessment system that actively combines human and intelligent computing techniques in the assessment process. The approach promises the integration of the better of two worlds i.e. human ability to make a judgement and provide feedback on a piece of work and the computer's strength in providing fast, consistent and reliable processing. The objectives of

the study are twofold: first, to provide an assessment environment that can assist students
to input their responses to word problems in elementary mathematics; second, to provide
teachers with support tools and techniques for detailed feedback and consistent marking of
students' submission for more complex tasks.

A few studies have adopted the semi-automatic approach for different domains of study; for
example student texts (Tselonis et al., 2005), in graph diagrams Batmaz and Stone 2009,
programming courses (Watanabe et al., 2001) . However, not many studies have detailed the
approach within a framework which may be used as a guide for developing CAA assessment
systems.

4.3   Overview of the framework

This study approaches the assessment process from the perspective of the two main stake-
holders (students and teachers). An overall architecture describing the series of steps is
shown in Figure 4.1.
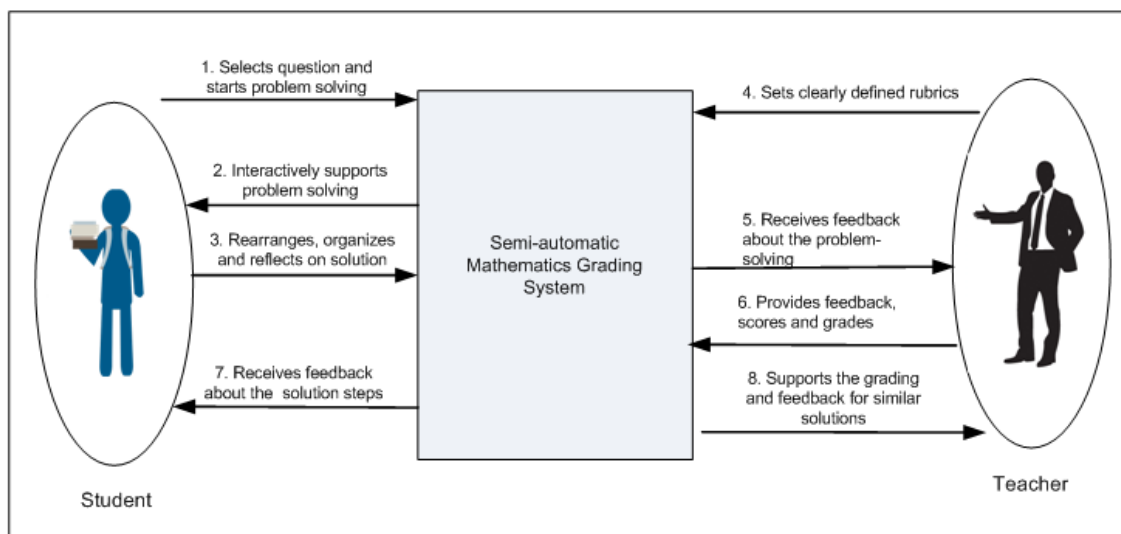
Fig. 4.1 A general overview of the semi-automatic system

As indicated in the figure, the process begins with a student answering a question retrieved
from the question bank. The computer supports the problem-solving effort by providing a
highly interactive environment offering innovative ways to express solution steps. The student
is provided opportunities to rearrange, reflect and make changes as desired to complete a

detailed response. The computer logs the interaction and activities and presents these to the teacher who receives them and proceeds to mark and provide feedback comments on the various parts of the student's response using clearly defined rubrics. After marking the work, the feedback is passed to the student. Where there is a pool of students' responses on the same question waiting to be graded the computer supports the teacher by reusing previous feedback comments and marks to similar students' responses. The teacher only makes judgments on new or different responses in the response pool. Given all these steps, the three main components of this framework are: interactive problem-solving environment, clearly defined rubrics, and re-use of marking experiences.

A detailed breakdown of the framework is presented in Figure 4.2. The figure shows the components of the framework and their interrelationships starting from when a student begins solving a word problem to when a teacher marks and provides feedback. The first part relates to the capture of problem solving steps with a user interface that facilitates interactive problem-solving. Interaction traces, logs of the student actions and text input are all captured and stored in a records database. This part is identified by number 1 in the diagram. The second component addresses the creation and application of analytic rubrics by the teacher to facilitate scoring and input of feedback comments. The second component, labelled number 2 in the diagram, closely links with the third segment (number 3) which aims to support the marking process by applying the case-based reasoning methodology using the case library and records database.
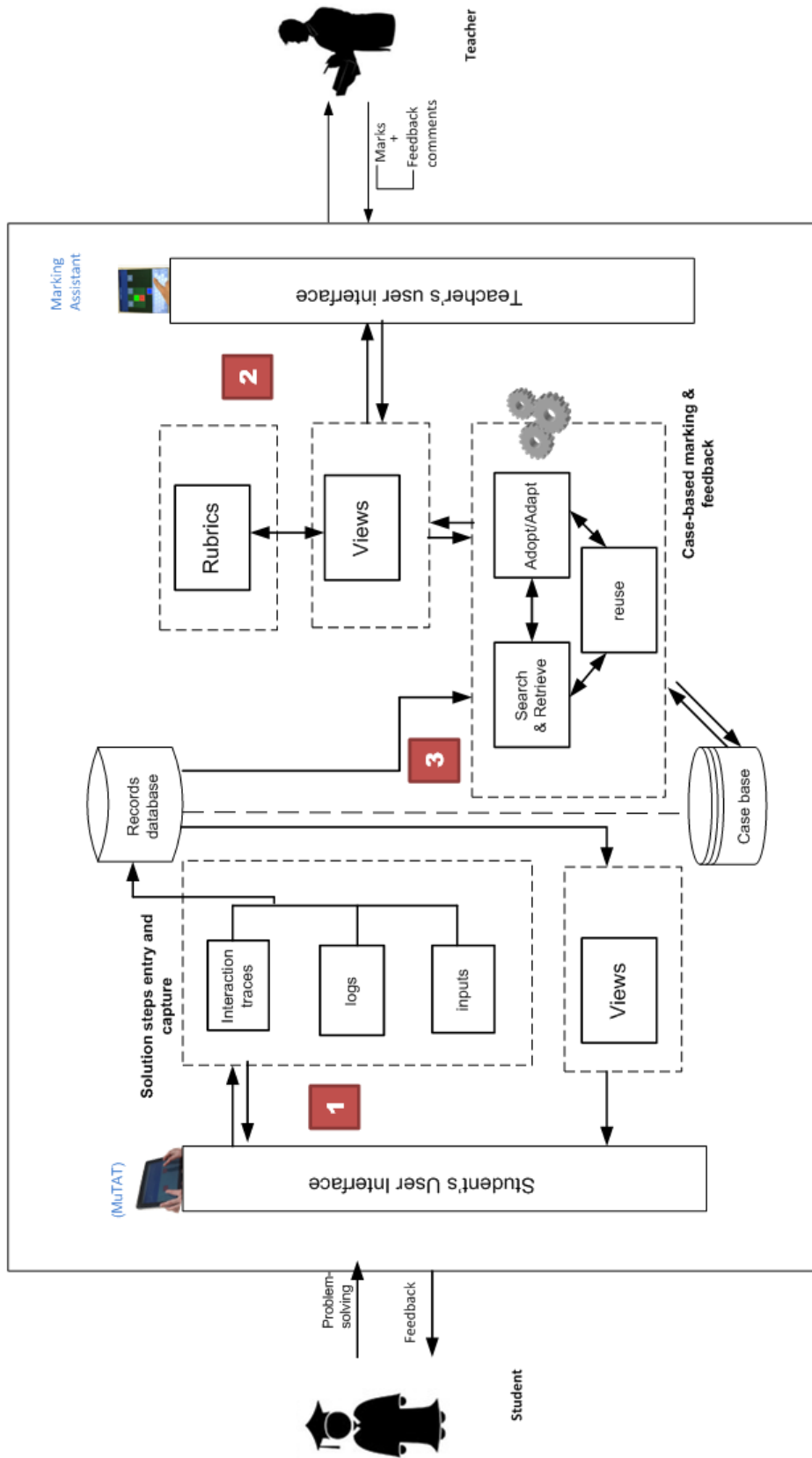
Fig. 4.2 The integrated semi-automatic assessment architecture

Discussions on the various aspects of the framework are presented in the following sections.

4.4   Interactive problem-solving environment

The first aspect of the framework involves interactive problem solving. Interactive environments in education are not new. In many educational settings, manipulative materials play a major role in students' learning, enabling then to explore mathematical and scientific concepts (such as number and shape) through direct manipulation of physical objects (Resnick et al., 1998; Zoanetti, 2010). Generally, manipulatives refer to physical objects specifically designed to foster learning. Moyer (2001) defines manipulatives as "objects designed to represent explicitly and concretely mathematical ideas that are abstract" (Moyer, 2001, p. 176). When used in a computing environment, they are called digital manipulatives (e.g Resnick et al., 1998; Zuckerman et al., 2005). Some researchers in mathematics education have suggested that these manipulatives enable students to explore a set of concepts, develop well-grounded, interconnected understandings of mathematical ideas and communicate their thinking processes in problem solving (Stein and Bovalino, 2001). Young children for instance often solve addition and subtraction problems by counting concrete objects, beginning with their fingers. They then go on to use concrete objects such as base-ten blocks and counters to develop more sophisticated problem-solving strategies based on what they know about counting (Siegle, 1996).

4.4.1   Theoretical underpinnings

Learning theorists have advocated for some time that students' concepts evolve through direct interaction with the environment, and materials provide a means through which this can happen. According to learning theory based on the ideas of psychologist Jean Piaget, children are active learners who master concepts by progressing through three levels of knowledge–concrete, pictorial, and abstract (Piaget, 1971). He suggested that concepts are formed by children through a reconstruction of reality, rather than through an imitation of it. Similarly, Dewey (2007) argued for the provision of first-hand experiences in a child's educational program. Bruner (1966) indicated that knowing is a process, not a product and emphasis should be placed on processes. Dienes (1971), whose work specifically relates to mathematics instruction agrees with Piaget and suggests that children need to build or construct their concepts from within rather than having those concepts imposed upon them. In the same vein, Pestalozzi (1803) asserted that children learn through their senses and through

a physical activity he argued that that things come before words, and the concrete comes before the abstract. These studies have contributed to the development of the constructivist learning theory which influences the development of educational tools.

Computation and communication capabilities play a critical role. They enable physical objects to move, sense and interact with one another – and make systems-related concepts more noticeable (and manipulable) by students. As Resnick et al. (1998) states "... when students work with virtual manipulatives, they not only become more motivated in science activities but also develop critical capacities in evaluating scientific measurements and knowledge, make stronger connections to the scientific concepts underlying their investigations, and develop deeper understandings of relationships between science and technology" . Computer modelling environments such as Stella, Starlogo and Model-It have made it possible for college students to learn by modelling and exploring systems phenomena (Ballera et al., 2015; Resnick et al., 1998; Zuckerman et al., 2005). While most of these studies have focused on the learning advantages provided by the computing environment few studies have considered their usefulness in assessments. These manipulatives can help reveal abstract ideas and make it easier to express symbolic operations.

4.4.2   Design process in problem solving

The constructivist theory of learning suggests that students need software tools to support their problem-solving tasks. According to Resnick (1998) and Williamson et al. (2004), a student's response or solution to a problem can be conceived as a design process just as professional designers routinely employ computer-aided design (CAD) systems in their design activities. Four reasons were advocated by Resnick for why design tasks can provide rich opportunities for learning.

1. Design tasks engage students in as *active participants*, giving them a greater sense of control over (and personal involvement in) the learning process, in contrast to traditional school activities in which teachers aim to transmit new information to the students.

2. Design activities are often *interdisciplinary* bringing together concepts from the arts, mathematics, and sciences.

3. Design activities encourage *pluralistic thinking* avoiding rigid right/wrong dichotomy, suggesting that multiple strategies and solutions may be possible.

4. Design activities provide a context for reflection. A student's construction may serve as external shadows of the internal mental models – providing an opportunity for the student to reflect on, revise and extend personal internal models of the world.

A common example of this is word problems in mathematics. Word problems have long constituted a major part of elementary school mathematics. Verschaffel et al. (2000) list reasons to justify this privileged position to include: providing opportunities for students to use mathematical tools, motivating the link between mathematics and real-life context, encouraging thinking and the use of problem-solving heuristics, and providing a platform to develop new concepts and skills. However, critics like Gravemeijer (1997), have questioned such justifications by suggesting that word problems are a mere disguise for practice in the four basic operations. Nevertheless, the characteristics of word problems such as length of problems, placement of questions and semantic relations provide rich opportunities for designing solutions in computing environments. Table 4.1 gives examples of word problems that contain five semantic relations.

Table 4.1 Examples of word problems with each semantic relation (Yeap and Kaur, 2001)

Semantic relation	Example
Change	Chris had 623 stamps. He gave his friend 572 stamps. How may stamps does Chris have now?
Group	Ginny baked 315 chocolate cookies. She also baked 49 vanilla cookies. How many cookies did she bake?
Restate	Royson has 316 marbles. Ray has 49 marbles more than Royston. How many marbles does Ray have?
Vary	There are 25 cookies altogether. Valerie puts them into 5 boxes. How many cookies are there in each box?
Compare	Mason has 120 stamps. Molly has 115 stamps. Who has more stamps?

As can be seen from the table the problems usually contain numbers, words and sentences. In this study, we argue that the numbers may be made "interact able" in a software environment and used in the design of a solution. This technique is illustrated in Figure 4.3.
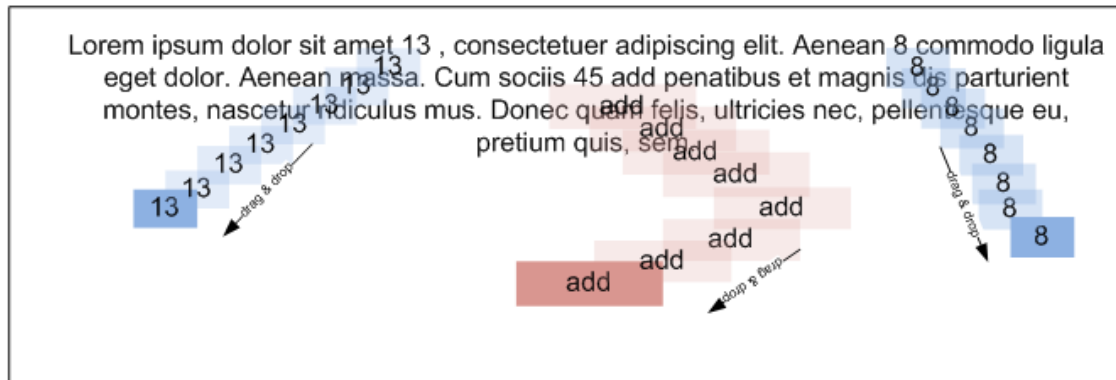
Fig. 4.3 Making elements of a word problem "interactable"

The numbers 13, 8, 45 and the word "add" can be selected a computer environment and transformed to become digital manipulatives which can be creatively used to provide a response to the question. The ability to select and drag items interactively reduces copying errors and provides some scaffolding to students as they solve the problems. Another advantage is that it provides graphic representation of problem-solving. Goldman (1989), stated that practical word problem solving requires the creation of a representation of the problem that mediates the solution. The graphic representation allows for the use of schematic diagrams that enable students to organise information in a problem to facilitate problem translation and solution. He further argued that the external representation provide scaffolds that may serve to reduce the student's cognitive processing load and make available mental resources for engaging in problem analysis and solution.

4.4.3 Providing interactivity with touch gestures

Grounded cognition and embodied interaction research suggest that action can play significant roles in perception, acquisition and thought. Gibson (1979), for example claimed that action underlies perception and the ability to perceive evolved from a need to interact with the world. Barsalou et al. (2003) found that there is a correlation between one's physical state and one's mental state. Gestural interfaces in computing environment provide a more hands-on experience and therefore could support cognition and create meaningful learning experiences by using a more direct manipulation of objects (Segal, 2011). There is also a growing body of research based on embodied cognition theory which argues that physical manipulation of objects support thinking and learning (Bara et al., 2004; Glenberg et al., 2004; Siegle, 1996). Also, studies about technological devices and learning provide evidence that incorporating

the haptic channel yields better learning performance (Chan and Black, 2006; Hallman et al., 2009; Jang, 2010). These findings suggest that incorporating gestural actions in educational tasks may enhance teaching and learning efforts. A list of common gestures used in computing environments is presented in Figure 4.4.
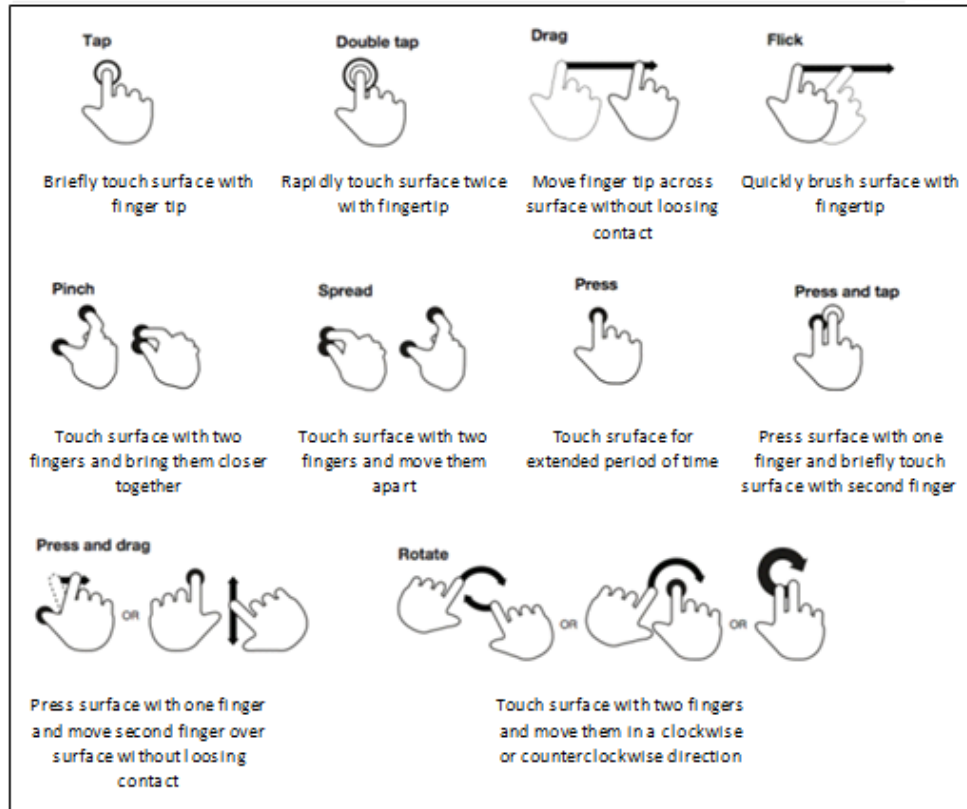
Fig. 4.4 Basic touch gestures (Bank, 2015)

This research suggests that using multi-touch gestural interfaces can provide needed inter-activity in the assessment process. Multi-touch technologies present novel opportunity to increase the input bandwidth between humans and the computer that can aid capture of problem-solving steps. Advances in computing hardware technology following Moore's law has led to higher processing speed, memory capacities, sensors and the number and size of pixels in digital cameras (Myhrvold, 2013). As a result, computers are getting smaller, cheaper and more capable of advanced functions. These developments are reflected in changes in the way humans interact with computers over the years. For instance the command line interface (CLI) was the primary way of interacting with the PC in the 1970s. The arrival of the graphical user interfaces (GUI) of the 1980s brought about the WIMP interfaces (windows, icons, menus and pointing devices) as an improved method of interac-

tion. In recent times the landscape is changing to natural user interfaces (NUI) multi-touch devices, voice control, gesture input and augmented reality (Wigdor and Wixon, 2011). van Dam (1997) described a Post-WIMP interface as one containing at least one interaction technique not dependent on classical 2D widgets such as menus and icons which ultimately will involve all senses in parallel, natural language communication and multiple users. The multi-touch technologies (handhelds, tablets, tabletops and whiteboards) are at the forefront of this revolution.

Multi-touch refers to the ability to simultaneously register two or more distinct positions of input touches on a touch screen or surface. The unique features and characteristics of multi-touch technologies present several advantages over traditional computers with mouse and keyboard. Some of these as highlighted by several studies are outlined below:

- *Abstraction*: Multi-touch removes abstraction from the interaction process. The direct manipulation of digital objects on the surface allows for working with virtual manipulatives directly instead of being mediated through another input device such as a mouse.

- *Fun:* The multi-touch environment provides a playful environment that engages users. Their interactive nature is engaging and appeals to a wide variety of learning styles (Heinrich, 2011). Skinner and Belmont (1993) argued that students who are engaged show sustained behavioural involvement in learning activities. Such students can exert intense effort and concentration in the implementing learning tasks.

- *Parallelism*: Multi-point interaction allows more parallel interaction. Multi-touch technology can capture multiple touches on a screen and convert these actions into events that can be interpreted by software. This permits performance of a complex task in a reduced time (Jiao et al., 2010). In the simplest case these might be mouse actions like mouse clicks and mouse drags, but because several touches can be combined into gestures it is possible for the user to give more information than just translations (that correspond to dragging the mouse) or (x,y) positions that correspond to clicking the mouse (Villar et al., 2009). For instance, with two fingers it is easily possible to rotate or scale or move several objects at once, each in a separate direction. With more fingers more degrees of freedom are available to input almost arbitrary transformations. This results in a reduction of task switching times.

- *Collaboration*: The split of concurrent touch points between multiple users makes it possible to create learning environments using large screens (for example multi-touch-

tables), which encourage collaborative learning and communication (Don and Smith, 2010; Mercier and Higgins, 2013).

- *Easier and faster*: Multi-touch provides a good illustration of Fitts's Law (Fitts and Peterson, 1964). The law states that the time needed to move to a target area is a function of the distance to and the size of the target. In effect, it means that big icons are easier to hit than little ones, top-of-screen menus are easier to click on than top-of-window ones, and pop-up menus are faster than pull-downs

- *Intuitive and natural*: The interfaces support intuitive and natural interactions and allow for rough motor skills and imprecise manipulation. These, in combination with the visually appealing interface elements and graphics, lead to an increased overall user experience (Moscovich, 2007).

- *Bi-manual interactions:* The multi-touch favours two-handed interaction which is common in the physical world (Bailey and Garner, 2010). Don and Smith (2010) showed that it allows for bimanual text entry.

However, multi-touch technology does have some challenges. Several authors have pointed out these limitations. For instance Villar et al. (2009), noted that the playfulness of the interface encourages ephemeral interactions. Jacucci et al. (2010) illustrated the limitation to the implementation of full gestural language; he argued that virtual objects in the 3D space require six degrees of freedoms to be manipulated in full detail (i.e. translations and rotations in all three dimensions). In contrast, the multi-touch input is sampled from a 2D surface, giving only 3 degrees of freedom (translation in the x/y dimensions and rotation around the z-axis). Other authors have argued that the interfaces lack the precision usually afforded by indirect pointing devices (Hansen et al., 2009), are inconsistent across different manufacturers (Kammer et al., 2010) and can result in task complexity (Davies, 2010). Some activity theorists have argued that conscious learning and activity are completely interactive and interdependent (Jonassen, 2000). In this thesis, we argue that multi-touch gestures can be used to to perform arithmetic work intuitively. Numbers from the word problems (see Fig. 4.5) may be further utilised for the solution by using multi-touch gestures to indicate intentions. In primary schools' operations and algebraic classes, for example, addition is taught as putting two numbers together and adding. Likewise, subtraction is often modelled as taking apart and taken from (Dolan et al., 2012). According to Resnick et al. (1998), it is vital that students have many varied experiences building number sentences (equations) through the use of concrete manipulatives. This may incorporate the tactile, visual, and abstract experiences and assist in developing conceptual understanding. A multi-touch computing environment provides varied opportunities for these. Figure 4.5 illustrates how

addition and subtraction may be modelled in using digital manipulatives and multi-touch gestures. Intuitive models of multiplication and division may also be conceptualised and implemented.
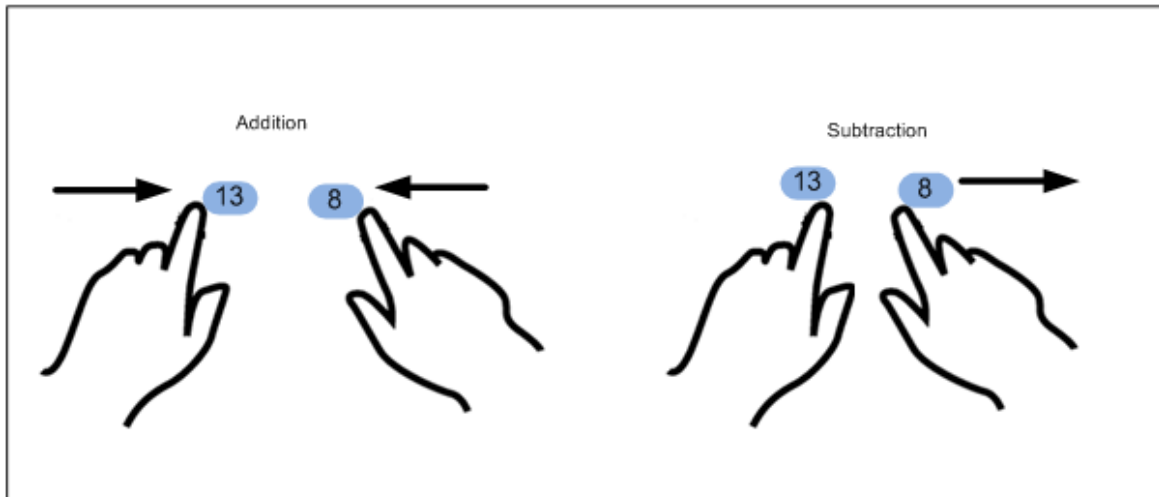
Fig. 4.5 Multi-touch gestures for arithmetic work

Because a multi-touch environment provides increased opportunities for expression through direct and intuitive touch gestures, it may be useful in computer-aided assessment environments, problem-solving steps can be made explicit and captured as they can enable the decomposition of problem-solving steps. Touch gestures

- Allow direct and intuitive expression

- Engage the student in problem solving

- Provide opportunities for decomposition of steps

- Allow for sequence detection

- Provide a non-intrusive method of obtaining assessment data

- May provide opportunities for collaboration in problem solving

The use of multi-touch gestures for problem solving is explored in Chapter 5.

4.4.4 A further look at the word problem example

Consider the arithmetic elementary mathematics problem outlined in Section 2.7: Alice has 34 black beads, 18 blue beads and 6 green beads. How many beads does Alice have altogether?
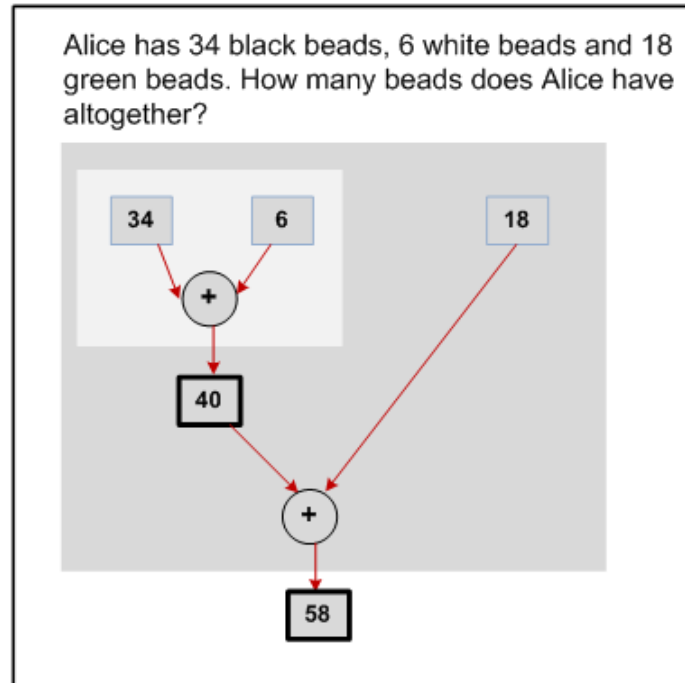
Fig. 4.6 Word problem example

The figure shows digital manipulatives being used to "design" a response to a multi-step word problem. In multi-step word problems, one or more problems have to be solved in order to get the information needed to solve the question being asked. Each step in the solution process potentially represents a solution case.

4.4.5 Extracting assessment information for interactive user-interface events

When interactive activities occur in a computer environment, they provide detailed information regarding student problem-solving behaviour which may be captured, searched, counted and analysed using automated tools (see figure 4.2). All these can be logged and stored in a

repository. However, there is a need to search, separate and select events of interest from the event/interaction streams. The results of the selections, transformations and analyses may then be visually presented. The process of extraction and visually presenting captured data is illustrated in Figure 4.7.
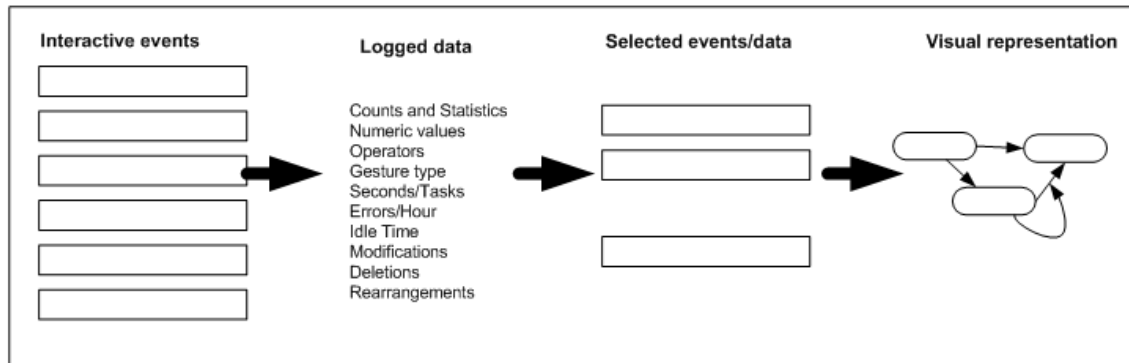
Fig. 4.7 Extracting assessment information from interactive user-interface events

4.5   Explicitly defined rubrics

Rubrics are important in assessment tasks (Campbell, 2005).Valuable assessment and eval-uation of any performance depend on the accurate and reliable measurement of major performance factors. As mentioned in Chapter 2, low-level understanding is conveniently in-vestigated with the help of simple, quantitative tools, such as multiple-choice, true-false, and other selected-response types. On the other hand, systems thinking, procedural knowledge, and attitude formation require more sophisticated measurement schemes. By clearly stating significant performance criteria, rubrics classify and organise performance observations with respect to different skill levels, behaviours, and/or product quality. Rubrics have been described as a "scoring tool that lays out the specific expectations for an assignment" (p.4 Stevens and Levi, 2011). They are usually aligned along clearly defined learning intentions or curriculum objectives ,and they can help simplify the grading process. Charles et al., (1987) referred to rubrics as a statement of characteristics associated with different levels or grades of performances by which a student work can be understood. Rubrics are designed to help teachers measure the ability to use and apply factual, conceptual, procedural, and metacognitive knowledge (Anderson and Krathwohl, 2001; Bloom, 1956).

4.5.1 Holistic and analytic rubrics

The two types of rubrics which are commonly used are holistic and analytic rubrics. A holistic rubric requires the person doing the measurement to score the overall process or product as a whole, without judging the component parts separately (Nitko, 2001). They are quick to use and provide the measurer with a snapshot of the performance at hand. One limitation of use is the inability to provide detailed and specific feedback of the performance. While holistic rubrics provide a single base or an overall impression of a student's performance on a task, analytic rubrics provide specific feedback along several dimensions (Jonsson and Svingby, 2007; Stevens and Levi, 2011). Analytic rubrics help both students and teachers identify strengths and areas for improvement. However, scoring and use may take longer than with a holistic rubric.

4.5.2 Rubrics application in mathematics

The semi-automatic assessment described in Figure 4.2 will require analytic rubrics to properly assess and provide feedback comments on different aspects of a problem-solving effort. For example in the arithmetic problem in Section 2.7, an assessor may be able to separate what the student is trying to do (based on his understanding) from his ability to perform calculations. This way, the student can get feedback on three categories i.e. calculation, conceptual understanding, and strategy used. This is in contrast to the assignment of a single score as in holistic rubrics. Using analytic rubric provides several opportunities for extensive feedback. Other assessment information such as time taken to complete a step, overall time, and count of interactions can also be obtained. Garfield (1994) suggests that these type of attributes need not be given a score or grade, but they can inform the teacher about understanding, feelings, and frustrations and can also serve as inputs to modifying instruction.

Table 4.2 Analytic scale for scoring three categories problem-solving (Szetela, 1992)

Understanding the problem
0 – No attempt
1 – Completely misinterprets the problem
2 - Misinterprets major part of the problem
3 - Misinterprets minor part of the problem
4 - Complete understanding of the problem

Solving the problem
0 – No attempt
1 – Total inappropriate plan
2 - Partially correct procedure but with major fault
3 – Substantially correct procedure with minor omission or procedural error
4- A plan that could lead to a correct solution with no arithmetic errors

Answering the problem
0 – No answer or wrong answer base upon an inappropriate plan
1 –Copying error, computational error, partial answer for problem with multiple answers. No answer statement, answer labelled incorrectly
2 – Correct solution

(Adapted from Charles et al. (1987); Szetela (1992)

Using a scoring rubric such as outlined in Table 1 to assign points (such as 0, 1, 2 …) illustrates how feedback comments to different components of the assessment provide opportunities for richer assessments and feedback.

4.6   Assessment Objects and Learning Objects

The use of analytical rubrics leads to the creation of educational objects.The concept of learning objects is not new in education. Polsani (2006) defined a Learning Object as independent and self-standing unit of learning content that is predisposed to reuse in multiple instructional contexts. Learning objects are a way of thinking about learning content design, development and delivery. Instead of providing all of the material for an entire course or lesson, a Learning Object only seeks to provide material for a single lesson or lesson-topic within a larger course. This fundamental idea is that instructional designers can build small (relative to the size of an entire course) instructional components that can be reused a number of times in different learning contexts. Examples of LOs include simulations, interactive data sets, quizzes, surveys, annotated texts, adaptive learning modules. Supporting the notion of small, reusable chunks of instructional content, Reigeluth and Nelson (1997) suggest

that when teachers first gain access to instructional materials, they often break the materials down into their constituent parts. They then reassemble these parts in ways that support their individual instructional goals

Wiley (2001) identifies the characteristics of learning objects to be

- Self-contained – each learning object can be consumed independently

- Reusable – a single learning object may potentially be used in multiple contexts for multiple purposes on multiple campuses

- Can be aggregated – learning objects can be grouped into larger collections, allowing for their inclusion within a traditional course structure

- Tagged with metadata – every Learning Object has descriptive information allowing it to be easily found by a search – which facilitates the object being used by others

For the semi-automatic assessment system the interactive nature of the problem-solving as mentioned in Section 4.4 helps to decompose a student's response into several components. In this research, a definition of a parallel/similar concept to learning objects will be made. This will be referred to as reusable "Assessment Object". An Assessment Object may be viewed as a component or section of the student's response which has been assigned marks and/or feedback comment by the teacher within the context of the question. An Atomic Assessment Object then represents the smallest piece of assessment data that correspond to an instructional objective and so may be marked or feedback comment provided on it.

Applying the concept of learning objects, it can be seen that smaller bits (i.e., atomic assessment objects) may be combined into structures that promote one object's combination with a second, and a possible third. This is illustrated in Figure 4.8 where the elements of the students responses are raw data components. When these are combined they form assessment objects which in turn may be combined to form aggregate assemblies and the collections. 0ne factor to consider is "what degree of granularity is the most appropriate for instructionally relevant assessment object combinations?"
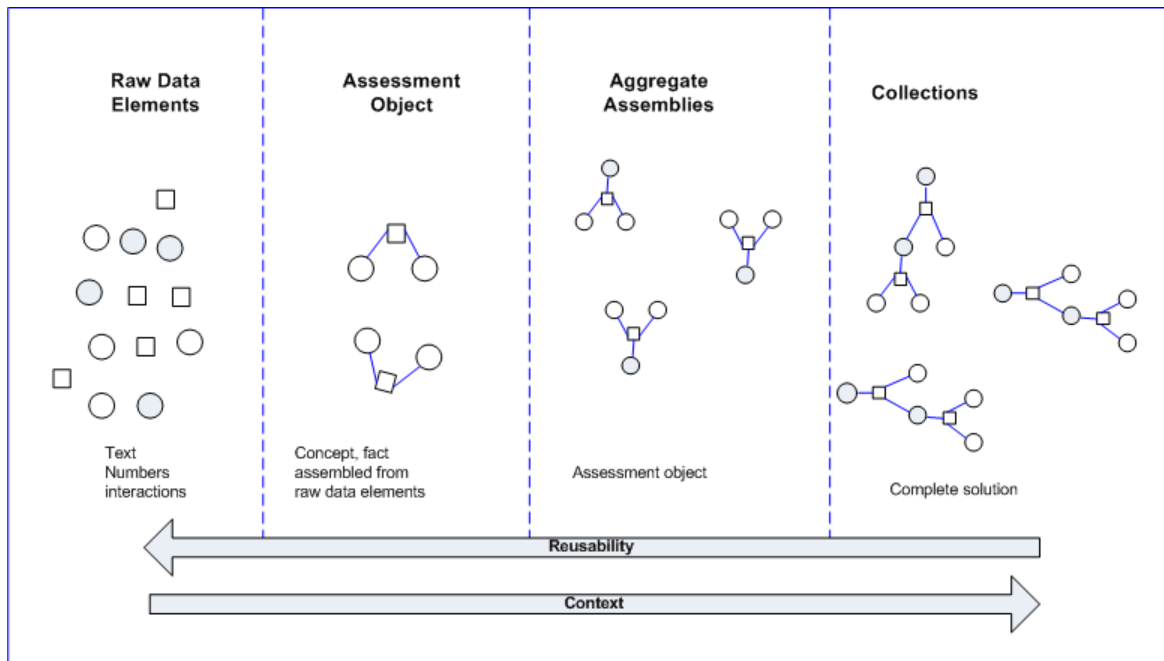
Fig. 4.8 Levels of granularity

As can be seen in Figure 4.8 the smaller the assessment object, the larger its capacity of being reused within and across question context. The larger the combined assessed objects the lower its capacity of being re-used. The application of this concept is presented in 6.5.

4.7 Re-use of teachers' marking experiences

This section addresses the third framework component which concerns how marking and feedback can be made more efficient and consistent

4.7.1 Reusability

Many teachers agree that in marking and feedback activities there are a number of repetitive activities. Repetition easily breeds boredom and may result in inconsistent results. To allow for better use of the teachers effort and time the semi-automatic framework exploits the reuse of already marked components thereby reducing repetitions (see Figure 4.9). By having assessed work stored as assessment objects or cases, reuse is made possible.
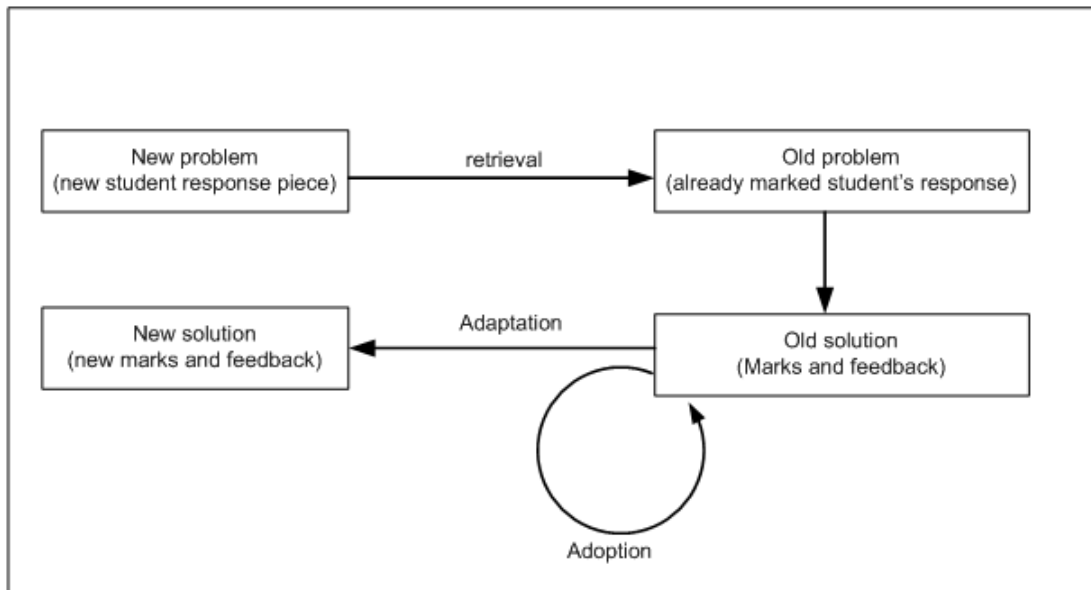
Fig. 4.9 Reuse principle

According to the figure, the process begins with the retrieval of students responses from the pool of students works to be marked. A check is made to see if there are similar submissions in the case library that have been marked. If a match exists, the marks and feedback comments on the already marked work is adopted for the retrieved work. If no exact match is found but a sufficiently similar marked solution exists in the case based, the marks and feedback comments can be adapted.

4.7.2   Utilisation of case-based reasoning

In Section 2.10, case-based reasoning, was described as an approach to developing knowledge-based systems that are able to retrieve and reuse solutions that have worked for similar situations in the past. CBR was chosen above other AI methods because it facilitates reuse and reasoning from past experiences.

A case typically comprises of two parts, a problem part and solution part (Figure 4.10).
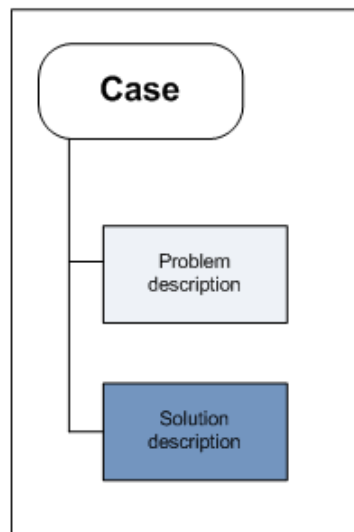
Fig. 4.10 Case components

Problem Part: This part of a case contains information characterising the problem situation occurred in the past. Due to the basic assumption of CBR, it is crucial that this description, in particular, includes information relevant to decide whether two problems are similar or not.

Solution Part: This part contains information used to reproduce the solution applied successfully in the past when being confronted with new problem situations. The solution part can also include additional information such as the way the solution was obtained, the solution's quality, constraints restricting the solutions application and alternative solutions.

4.7.3   Case representations

A CBR system depends heavily on the way its cases are represented (Bergmann et al., 2005). The case representation needs to enable the system to reason with concepts in the problem domain, and facilitate execution of each of the steps in the CBR cycle. Cases are usually divided into problem descriptions and solutions, either explicitly or indirectly as part of the same case structure. Richter and Rosina (2013) states that representations vary from simple feature-value vectors, to complex data-structures specially tailored for a domain. The major representations in CBR are:

- Flat attribute-value representation

- Object oriented representation

- Trees and graphs

- Hierarchies and taxonomies

- Generalized (set-oriented) cases

The attribute value based representation is known to be flexible and the most commonly used approach to representing case knowledge in a structured way. The basic idea of this approach is the so-called local-global principle which states that the entire similarity assessment can be subdivided into a set of local parts and a global part. The local parts reflect the contribution of individual attributes to the entire similarity assessment. Therefore, for each attribute a so-called local similarity measure has to be defined. The global part is responsible for estimating the actual utility of complete cases. It is based on the outcome of the local similarity measures and so-called attribute weights that reflect the relative importance of the individual attributes.

4.7.4 Distance and similarity measurements

Similarity and distance measurements look at the same objects from different points of view: similarity is used to express how close objects are and distance reveal how far apart they are. These two measures are described below:

I **Distance**

Distance metrics are used to check how far different cases are to each other. The distance is a quantitative variable that usually satisfies the following three conditions (Kuhn and Johnson, 2013):

    i $d_{ij} \geq 0$ - always positive or zero

    ii $d_{ij} = 0$ - zero if and only if it measure to itself

    iii $d_{ij} = d_{ji}$ - is symmetrical

The distance measurement between two case is calculated based on the several feature variables of the cases. The distance is calculated for each feature variable and aggregated together into a single similarity index between the two objects (Harrington, 2012). A list of common distance measurements is presented in Table 4.3.

Table 4.3 Distance metrics

Distance measure	Description	Formula
Euclidean	The most common distance metric. It is simply the root of squares differences between coordinates or a pair of objects	$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ij} - x_{ij})^2}$
City Block or Manhattan	Computes the absolute value distance. It examines the absolute difference between coordinates of a pair of objects	$d_{ij} = \sum_{k=1}^{n} |(x_{ik} - x_{jk}|$
Chebyshev or Maximum	Examines absolute magnitude of the differences between a pair of objects. This can be used for both ordinal and quantitative variables	$d_{ij} = max_k|(x_{ik} - x_{jk}|$

In the table, the distance of two objects or cases is expressed through a linear contrast of weighted differences, which exist between their common and varying attributes. For real-valued data, the Euclidian distance is the most widely used (Harrington, 2012). For other types of data such as categorical or binary data, the Hamming distance can be used (Norouzi et al., 2012). In the case of regression problems, the average of the predicted attribute may be returned.

II **Similarity**

A similarity measures assigns a numerical value to the case, expressing its degree of similarity with the given query, inducing a partial ordering on the set of problem descriptions and, consequently, also on the case base (Hüllermeier and Minor, 2015). These measures are often based on a geometrical interpretation, where the cases are considered in an n-dimensional space. Each dimension corresponds to one attribute the problem description of the cases. The concept of similarity consists of two parts (contents-based similarity) and similarity between concepts (structure-based similarity). The contents-based similarity of a concept is computed on the attributes defined in the concept. The structure-based similarity defines similarities between concepts independent of their contents. Inside a concept hierarchy, the similarity of concepts to each other may be

explicitly or implicitly defined by using a taxonomic view of the hierarchy (Hüllermeier and Minor, 2015).

Let *CBP* denote a set of input descriptions *P* a query for which a solution *Q* exists such that (P, Q) is in the case base *CB*. A similarity measure is a mapping (Richter and Rosina, 2013).

To reduce arbitrariness, some assumptions are made (Harrington, 2012):

i $0 \leq sim(x, y) \leq 1$ (Normalisation)

ii $sim(x, y) = 1$ (each object is itself the nearest neighbor)

iii $sim(x, y) = sim(y, x)$(symmetric property)

Where *sim* represents the similarity measure which approximates the usefulness of the solutions for the query.

A general similarity measure is defined as function

$$similarity \quad p \times q \longrightarrow [0, 1] \tag{4.1}$$

In general, two retrieval techniques are used by the major CBR applications: nearest neighbour retrieval algorithm and inductive retrieval algorithm. According to Watson (1998) and Kolodner (1993), the nearest neighbour retrieval technique is simple but has slow retrieval speeds when the case base is large. Inductive retrieval, on the other hand, is fast but depends on pre-indexing which is time-consuming, and it is impossible to retrieve a case while data is missing or unknown. The choice between nearest-neighbour retrieval and inductive retrieval in CBR applications requires experience and experimentation. Usually, using nearest-neighbour retrieval without any pre-indexing is a good choice (Watson, 1998). If retrieval time becomes an important issue, inductive retrieval is preferable. The simplicity and multi- feature calculation capabilities of the nearest neighbour techniques makes it suitable for the present study.

4.7.5   The Nearest Neighbour

The nearest neighbour algorithm stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions) between the new cases in the case library. It has long been used in statistical estimation and pattern recognition systems. The *K*-Nearest Neighbour algorithm is similar to the Nearest Neighbour algorithm, here a new case is classified by a majority vote of its neighbours, with the new case being assigned to the

most common amongst its K nearest neighbours measured by a distance function such as Euclidean distance. (K is usually assigned an odd number to avoid ties in determining the nearest neighbour). Generally, where K is large (more precise, it reduces the overall noise). Historically, the optimal K for most datasets has been 3-10. $K = 1$ will be the same as nearest neighbour, as it only looks at the 1st closest neighbour.

The following presents descriptions and formal expressions for the nearest neighbour CBR system for objects p,q with attribute-value vectors (Leake, 1996; Riesbeck and Schank, 2013; Watson, 1998).

$$p = (p_1...., p_n) \quad and \quad q = (q_1...., q_n): \tag{4.2}$$

$$Attributes : Attribute_1, Attribute_2, ...Attribute_n. \tag{4.3}$$

The similarities between values of attributes are the local similarities:

$$similarity_1(p_1, q_1), similarity_2(p_2.q_2)...., similarity_n(p_n, q_n); \tag{4.4}$$

The overall (global) similarity is:

$$similarity(p, q) = \sum_{i=1}^{n} f(p_i, q_i) \times w_i \tag{4.5}$$

Where

- P = the query case

- q = source case

- n = the number of attributes in each case

- i = an individual attribute from 1 to n

- f = a similarity function for attribute "i" in cases p and q

- w = the importance weighting of attribute i

The equation represents a typical nearest-neighbour technique that describes a situation for which p and q are two cases compared for similarity, n is the number of attributes in each case, i is an individual attribute from 1 to n, and w_i is the feature weight of attribute i.

Generally, the similarity calculation continues until all cases in the case library have been compared, and ranked according to their similarity to a target problem case.

As mentioned earlier similarities are usually normalised to fall within the range 0 to 1, where 1 means a perfect match and 0 indicates a total mismatch.

Figure 4.11 illustrates the nearest-neighbour similarity comparison for two feature attributes.
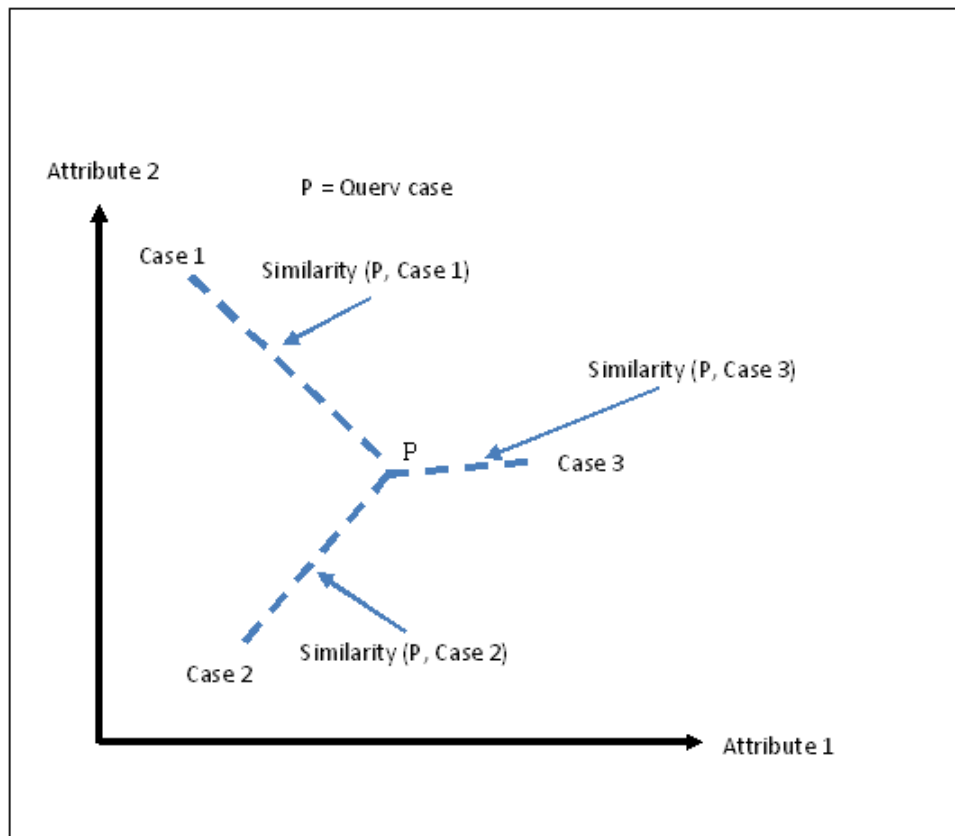
Fig. 4.11 Nearest neighbour discovery of the new case NC (adapted from Watson (1998))

The figure shows a query case q for which similarity comparisons are made with three other cases in a 2-feature attribute dimensional space. The objective is to identify the nearest neighbour. In this 2-dimensional space, case3 is selected as the nearest neighbour because similarity(p, case3)> similarity(p, case1) and similarity(p, case3)> similarity(p, case2).

4.7.6 Example of an attribute-value representation with the nearest neighbour algorithm

As a way to illustrate the nearest neighbour concept for assessment of a mathematics problem, Table 4.4 shows an example of four cases with nine attributes.

Table 4.4 Example cases of students solutions

Case	X1	X2	X3	X4	X5	X6	X7	X8	X9
Case1	54	+	13	67	36	-	31	3	15
Case2	13	-	36	24	54	+	98	12	42
Case3	54	+	13	68	36	-	32	5	20
Case4	36	-	13	23	54	+	777	7	29

X1 represents operand 1, X2 represents operator 1, X3 represents operand 2, X4 represents the intermediate answer, X5 represents operand 3, X6 represents operator2, X7 represents the final answer, X8 represents the task time in minutes and X9 represents the count of interactions. Each case represents the answers given by a student to a hypothetical question. The attributes are features of students answers such as operators, operands and calculated results entered, and so on. An answer case is selected (e.g., case1), and marked with appropriate feedback comments given. Similarity comparisons are then made with other cases in the case library so that the marks and comments may be reused.

Table 4.5 Result of similarity calculations

Case	Euclidean distance	Similarity index
Case 1	0	1
Case 2	98.41	0.27
Case 3	05.57	0.92
Case 4	70.10	0.52

As can be seen in Table 4.5, Case 2 has the largest distance (98.12) and smallest similarity value (0.27) and so is very dissimilar to Case 1. Case 3 with the highest similarity value (0.92) is very similar to Case 1. Case 1 is an exact match with itself (0 distance and similarity value of 1). In the pool of student answers, items that are exact matches to Case 1 can automatically be selected and marks with the comments already attached to Case 1 reused. Cases that are reasonably similar can also be retrieved, and the solutions (i.e. marks and comments from Case 1) adapted to suit the respective cases better. This way consistency and reduction of repetitive marking is achieved.

The approach described produces a semi-automatic system where the human makes judgements on marks and provides comments, while the computer system computes and attempts to reuse the marks and feedback in similar cases.

4.8   Chapter summary

This chapter has presented the semi-automatic assessment framework which provides understanding how to create interactive assessments that that help to capture students' solution processes and supports teachers in marking and feedback activities. The framework is guided by three underlying principles: design problem solving in an interactive environment, clearly defined analytic rubrics and re-use of marking and feedback work with case-based reasoning. An important concept that emerged from the framework is a repository of assessment objects which facilitates the reuse of marking and feedback activities. An arithmetic word problem example was used to illustrated the components of the framework.

Chapter 5

Design and Development of MuTAT

The previous chapters form the basis for the design and development part of the thesis. They established the need to have CAA systems that allow for solution process evaluation. With new advancements in computer technology the possibilities of supporting students in entering their responses is increased. Chapter 4 situates the capture of student solution responses in the semi-automatic framework and identifies it as a key element in the framework. It also outlined the approach adopted for the research. The development of the approach using a prototype tool will help to demonstrate the feasibility of the design and describe the functionalities of the system. The process of implementation can provide insights into the advantages and disadvantages of the concepts and the chosen design alternatives. This chapter discusses the requirements, design, development and demonstration of the tool referred to in this work as Multi-Touch Arithmetic Tool (MuTAT). The tool focuses on primary mathematics and mainly addresses addition, subtraction, division and multiplication word problems.

5.1   General requirements

Effectively capturing problem-solving steps requires a tool that is educationally justified. It should follow sound pedagogic principles and therefore contribute to learning. Also, the tool needs to provide an environment that freely allows creative problem solving without introducing extraneous cognitive load. Chandler and Sweller (1991) have argued that inefficient and complex interactions in designs add unnecessary load to the working memory during problem solving and so should be avoided. Also, Jonassen (1999) and Savery and Duffy (1996) have suggested principles for designing effective learning environments. Their

suggestions include making interfaces intuitive, flexible and structured such that students can construct meaning based on their conceptual understandings. These general requirements serve as a guide for the implementation of the tool.

5.2   Design goals and functional overview

The purpose of the prototype tool is to provide a hands-on assessment environment that can engage students in problem solving, provide opportunities to capture the solution steps with limited restrictions as often encountered in fully constrained assessment environments. The solution steps captured should take place unobtrusively without interrupting or distracting the problem solver. Data obtained from the prototype tool should be useful for process-level type feedback.

There were five design goals:

1. provide a highly interactive environment that engages students in problem-solving

2. allow the breaking down of the problem-solving effort into small steps

3. capture all the traces of problem-solving

4. present a visual representation of the solution steps

5. provide an easy-to-use application that accommodates users of varying computer experience

An the end, the output from the tool should provide information from which inferences about the problem solver's thinking may be deduced.

5.3   Devices and technology choices

Before the coding and implementation of the tools begin, the following choices need to be made: type of device, web app or native app, operating system type and type of interactions.

5.3.1   Devices

Laptops and desktop computers are widely used in educational settings. Students typically use a mouse and keyboard as inputs; trackpads are also common. Since 2007 when Apple

launched the iPhone, touchscreen devices (smartphones, tablets and phablets) have become common (Allen et al., 2010). They provide gestural interactions not readily available with other input methods. They can give more freedom in the design with less reliance on visible buttons or other user interface elements. The tablet form factor has been chosen as the primary target device because of its movability and good screen size. As mentioned in Section 4.4.3, gestures are becoming a major way of interacting with computing devices.

5.3.2 Web and Native Apps

There are two primary ways applications are delivered to devices: as a web app or website or a native app for a specific platform such as iOS or Android (Chien et al., 2014).

I **Web app**

A web app usually has a high level of interactivity and allows users to complete some task. Irrespective of the device, access to the applications content is through the web browser. Since almost all devices have a browser, users can access the web app without having to install anything new (Charland and Leroux, 2011). A web app relies on a single code base; it makes updates much faster and cheaper. Also, the web is an open platform; this means anyone can access it without any proprietary software or operating system. This openness ensures that the application will be available in the future than a native application. However, web apps usually developed using HTML5 and CSS may suffer from slow performance. This is especially the case when the web app makes use of images, audio, and high interactivity. Also, in most cases, it is not possible to use the web app in an offline mode since the web requires an internet connection. A web app is constrained by the capabilities of the browser, this means that for touch interfaces on smartphones and tablets, only basic gestures like tapping to follow a link or swiping between images in a slideshow can be used. Other gestures do not work well in the browser or may be reserved for navigating the browser, itself. The other drawback to a browser-based web app is that mobile browsers only have limited or partial access to a device's capabilities like the GPS, accelerometer, or notification system.

II **Native app**

A native app, on the other hand, is one that is built specifically for one operating system, such as IOS or Android for mobile devices, and Windows or Mac OS for desktop computers (Allen et al., 2010). These are self-contained applications that do not require a web browser, but they can only be used on the operating system for which they were built. These are apps that are downloaded or installed directly onto a device, and they

will only work on that particular device's operating system. The main advantage of a native app is the potential for much faster performance. Because the app is already installed on the computer or device, they only have to download data when needed. Another advantage to having an app already installed on a device, is that it gives the ability to work offline, when there is no internet connection. Native apps also provide the ability to use more gestural interactions on touch screen devices (Charland and Leroux, 2011). This is very important for this research as it provides the opportunity to implement and customise custom gestures. The biggest disadvantage of native apps is the cost and time investment required to create them. In many cases, native applications have to be developed for the major platforms and screen sizes to provide wider access.

Unlike the web, it is not possible to publish a solution directly to end users. Once a native app is built, almost all operating systems require the app to go through a submission and review process before it is ready for full use and available in the app marketplace. Often the rules are not always clear. Each platform has its patterns for typical uses of different touch gestures. This often makes the development painful and time-consuming.

For the MuTAT application, the choice of a native app was made so as to have very good performance in the implementation of custom gestures and the ability to work offline when there is not internet connection.

5.3.3   Native apps: iOS, Android, or Windows Phone?

An important decision is on which platform the development should be done. For smartphones and tablets, there are four major players: Google's Android, Apple's iOS, Blackberry and Microsoft's Windows Phone (Grønli et al., 2014). Amazon has also released their Fire Phone, but they are yet to become mainstream. At the end of 2013, Android led with almost 61% of the market share, followed fairly closely by IOS with just over 36%. The next players are much farther behind, with Windows Phone at just 2% (gartner.com, 2014). Android and Apple are the major market leaders. Because Apple builds its hardware, iOS has relatively few devices. However, this makes it much easier to design for and test. Android, on the other hand, has hundreds of different devices on the market, all with varying screen sizes and resolutions.

The present study does not aim at immediately producing commercial software available on the major and best platforms but rather focuses on providing a tool for testing the concepts of the semi-automatic framework on real student users. Getting a tool out quickly on the tablet

first gives important feedback and data from real users. The iPad was chosen as the hardware platform for the application because of its smooth implementation of multi-touch functions and the increasing availability of tablet types devices in many schools (UPI.com, 2013). The application was developed with Apple's native programming language, Objective-C, using the iOS development frameworks.

5.3.4 Interactive touch gestures

Gestures and Interactions are the basis of touch device designs (Shneiderman, 2010). The most common gestures are tap, double-tap, swipe and pinch gestures. The tap is the simplest of interactions and the most commonly used. The tap is used to select an object or advance in the navigation hierarchy. A student can double-tap a section of content to zoom in on that area of the screen. It is also sometimes used to open something, such as an application on a device. The double-tap is not always immediately discoverable since there are almost never visual cues to indicate that the action is possible. The swipe gesture occurs where the finger slides across a particular item on the screen. Doing this will often reveal options for interacting with that item, or it can be used to delete the item from a list. This is widely used in lists on the iOS and Android platforms. Swipe is also used to move forward and backwards through a series of photos or screens.

The drag sometimes called pan gesture occurs when the surface of the screen is touch with the finger or thumb and then moving the finger or thumb in some direction, without releasing it. As this occurs, the content that was touched is dragged around with the finger. The long press gesture also called touch and hold gesture occurs when an object on the screen is selected and held for a few seconds before being released. It is used primarily to trigger more options to appear for whatever is being targeted. It can give access to actions such as delete, edit and share. On all platforms, it can be used to enter an editing mode for an item or screen, such as the home screen. Usually, users need to learn what that tap and hold does before it is understood. The pinch and spread gestures are very common interactions and are usually used to zoom out and zoom in when looking at a website, a photo or some other type of content. The different interactions and subtle animations will be implemented in the MuTAT to provide a rich variety of interactions.

5.3.5 Design considerations for touch-based applications

With the new interfaces emerging, additional considerations are needed for touch devices. A mouse cursor is considered a fine input because it can easily point to and click on smaller target areas. A finger, or thumb, although more direct and natural are not nearly as accurate at targeting a discrete point on a screen and so are considered coarse inputs. Designing for touch means resizing and respacing interactive elements to allow fingers and thumbs to hit targets accurately. For a user to easily activate elements on a touch screen, user interface elements are designed with larger touch targets. It also requires adding enough space around the links and buttons, so that the user does not accidentally tap on the wrong element. Also, font sizes are designed big enough to be selected. Apple recommends a minimum touch target of 44 points by 44 points in the iOS human interface guidelines (Apple Inc., 2015). Having a good amount of space between interactive elements and having them all be large enough that they are easily targeted will provide good user experience.

5.4 Development

Regarding the first design goal which centred on providing a highly interactive environment, a multi-touch capable environment was selected for the implementation. The justifications for an interactive environment and multi-touch gestures were broadly presented in 4.4. The gains, as stated of an interactive multi-touch environment include intuitiveness, ability to engage students and the recent popularity of the devices.

The second design goal aims at providing the ability to build arithmetic expressions with operators and operands. This is made possible by having the operands dragged down from the question context as described in Section 4.3. A contextualised menu pad is used to provide access to the arithmetic operator symbols. The third goal seeks to have captured all the student interactions, decisions and processes. The use of log files helps to achieve this. For each solution effort, key elements of problem-solving, such as time-stamps for each user-interface event and interaction, operators and operand selected, position and placement of the operands, cancellations, deletions and choice of operators are all recorded. The fourth objective is to provide a visual representation of solution steps. This is made possible by the use of user-interface shapes, views and symbols which are integrated with the major problem-solving elements such as the operators, operands and answer responses. These are visually presented in a summary screen detailing connections, relationships and sequence.

The last design goal looks to provide an easy-to-use application that accommodates users of varying computer experience. This non-functional objective aims at removing barriers the technology may impose on users if cluttered by too many (and intimidating) screen menu elements. The barriers when present shifts focus from the task at hand to understanding and using the technology.

5.5 MuTAT Architecture

The MuTAT was designed to be used with only touch gestures. Simple designs provide both aesthetic and functional benefits (Mullet and Sano, 1994). Concepts are added to the solution space by touching and dragging the concept icons in an open screen.

5.5.1 User Interface

The user interface is a crucial part of the design of the assessment environment. The MuTAT prototype assessment environment is based on multi-touch technology described in Section 2.5. The tool provides support for four arithmetic operators (addition, subtraction, multiplication and division) and use of digits 0-9. It can support questions of varying complexities.

The user interface has a simple and smooth design (see Figure 5.1) following Savery and Duffy's (1996) suggestion for free and unrestricted learning environment. The workspace goes edge to edge on the screen with very few elements on top and it does not display menus. Touch gestures are used to interact with elements on the screen and navigate to the different parts of the tool. The workspace is divided into two, the problem pane and the solution pane. The problem pane displays specific and relevant word problem tasks suitable for students while the solution pane is used to display constructed responses made by the student.
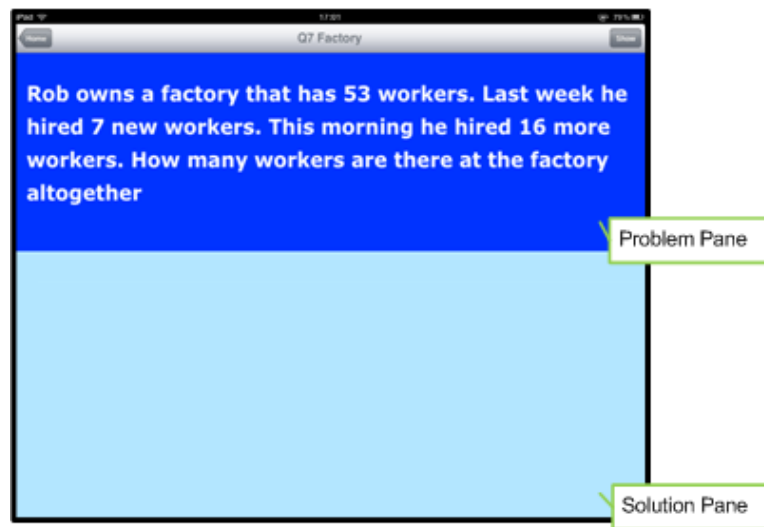
Fig. 5.1 MuTAT user interface

Both panes are placed together on the workspace to allow movement of constructed responses between the problem and solution panes. This arrangement is similar in design to that used by Stone et al. (2009) in an Entity Relationship (ER) rationale design system. The layout helps the student to focus on problem details without having to switch back and forth between screen views.

5.5.2 Problem-solving process

The problem pane contains word problems with draggable numeric values. The pane provides scaffolding for students because they will not have to write or type the numbers that are listed in the question. One or more numbers/operands can be dragged from the problem pane down to the solution pane with one or both hands (Figure 5.2). The pan (drag) gesture is used for this stage. The operands may be dragged around the drop area.
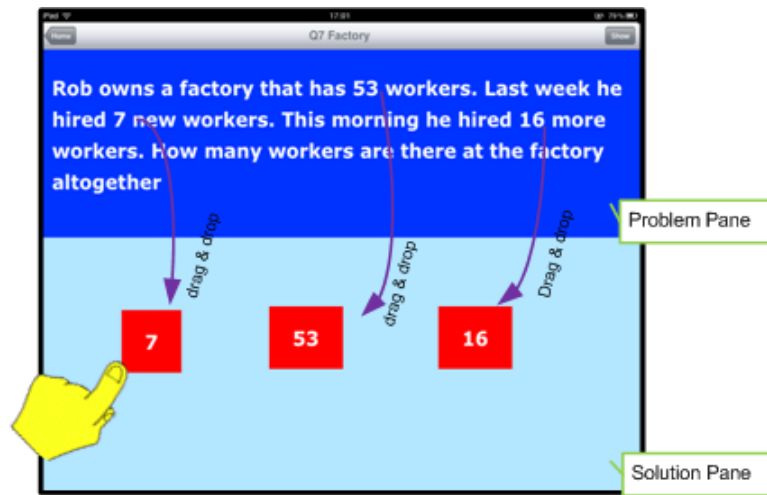
Fig. 5.2 Drag and drop interactions with question elements

To perform an arithmetic action, a multi-touch/multi-finger gesture is used. A user touches and holds two chosen numbers simultaneously (Figure 5.3). The action of touching and holding an interface element for a few second is referred to in the literature as touch- and-hold gesture/long press gesture Bailly et al. (2012); Wigdor and Wixon (2011). This action in the MuTAT environment activates a menu containing four arithmetic symbols (+ - ÷ ×) for selection. This is shown in Figure 5.3.
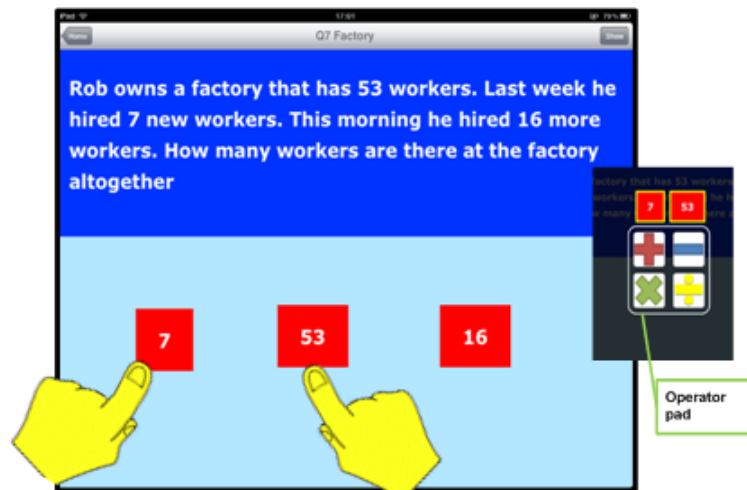
Fig. 5.3 Multi-touch interactions to create arithmetic expression

The student makes a decision on which operator to use for the calculation and selects the one deemed appropriate for the task. The action of selecting an operator brings up a numeric pad from which digits of calculated results are entered. This action is shown in Figure 5.4
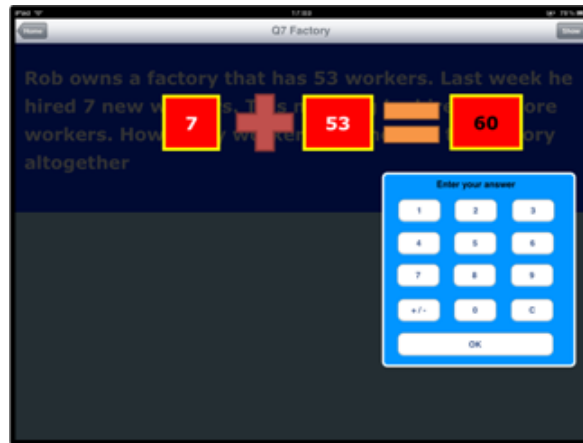
Fig. 5.4 Enter results with numeric keypad

This concludes the first step in the problem-solving process. For multi-stage arithmetic problems, the product (answer) of the first step becomes an intermediate result for subsequent steps. This is illustrated in Figure 5.5. The process of selecting operands and operators may be repeated in subsequent steps. In this way, intermediate and final answers are made more explicit.
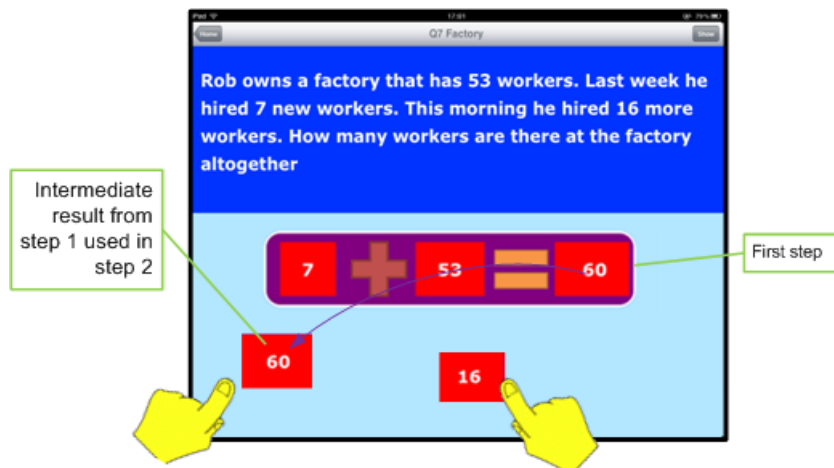
Fig. 5.5 Using intermediate result in other solution steps

Figure 5.6 shows the arithmetic expressions for the two steps involved in the solution process. It provides information on the steps, intermediate and final answers.
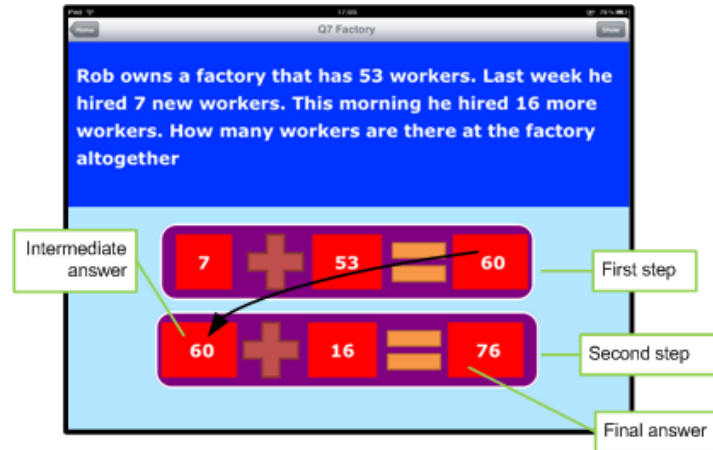
Fig. 5.6 Details of solution process steps

Further details on the solution process can be assessed in another view made available by the selection of the *show* icon on top of the screen. This is shown in Figure 5.7, provides a visual representation of the problem-solving steps. The figure shows how the numbers have been paired. All entered responses can be viewed. As shown in Figure 5.7 the simple two-step problem has four gradable items (correct use of the arithmetic operators and correct computations).
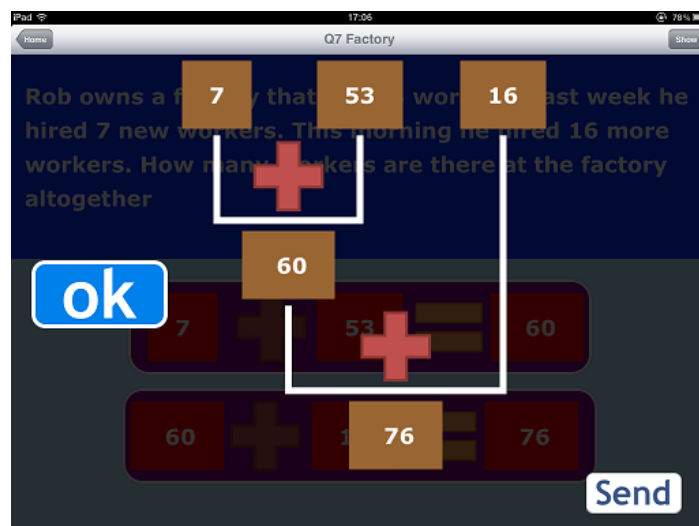
Fig. 5.7 Tree representation of solution process

The process and sequence in which the problem was tackled may be deduced. Apart from the final answer, the intermediate answer and the two correct choices of the arithmetic operator can be scored separately. Furthermore, the way the numbers are paired may be used to investigate strategies and understanding of number concepts.

5.6   Programming of the MuTAT

There is no direct access to any of the hardware of the iPad. Therefore, all hardware interaction take place exclusively through the layers of software which act as intermediaries between the application code and device. As mentioned in Section 5.4, Apple's iOS was the platform used for the development.

5.6.1   The four pillars of iOS development

Developing an application on the Apple iOS platform rests on four core pillars (Allan, 2013; Nahavandipoor, 2012; Neuburg, 2012). This is illustrated in Figure 5.8.
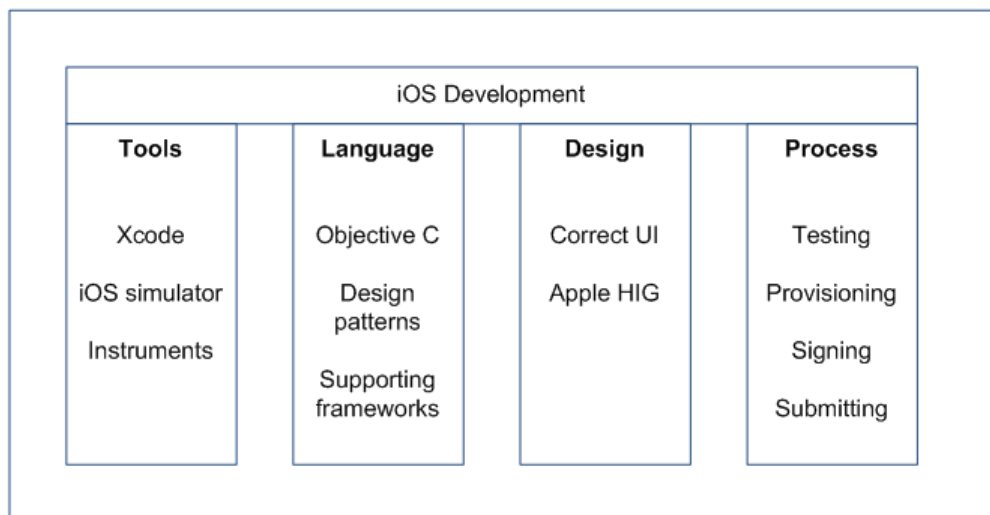
Fig. 5.8 The four pillars of the iOS development

The first pillar is the tools used which includes the Xcode IDE, iOS simulator and instruments. The second pillar is the Objective-C native programming language for iOS devices which includes the iOS development principles like protocols and delegation, and the supporting

frameworks and all the prewritten code Apple have provided to make iOS app development easier. The third pillar is the overall design of the application showing how the application functionality and data is broken down across multiple screens, and the flow through the screens. Also important here is the choice of the right user interface elements in line with standard Human Interface Guidelines. The fourth pillar is adhering to the rules and processes set by Apple for development. This includes the provisioning, signing and testing and submitting. These are all essential for successful development of the application.

5.6.2   iOS Development Architecture

The iOS consists of a number of different software layers, each of which provides programming frameworks for the development of applications that run on top of the underlying hardware. These operating system layers is presented diagrammatically as illustrated in Figure 5.9:
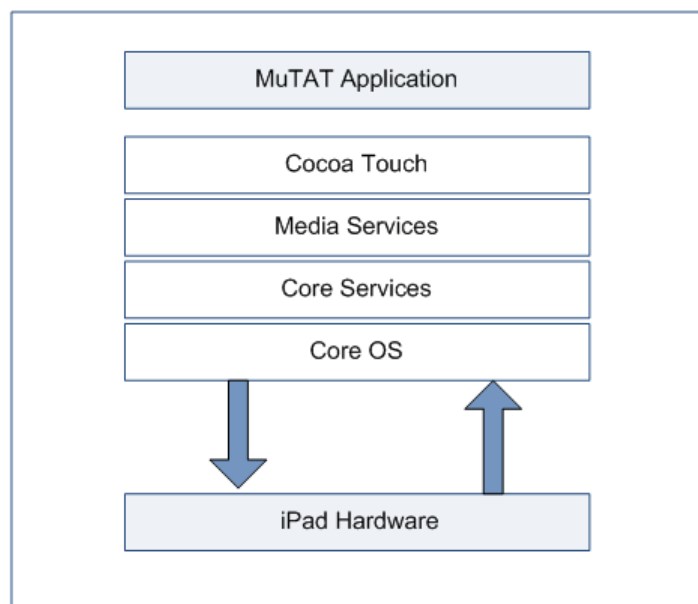
Fig. 5.9 iOS Architecture

The MuTAT application makes direct calls down any of the layers of the stack to perform tasks on the iPad. The Cocoa Touch layer sits at the top of the iOS stack and contains the frameworks for user-interface creation and management (text fields, buttons, labels, colours, fonts, etc.). The role of the Media layer is to provide iOS with audio, video, animation and graphics capabilities. As with the other layers comprising the iOS stack, the Media layer

comprises a number of frameworks which are utilised when developing the application. The iOS Core Services layer provides much of the foundation on which the previously referenced layers are built and includes frameworks for network and data management. The Foundation framework is the standard Objective-C framework which consists of Objective-C wrappers around much of the C-based Core Foundation Framework. Development of native iOS application requires strict compliance with standard design patterns.

5.6.3 The Model-View-Controller design pattern

Model-View-Controller design (MVC) refers to a high-level design pattern that concerns itself with the global architecture of an application. This design pattern was chosen for the MuTAT programming because it allows the application objects to be more reusable, it removes duplication, and allows applications to scale without having to re-write major portions of the system. Also, the Cocoa framework is based on MVC. Within this design pattern, objects are classified by the type of role that they play within the system. There are three role types that an object can fall into, model, view, or controller. MVC separates all of the displays and interface code from the programme logic and processing functions. The separation of these roles also defines how the three communicate with one another. This is illustrated in Figure 5.10.
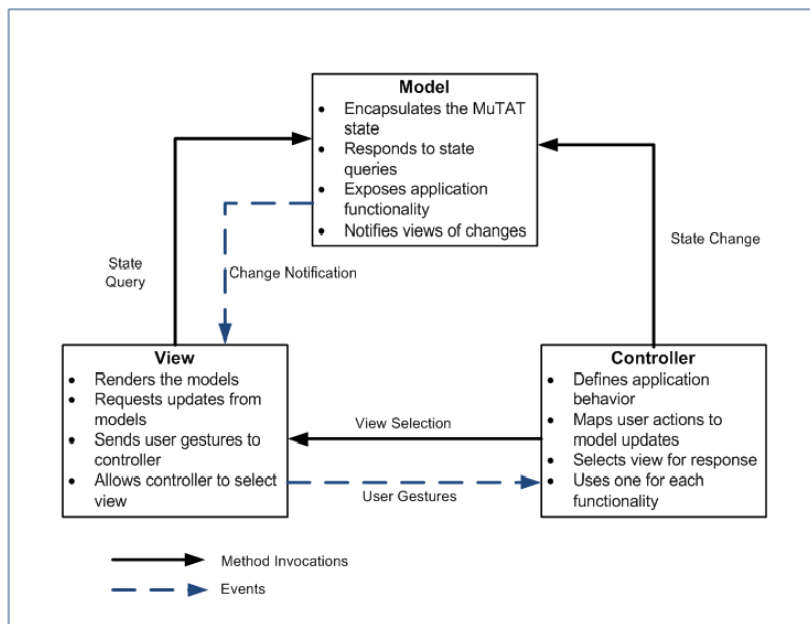
Fig. 5.10 Model-View-Controller (MVC)

- **Model** – Holds all logic and core data. This includes the variables, detailed functions, and number crunching. This layer is completely detached from the views so that the views could easily be changed and the same data working.

- **View** – Is the display style of the application. These are the user-interface elements that the students interact with. They include the question and solution workspaces, the table list displaying the questions, profile page and visual representation page. All these may be changed the different styles used while still working with the same data model.

- **Controller** – Acts as an intermediary between the other two. Objects in the view are connected to the View Controller, which passes the information to and from the Model. So in this way, it is possible to have a student drag on a number, and it is registered in the model.

5.6.4 Making the MuTAT question elements intractable

The multi-touch interface of iOS devices allows a user to interact with a device using one or more fingers. For an iOS device to detect gestures, a gesture recognizer must be attached to the view the user is touching. Delivery of events follows from the operating system to the application object then to the window object representing the window in which the touches are occurring. This path is illustrated in Figure 5.11.
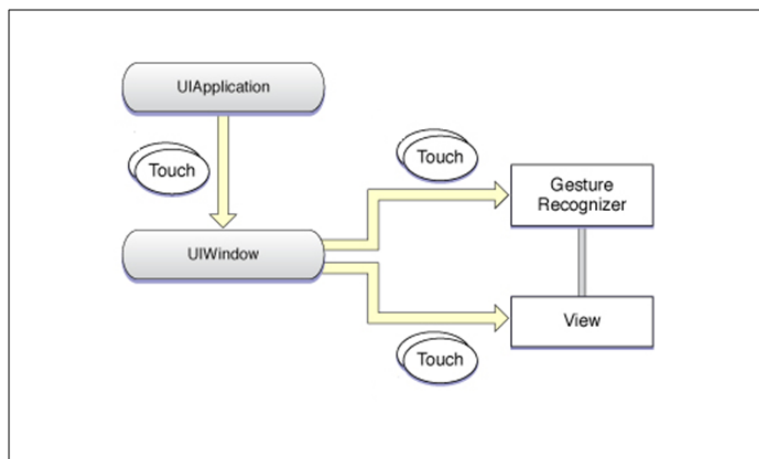
Fig. 5.11 The touch paths

The multi-touch gestures are a combination of touch events which is achieved by adding a Gesture Recognizer to a user-interface view. The following listing shows how this is achieved in code.

```
#pragma mark − multi touch
− ( void ) addGestureRecognizersToShape : ( UIView ∗ ) piece
{
    UIPanGestureRecognizer ∗newPanGestureRecognizer = [[
        UIPanGestureRecognizer alloc ] initWithTarget : self
        action : @selector ( panDetected : ) ];
    [ newPanGestureRecognizer setDelegate : self ];
    newPanGestureRecognizer . maximumNumberOfTouches = 2;
    [ piece addGestureRecognizer : newPanGestureRecognizer ];

    UILongPressGestureRecognizer ∗
        newLongPressGestureRecognizer =
    [[ UILongPressGestureRecognizer alloc ]
      initWithTarget : self  action : @selector ( longPressDetected
          :) ];
    [ newLongPressGestureRecognizer setDelegate : self ];
    [ piece addGestureRecognizer :
        newLongPressGestureRecognizer ];
    newLongPressGestureRecognizer . numberOfTouchesRequired =
        1;
    newLongPressGestureRecognizer . allowableMovement = 150.0 f
        ;
    newLongPressGestureRecognizer . minimumPressDuration =
        0.8;
}
```

This shows how the pan (drag) and long-press gesture recognizers being added to solution pieces to make them active and usable for problem solving in the MuTAT. Code snippets that show various implementations of the design on the MuTAT can be found in Appendix F.

5.7 Chapter summary

In this chapter, the design decisions regarding the software tool MuTAT were outlined. The requirements are based on research literature, and the research approach presented in the previous chapters. Within the chapter, the implementation of the conceptual design, as it relates to the solution steps capture part of the framework, was illustrated using a practical example. The main focus of the next chapter will be on the design and development of the Marking Assistant tool to assist in grading students' work. The design and development chapters provide a base for empirical studies on usabilities and the functionalities of the approach which can only be performed after the system has been designed.

Chapter 6

Design and Development of the Marking Assistant

Chapter 5 described the design and development of the MuTAT prototype which captures and stores students' solution processes. The next step in the framework is the marking and production of feedback comments on the submitted work. Section 4.7 described the principles adopted for the marking of students' submitted work. Implementation of these principles requires a software tool. This chapter describes the design and development of the research prototype tool for marking, called Marking Assistant (MA). The MA aims to support teachers in the marking and feedback activities with intelligence techniques so that the problems of marking inconsistencies and time intensive repetitive activities are reduced. The basic approach, requirements, design goals/considerations, development and demonstration of the tool is discussed in the following sections.

6.1   The marking approach

The background for the marking software was described in Section 4.7. The MuTAT helps to provide information on the problem-solving steps for "intermediate constrained" question types. Usually, to support students' learning, teachers are required to mark and provide detailed feedback comments on submitted work using preferably an analytic rubric that aids the examination of the work in detail. The marking and feedback load can be high, especially when there are large numbers of students. Also, inconsistencies may occur due to human elements of fatigue and lethargy that may result from the drudgery of the marking. The idea

is to have a support tool that reduces repetition in the marking exercise. When the teacher makes a judgement on the segment(s) of submitted response it is stored as cases in a case library and reused for similar submissions made by other students. The Marking Assistant provides this support and can help to identify sufficiently close submissions for the cases marked so that the marks and feedback may be adapted and reused. Figure 6.1 illustrates this basic idea.
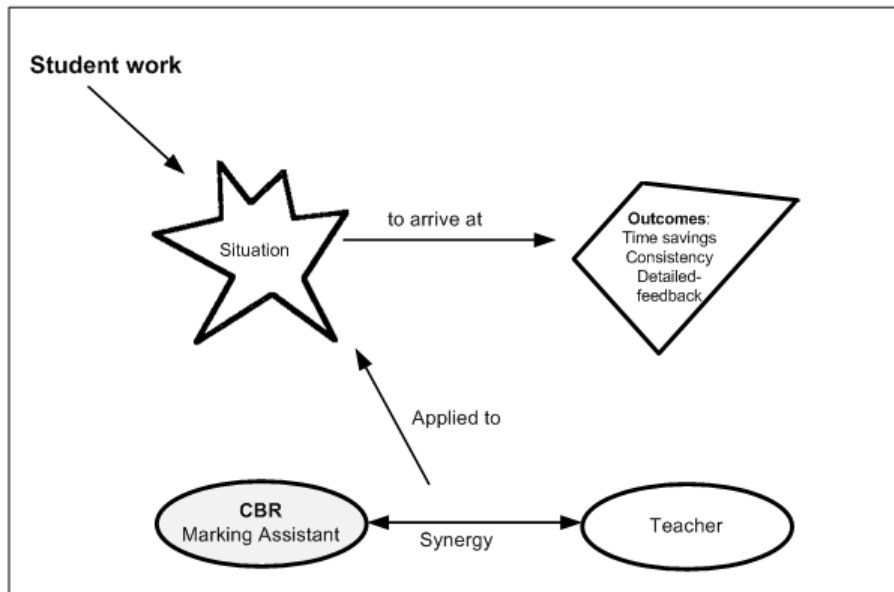
Fig. 6.1 CBR Marking Assistant scheme

The idea is essentially an active human-computer collaboration approach to marking and feedback.

6.2 Design requirements

There are three main functional design requirements for the MA. These are to:

- provide a user interface to apply an analytic marking rubric

- reuse marks and feedback comments to similar occurrences using case-based reasoning that implements the nearest neighbour retrieval algorithm

- provide summary reports

6.3 Web applications development architecture

Section 5.3.2 contrasted the options of web applications and native apps for application development. For the Marking Assistant there was no need to implement custom interactions; therefore, the web apps option was chosen for the development. This section presents key attributes for the web application architecture used for the Marking Assistant. This includes the layered structure; guidelines for performance, security, and deployment; and the key patterns and technology considerations (MSDN, 2016).

A web application is an application that can be accessed by the users through a Web browser or a specialised user agent (Charland and Leroux, 2011). The browser creates HTTP requests for URL that map to resources on the Web server. The server renders and returns HTML pages to the client, which the browser can display. The core of a web application is its server-side logic. The application utilises a three-layered architecture which comprises of presentation, business, and data layers. Figure 6.2 illustrates a typical web application architecture with common components grouped by different areas of concern.
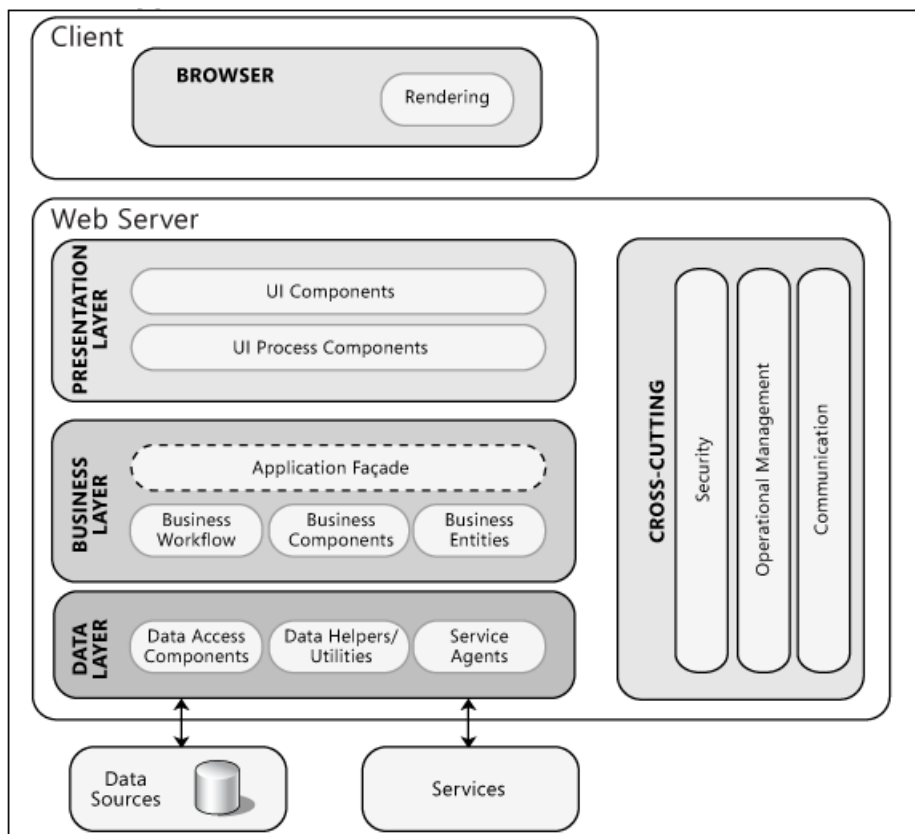
Fig. 6.2 Web applications architecture (MSDN, 2016)

From the figure, the presentation layer contains the user oriented functionality responsible for managing user interaction with the system. The layer includes the user interface and presentation logic components. The business layer implements the core functionality of the system and encapsulates the relevant business logic. It consists of components, some of which may expose service interfaces that other callers can use. These includes business logic, business workflow and business entities components. The data layer provides access to data hosted within the boundaries of the system, and data exposed by other networked systems. The data layer includes data access and service agent components. The implementation of the Marking Assistant web application was done using open source development technology tools. Open source software allows for greater freedom in development without severe vendor lock-in that comes with the use of proprietary packages. In addition, open source packages are less resource intensive and can run well on even older hardware (Weber, 2004).

6.3.1   Technologies used in the development

As mentioned in the previous section, open source technology development tools have been used because of the freedom advantages. Javascript was utilised on the client side while PHP was employed on the server side. This is because they are both cross-platform, well supported, have good performance and come with an extensive library of functions (Flanagan, 2006; Welling and Thomson, 2003). To achieve the design goals appropriate technologies were employed, PHP, MYSQL, JavaScript, Ajax, jQuery and CSS were used. Description of these main scripting tools together with other tools commonly used with them are outlined below.

- **PHP** – is the server-side, scripting language used for the design with HTML. It provides greater flexibility in the design of websites by enabling the creation of dynamic pages. Page contents are changed based on interaction with the user or data stored in the database. As mentioned earlier, PHP offers many advantages because it is open source and can be used across different platforms.

- **MYSQL** is the database used in the design. It is relational database management system in which data is stored in multiple tables by the sharing of keys. MYSQL uses SQL (structured query language) to create, manage and retrieve information from the database. It also controls access to the stored data.

- **Cascading Style Sheets** (CSS) were used to control the look and feel of the web pages by the application of set styles. These allow the application of one style to many elements and to different pages.

- **JavaScript** is the scripting language that is used to add interactivity to the Web App; the codes are interpreted and run by the web server. JQuery AJAX and JASON were used to ease the manipulation of data and to enhance the user experience.

- **Apache web server** is the open source web server used to serve the pages from the Marking Assistant.

The flow of activity that results when an assessor, through a browser, requests a student record to assess, is depicted in Figure 6.3
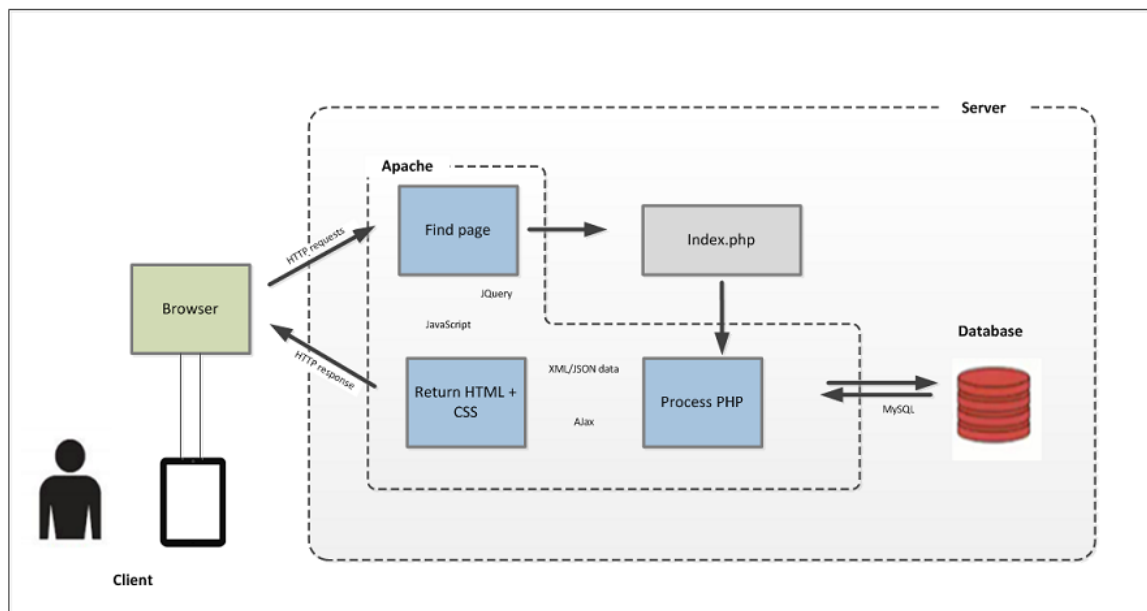
Fig. 6.3 Marking Assistant development outline

The browser makes a request to the Apache web server to find the main page (index.php). The page is located on the server. Because it has the .php extension, it tells the server to process the file as PHP and look for the PHP tags, this is followed by interactions with the database by the sending or receiving data. In the end, HTML is generated and is returned to the user's browser. The marking process model utilises these tools is discussed in the next section.
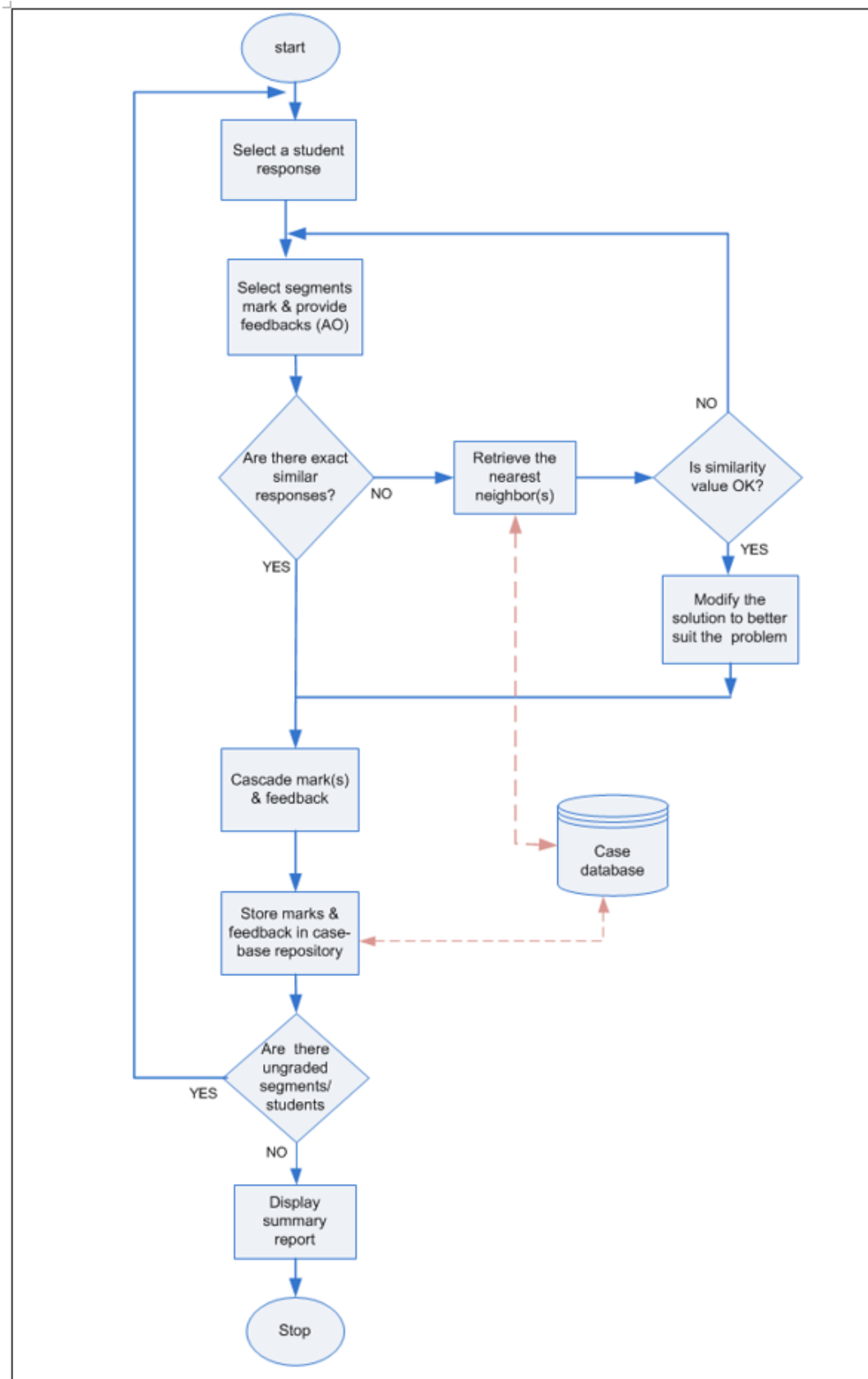
Fig. 6.4 Marking Process Model

6.4   The marking process

There is a set of student solutions available. Each solution has identifiable segments which can be marked separately, producing marks and comments. Marked segments together with marks and comments constitute Assessment Objects which are stored in the case database. Following the requirements set are shown in Section 6.2, steps to meeting them may depicted in a process flow diagram. Figure 6.4 shows the marking process model. As shown in the figure, the process begins by a teacher selecting a student's response to a question. The teacher then selects a segment of the student's work and marks it, providing feedback comments along with the mark for that segment. The system checks if there are similar responses by other students for that segment. If an exact match is found the marks and feedback comments are immediately applied to the response(s). The successfully marked segment is stored in the case library. Where no exact match is found, the system retrieves the nearest neighbour displaying its similarity value. If the value is sufficiently close to the marked case, the marks and feedback comments are retrieved to suit the retrieved record better. A new case is inserted with the marks/comments which have been modified/derived from the nearest neighbour. The system utilises the nearest neighbour algorithm for case retrieval. The process is repeated until all the segments, and all students' responses to the questions are marked.

6.5   Case definitions, representations and organisation

Case representation involves deciding what to store in the case base for retrieval and re-use. Section 4.7 discussed the different case-representation schemes. The attribute-value representation according to Watson (1998) is adequate for many case-based reasoning systems. This type of representation is used in the MA.

The typical case attribute-value representation comprises of the situation description and outcome. In the MA, the segment(s) of the students' responses represents the situation, the teachers mark(s) on the segment together with the feedback comments constitutes the outcome part. The case representation for the mathematics word problem described in Chapter 4 is laid out in Figure 6.5. Only key attributes have been selected for simplicity.
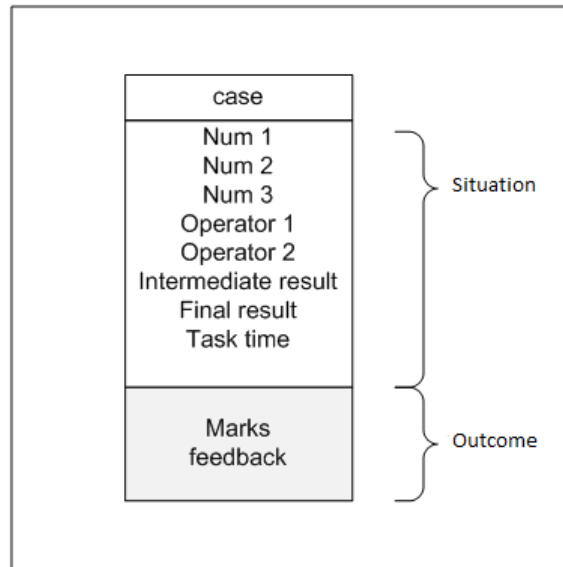
Fig. 6.5 Definition of an individual case

The figure shows numeric operands and arithmetic operators used as well as the integer responses entered by the student to constitute the situation of the case while marks and feedback make up the outcome. Table 6.1 shows definitions for five cases having different levels of granularity.

Table 6.1 Five cases at different levels of granularity

Case No.	Num1	Num2	1st operator	Interm. ans.	Num 3	2nd operator	Final ans
Case 1	6	18	+				
Case 2	6	18	+	24			
Case 3				24	14	+	
Case 4				24	14	+	38
Case 5	6	18	+	24	14	+	38

The table shows that some cases may be small. The smaller cases can be subsets of the larger ones. For example, case 1 is a subset of case 2, which in turn is a subset of Case 5. Thus, the

cases may be viewed as having a hierarchical structure. The case hierarchy is a representation schema for a case and corresponds to the decomposition of the student's solution processes. This is illustrated in Figure 6.6
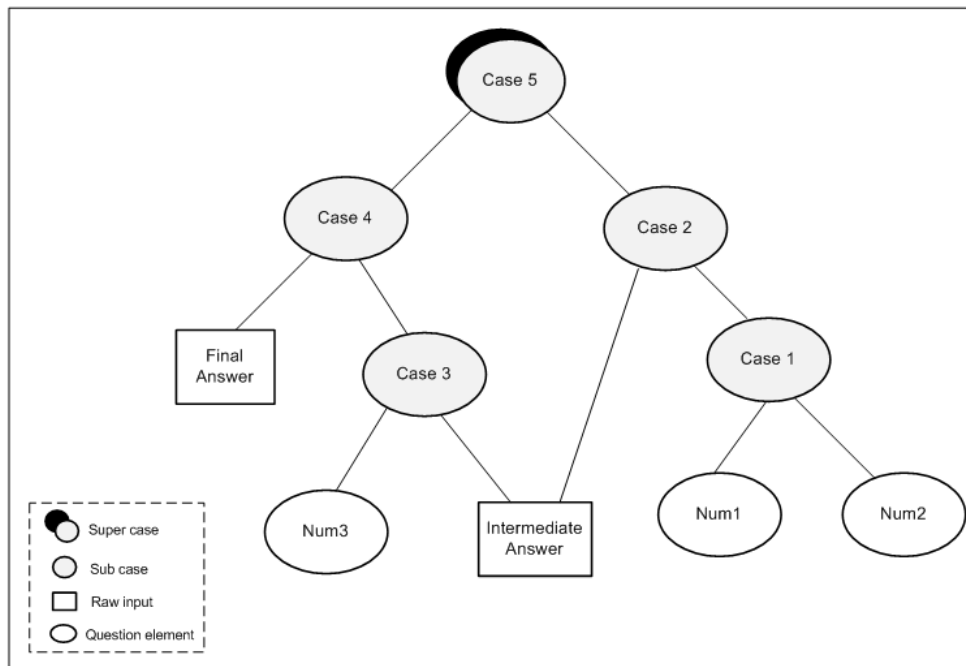
Fig. 6.6 Hierarchical case representation

Each marked segment is stored declaratively in a hierarchical structure where each node in the hierarchy comprises of attribute-value pairs. A full solution to the problem may be viewed as a super case which consists of multiple levels of subcases. The figure shows Case 5 to be a super case for the question containing all the attribute elements of the student's solution. Subcases are directly indexed individually with links that can be used to construct the whole case. Case 1 is the smallest segment of the students' response. It shows that the student recognised two operands and used an arithmetic operator. This is an *atomic assessment object*. An atomic assessment object as defined in Section 4.6 is the smallest assessed segment in a question that may be assigned a score and/or feedback comment. This follows from the definition of an atom and suggests a fundamental idea that atoms can be grouped, nested and combined.

6.6 The prototype user-interface development

The implementation of the approach outlined was realised using web and database application technologies. These technologies are widely used and are platform independent. The students' record and case base were implemented with MySQL database. JavaScript was used for the client side scripting while PHP was used for the server scripting. Like the MuTAT, the MA was implemented on the iPad but as a webApp for speed. Similarly, Interactive touch gestures were also utilised to facilitate easy input of marks and feedback comments.

Figure 6.7 shows possible marking for a student response to the example word problem used in this chapter .
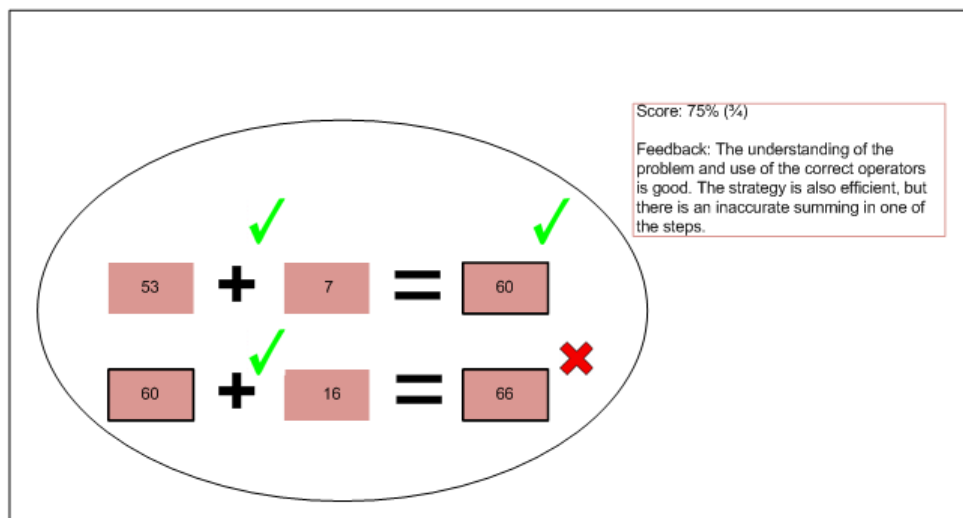
Fig. 6.7 A sample analytic marking and feedback

The first functional requirement is for a user interface that allows the application of an analytic rubric. The Marking Assistant user interface is shown in Figure 6.8. Like the MuTAT, the principle of simplicity was applied.
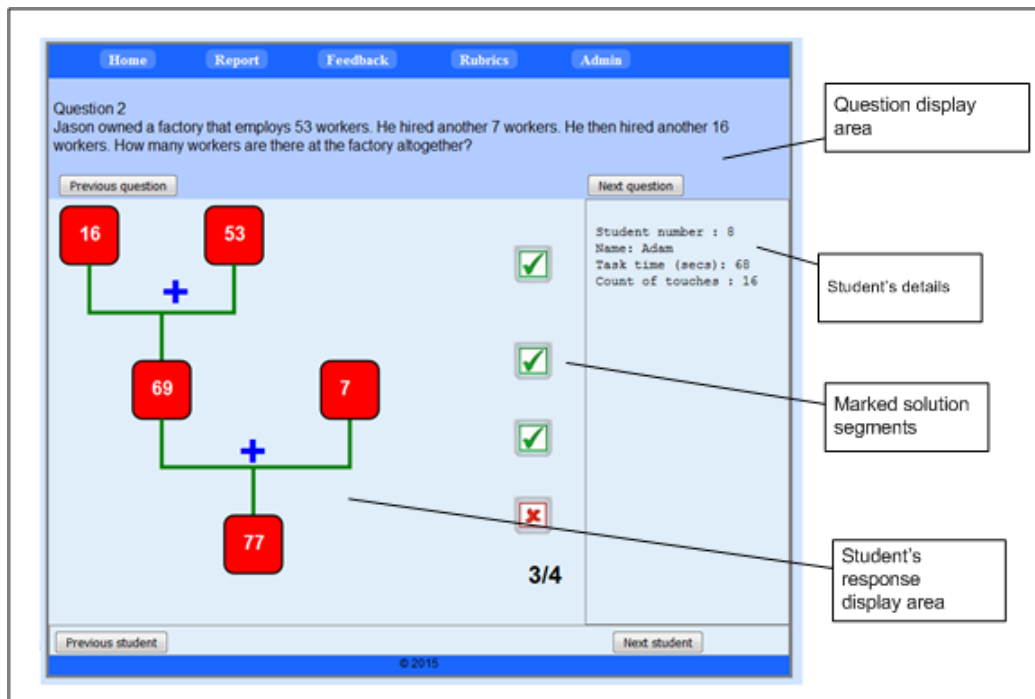
Fig. 6.8 The Marking Assistant user interface

It has a question display area which displays the question for the teacher. Each student response is displayed one at a time in the student response display area. Solution segments may be selected and marked by simply selecting the segment(s). When a segment is marked as correct, a tick mark indicates this. Otherwise, a cross mark indicates the response to be wrong. If any segment matches an already marked segment in the case base, then the marks/comments will be displayed.

Feedback may be entered by clicking the segment where the feedback is desired. This opens a contextual feedback area which accepts written feedback. This is illustrated in Figure 6.9.
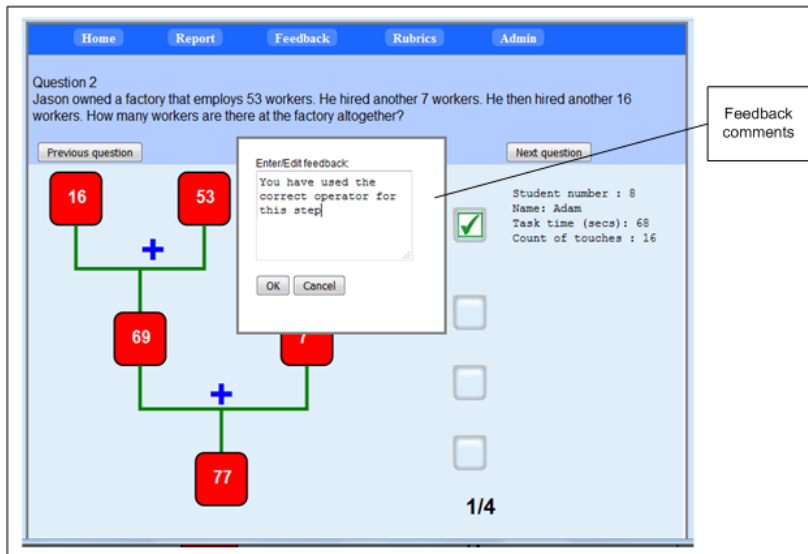
Fig. 6.9 Entering feedback comments

When the OK button is pressed, a message indicating that it has been added is displayed in the information area. After all the segments of the work have been marked, a detailed view of the students' work may be obtained by selecting the feedback button. The view shows the question, student's response, tick marks, and the feedback comments. This is shown in Figure 6.10.
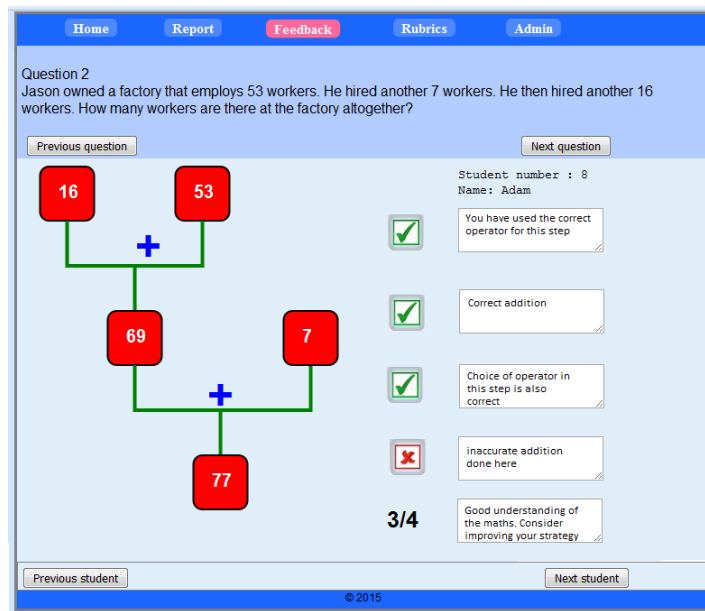
Fig. 6.10 Detailed feedback view

As described in Section 6.4, at the completion of the marking of each segment set the marks, and feedback comments are stored in the case base for subsequent retrieval and re-use. Consequently, the assessment/grading should be consistent across multiple grading sessions. When the marking exercise has been completed a report detailing key information on the marking exercise may be assessed. A sample report is shown in Figure 6.11.
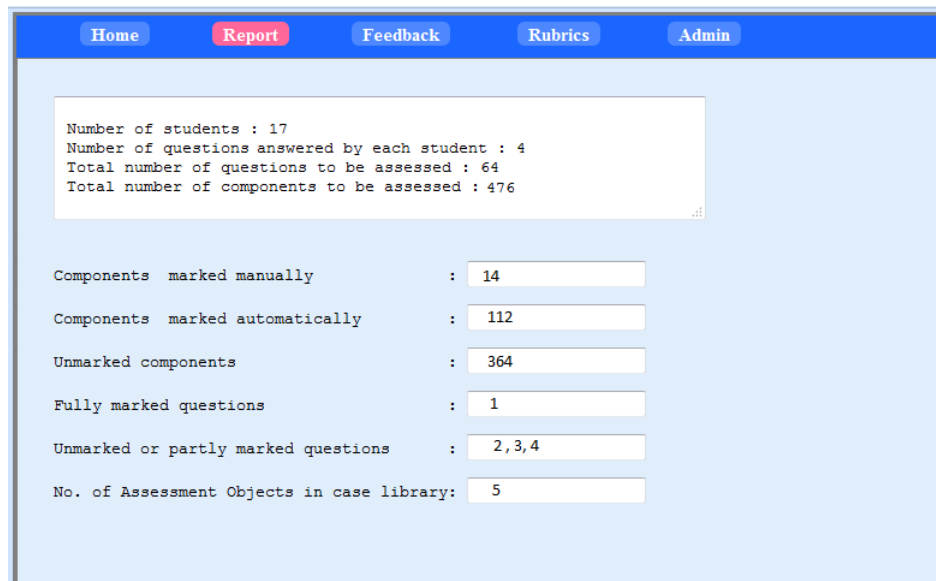
Fig. 6.11 Report on the marking process

The report shows among other parameters, the number of items marked manually and those marked automatically. It also provides information on the nature and count of student submissions. The rubric button presents the marking guide to the teacher. It records the model answers for the questions and may be checked regularly to aid consistency in the marking. The Admin button serves to assist in the technical administration of the tool and setup of access to the records.

6.7   Chapter summary

This chapter has presented the design and implementation of the Marking Assistant. It showed how cases are represented and the described the workflow for the marking and feedback processes. The chapter demonstrates the application of the case-based reasoning approach to marking students' solution steps. The next phase regarding the MA tool will be

the evaluation of the time-saving potential of the prototype. This will be presented in Chapter 8.

Chapter 7

MuTAT Evaluation

The design and implementation of the MuTAT were presented in Chapter 5. When innovative systems and products are designed with little regard to usability and acceptance, frustration, wasted time and errors often occur. The next phase of the study is the evaluation of the prototypes in terms of functionality, performance, and usability as outlined in the requirements definition. The test results need to be interpreted and evaluated based on the semi-automatic conceptual framework presented in Chapter 4 to see whether the anticipated benefit is achieved or not, and what problems may occur. Experiences gained here will lead to further development and enhancement of the system. This chapter presents the evaluation studies carried out to investigate the usability and potential benefits of the tool. The studies carried out to evaluate the MuTAT are in three parts. The first is a pilot study conducted to assess quickly the feasibility of using the prototype tool. The second is a study carried out in primary schools with real students in their study environment. The third study is with teachers to investigate their opinions on the usefulness and feasibility of using the tool in school classrooms. For each study, the design and description are first presented. After this, the results and discussions are then presented.

7.1   Study 1: Pilot study

The pilot study was the first study conducted after the first version of the MuTAT was completed. The purpose of the study was to examine the feasibility of the approach implemented on the tool before undertaking a bigger study. A pilot study provides the opportunity to develop consistent practices to enhance data integrity and the protection of human subjects

(Flick, 2015). Also, it offers a means of collecting preliminary data and the results assist in guiding the design and implementation of larger scale studies. In addition, a pilot study helps to discover errors and provides some practice at moderating and conducting live sessions (Bryman, 2012).

7.1.1 Study goals

The goals of the pilot study are:

1. Find out whether the study participants are able to successfully use the MuTAT to solve problems

2. Find out whether the multi-touch interactions implemented works as expected and are usable

3. Find out how satisfied the participants are with the general functionality of the tool

.

7.1.2 Participants

The population used for the pilot study comprises of Loughborough students in non-numerate programmes. Although the participants are university students, it was believed that the findings may be relevant to younger learners as well. This agrees with McNeil et al. (2010), who argued that findings from an adult participants may be applicable to children to in arithmetic knowledge studies. The sampling frame were Loughborough University undergraduate students studying in the School of Arts, English and Drama. Students in the school usually do not have a high degree of mathematics in their core subjects and so are less likely to regularly undertake mathematics tasks compared to students in numerate disciplines like Maths, Physics, Computer Science and Engineering. As indicated in Section 3.5.2. Convenience sampling was used to access readily available participants (see Section 3.5.2). A study table was set up at a regular lecture area for the students. After their lectures the students were asked to voluntary participate in the study. Students that do not meet the stated criteria were screened from the study. The final realised sample included a total of 17 participants (7 males and 10 females).

7.1.3   Task design

An introduction session was given to each participant on sample questions to show them on how to use the tool to solve problems as described in Section 5.5. After this, the participants were asked to solve four word problems using the techniques demonstrated on the MuTAT. The full range of questions can be found in Appendix C. The arithmetic problems essentially contained three numbers which are to be summed together to produce a final answer. After the participant had finished the task they were asked to fill a usability questionnaire (Appendix C). The questions on the usability questionnaire are typical questions from usability studies investigating ease of use of applications (Lewis, 2006). The allotted time for each participant session was 20 minutes.

7.1.4   Data Collection

There were two ways data was collected. The first is from click/touch screen interactions in the log files of the MuTAT while the participants were solving the arithmetic problems (see Section 3.5). The second was from the usability questionnaire the participants were asked to complete after performing the study tasks.

7.1.5   Results and Discussions

All the students were able to solve the problems, and all responded to the affirmative on the questionnaire that they were able to do this. An analysis of the detailed results generated on the tool showed that 98% of the participants had correct answers. Only one participant approached a question using subtraction rather than addition. This may be due to his lack of proper understanding of the question rather than not being able to use the tool. While the study did not set out to test arithmetic ability, this finding suggests that the tool did not prevent the students from solving the questions and inputting answers thought to be correct. To assess the usability of the multi-touch approach, the participants on completing the tasks were asked to respond to a Likert-type question with a six-point scale, asking how easy it was to complete the tasks with the MuTAT. The responses from participants are shown in Figure 7.1.
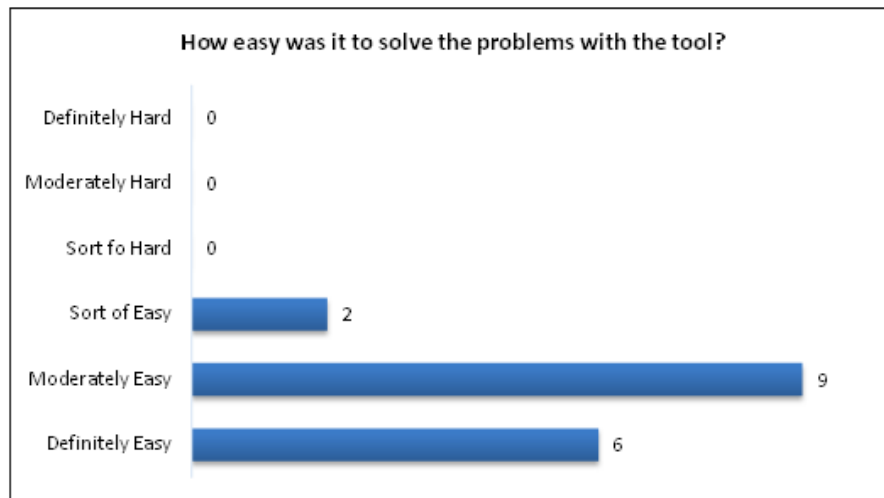
Fig. 7.1 Ease of use of MuTAT

The overall response to this question was very positive, all the participants expressed that they were able to successfully carry out the tasks. Over half (53%), responded that the found the tool moderately easy to use, 35% found it very easy while the others (15%) reported using it was sort of easy. The results further confirms that the MUTAT did not prevent the students from solving the questions and inputting answers thought to be correct.

Table 7.1 presents the open ended questions on the questionnaire and the responses given by the participants.

Table 7.1 Comments made by participants

Participant	**What do you like about the tool?**	What don't you like about the tool
1	Very interactive and easy to use	It was ok
2	Easy and clear	-
3	It was definitely easy and convenience to do calculations.	I could not use two fingers well, and it should be designed to do some more complex calculations
4	Fun and educative	Hard to input numbers
5	Easy to learn and intuitive	Not like using the PC to add
6	Cute, easy, handy and easy to understand	It is not very suitable for children to use fingers when the figures shakes you should improve on this part
7	Funny software	-
8	A good way to teach pupils	You need to be able to rearrange the numbers as you like. Use different colours for the numbers
9	Interesting	It's OK
10	-	Too much animation Improve the colour scheme
11	Idea interactive more exciting than paper	Too sensitive to touch
12	It's useful for calculations	I prefer to write down than drag
13	Fun to do mathematics	Touch and hold appears to be "clunky"
14	Dragging is good	-
15	I like the design, and the idea. I think it is feasible	The design could be more colourful as it is aimed at children. The button could respond better
16	Interactive	-
17	It is an easy way of solving problems	Mental sums can be faster than using this tool

Regarding what the participants liked about the tool, over half to the participants responded that they appreciated the way they were able to interact with the questions. The hands-on approach of inputting responses was described as being engaging and exciting. The ease with which they were able to use drag-and-drop interactions was welcomed. The novelty

of the interactions may have appealed to the participants however it supports the view that students are generally more engaged in task when they are active in problem-solving.

On what the participants did not like about the tool. The main concern was on the use of the long-press gesture to bring up the operator menu pad. Half of the participants struggled at the point of activating the operator pad. This was probably because most of the participants with the problem preferred to hold the iPad with their hands while solving the problems rather than setting it on a table. This highlighted a potential problem for the use of the touch-and-hold gesture when an iPad is not held in a steady position. The iPad needs to be on a flat and undisturbed surface before the multi-touch action on the numbers is activated, without this the numbers become "jumpy". At this point further refinement is needed on the gestural actions to minimise the problems.

7.1.6   Findings

The pilot study provided promising results. University students from non-numerate disciplines were all able to successfully carry out the arithmetic problem with the MuTAT. They were generally satisfied and valued the use of touch gestures and multi-touch interactions as the primary means of interacting with the system. However, the study revealed that the touch and hold-and-hold gesture needed refinement for smoother user experience.

7.2   Study 2: Primary schools study

The pilot study described in the previous section was a precursor to the study in primary schools. The findings regarding the need for a smoother long-press multi-touch interaction guided further development and coding efforts on the MuTAT. Although the pilot study provided promising results there are many tangible benefits to seeing how target primary school students use the tools in their normal study environment. Running studies in school locations provides a naturalistic setting which can provide authentic insights on how students may actually use the tools (Lewis, 2006). The target population is young students in the age range of 9-10 years. Typically at this age range the students are able to read and solve problems (Borthwick and Harcourt-Heath, 2010). The details of the evaluation study are presented in this section.

7.2.1   Study goals

The goals of the primary schools' study are to:

1. Find out whether primary schools will be able to use the MuTAT to solve problems

2. Compare performance on the MuTAT results with that of similar paper-and-pencil test

3. Find out whether higher-order level cognitive skills like arithmetic strategies may be deduced

4. Study the relative efficiencies of the different students' solutions

5. Observe the attitudes and behaviours of the students while working on the problems

7.2.2   Participants

The participants were 10-11-year-old children (24 girls and 15 boys) attending Year 5 classes in two rural schools in England. The schools were attended by children of mixed socio-economic background. Year 5 children are expected to have knowledge of elementary multi-digit addition and word problems (Borthwick and Harcourt-Heath, 2010). The typical school year runs from September to July. The study was carried out during the last month of the school year. The experiment was divided into two parts: the first involved the participants solving word problems on the MuTAT application while the second part involved solving similar problems on paper.

7.2.3   Experimental design

In the first part of the study, the participants were required to solve word problems using the MuTAT application. Well-researched word-problems specially designed to allow detection of strategies were chosen (Carpenter and Moser, 1984,?; Yeap and Kaur, 2001). The problems contained numbers that were chosen to support the use of the place-value strategy by students as discussed in Section 2.7. In each problem there are two large two-digit numbers, and one single-digit number. Two of the numbers bond to multiples of 10 as shown in questions 3, 6 and 7 in Table 7.2. In each problem the two numbers that bond to multiples of 10 are presented in different positions: 2nd and 3rd in question 3; 3rd and 1st in question 6; 1st and 2nd in question 7. The particular values were selected so that adding the two-digit numbers requires a carry over. The large numbers were selected such that the computational burden

is minimised by starting with the large number, and then one of the smaller numbers or vice versa. Question 7 presents the numbers in strategic order, with the large and the small numbers placed together. This was a control question which served to help us work out if any participants consistently either (i) just chose numbers from left to right or (ii) just choose numbers arbitrarily.[1]

Table 7.2 Four representative questions from the study

Question No.	Question
3	Joe had **17** toy cars. His father gave him **26** toy cars. His mother gave him **4** more toy cars. How many toy cars did Joe have altogether? [a]
6	Sara has **8** sugar donuts. She also has **15** plain donuts and **32** jam donuts. How many donuts does Sara have altogether? [a]
7	Jason owned a factory that employs **53** workers. He hired another **7** workers. He then hired another **16** workers. How many workers are there at the factory altogether? [a]
10	There are **22** sheep in the field, **13** more are put there and some moved away. At the end there are **28** sheep in the field. How many were moved away? [b]

Drawing on Gray and Tall (1994) and Gilmore and Bryant (2006), it was hypothesised that some students who understand arithmetic conceptually would consistently use the place-value strategy in the questions to ease the computational burden. Students with a less secure conceptual understanding would only make partial use of the place-value strategy. Students who do not view arithmetic conceptually would consistently go from left to right, or select numbers arbitrarily in no particular order.

The first two questions the students attempted were practice questions not included in the analysis. The next five questions were similar to Questions 3 and 6 and 7. In addition, there were five exploratory questions similar to question 10 above. The exploratory questions required the students to make use of addition and subtraction operations. They are distractors and served to help check if the students understood the requirements in the problem text to use a mix of plus and minus operators.

[1] Adapted from Carpenter and Moser (1984)[a] and Gilmore and Bryant (2006)[b]

In the second part of the study, the students worked on a paper-and-pencil test. The paper tests were a research instrument to check the performance scores of the students in similar tasks. In the test word problems were not used, instead, the same question structure was presented symbolically (e.g. 36 + 4 + 18). This was mainly to reduce the amount of work the participants had to do within the time frame for the study.

7.2.4 Measurements

The variables recorded by the tool are question number, type of gesture interaction (i.e. drag, long-press, and touch), numbers selected, operators used, X and Y coordinates of each element on the screen to obtain data about the physical and logical locations of the solution items. The time-stamp for each operation was also recorded to establish the sequence of events. For the paper-and-pencil tasks, the workings and answers were collected for comparative analysis with data collected with the MuTAT application.

A total of 546 questions were answered by the 39 participants (390 on the MuTAT application and 156 on paper). All the responses from the MuTAT were recorded in a log file. The file contained a total of 8,221 records. Where each record contained the current values of all the variables.

7.2.5 Results

The summary results of the experiment are presented in both qualitative and quantitative forms in this section. Further results have been presented in Appendix C. The analysis of the responses on the MuTAT and paper are first examined. This is followed by a discussion of the strategies used by the students.

I **Performance scores**

On the MuTAT application, all the participants were able to complete all word problems. Also, all the students made the correct choice of arithmetic operation (addition) in answering the study questions. Most of the students used a mix of the plus and minus operators for the distractor questions. This suggests that the students were able to use the tool to enter their desired solutions and understood the semantic relations in the problem text. This claim was also confirmed by observations while the participants were working on the experiments. The participants also successfully completed the paper-and-pencil questions.

Figure 7.2 compares the results obtained from the MuTAT application and paper-and-pencil task in a bar plot with standard error as error bars.
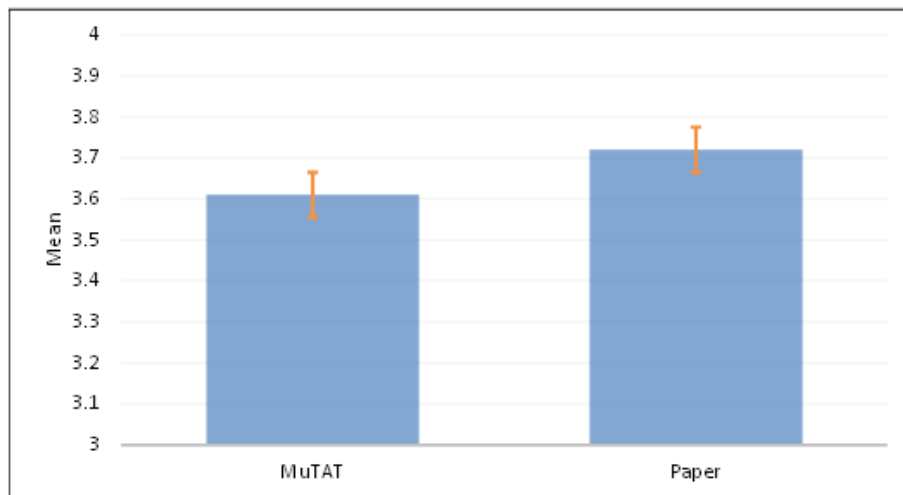
Fig. 7.2 Bar plot with standard error as error bars of mean scores

This figure showed that the mean scores on the MuTAT software and paper are comparable (MuTAT-> M = 3.61, SD=0.59; Paper ->M = 3.72, SD = 0.51). Both treatments had a mode score of 4. However, the performance on the MuTAT application appeared to have a greater variance. To test for significant differences between the two methods, a paired t test suggested that there were no significant differences between the means of the two groups (t-(38) = 0.94, p = 0.35).

II Time on task

An analysis of task time can indicate problems in using a tool or method. Scholars like Bailey and Garner (2010) and (Lewis, 2006) have argued that task times can be very useful in diagnosing usability problems such poor interactions and inefficient procedures. Table 7.3 provides univariate descriptive statistics for the times the participants spent on completing the tasks on the MuTAT platform. It can be seen that the average time spent on all the questions was 38 seconds.

Table 7.3 Univariate descriptive statistics for the task times (n = 39)

	Minimum	Maximum	M	SD
Completion time (seconds)	19	92	38.46	0.25

The mean score of 38 seconds suggests that the participants did not spend an excessive amount of time in using this approach. Because the current research was not specifically designed to measure the task time for the paper-and-pencil test a comparative task time analysis could not be made between the two methods. However observations made during the experiments suggested no notable difference in the task times.

III Detailed feedback

As mentioned in Section 2.4.1, traditional paper-and-pencil tests may capture detailed solution steps, but the absence of consistent notation for representing solution steps and the difficulty in enforcement may present some challenges. Figure 7.3 shows the responses from three students Lucas, Julia and Mark.

Fig. 7.3 Paper-and-pencil responses from three students

In the figure, Lucas appeared to have used as-presented strategy, arranging the solution items. Julia appeared to have used the place-value strategy in all the questions, and intermediate answers are shown. In contrast to Lucas and Julia, Mark did not show his workings hence it may not be possible to provide accurate feedback on where and why he missed the correct answers for questions 2 and 4.

The MuTAT application provides rich and consistent data that can be used to assess in detail the solution of each student. Figure 7.4 shows the diagram of the solution steps obtained from the tool for one of the study questions. Here Alan paired 6 and 14 as a first step, then entered an intermediate result of 30. He used this answer in the second calculation stage pairing it with 18 before entering a final result of 48.
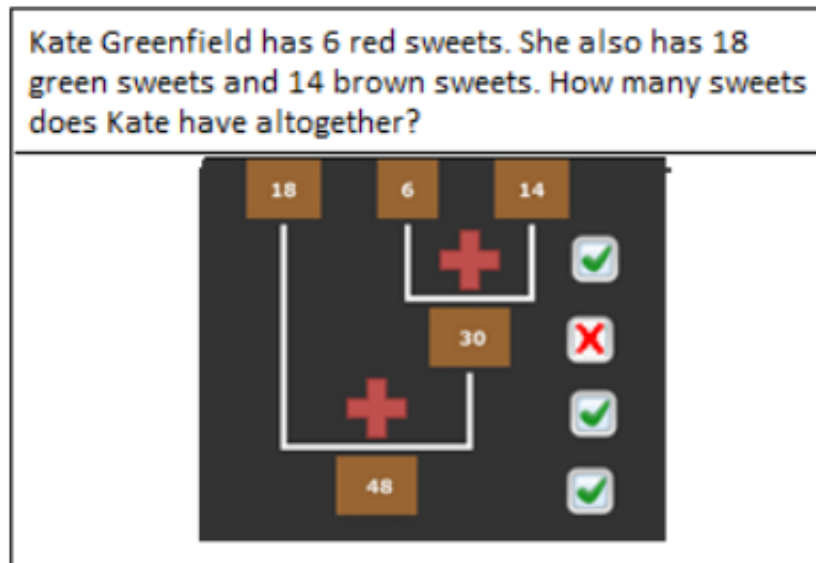
Fig. 7.4 MuTAT Diagrammatic representation of solution steps from Alan

From this figure, it is easier to see why and where Alan had difficulties in answering the question correctly. While he understood the problem and selected the correct operators, he got the first addition step wrong. A more accurate feedback can be given to him and he may be assigned partial credit for the parts of the solution he got right.

.

7.2.6   Strategies

As discussed in the literature review, performance on a problem-solving task is dependent on both strategy selection and strategy efficiency (Mayer and Wittrock, 1996). It is possible to use the data from the MuTAT application to examine these two characteristics. The following sub-sections discuss the strategies used by the participants.

I **Strategy selection**

As discussed in Section 2.7, strategy types can be inferred from the way the students paired the numbers. The students all completed five study questions. The strategies obtained from the MuTAT platform, ignoring pairwise commutations is summarised in Figure 7.5. In this figure, the five study questions are identified as 3, 4, 5, 6 and 7 respectively. Question 7 is the control question described in Section 5.2. Also, it can be seen that the use of strategies varied across questions for the 39 participants. More

students used the place-value strategies on questions 4, 6 and 7, while the as-presented strategy was used more frequently on questions 3 and 5. Most of the students used as-presented on question 3 possibly because it was the first study question and they were starting to get used to the application. The reason why no participants were coded as using the as-presented strategy in the control problem (question 7 in Figure 7.5) is that it was designed such that the as-presented and place-value strategies could be counted together (both were coded as place-value).
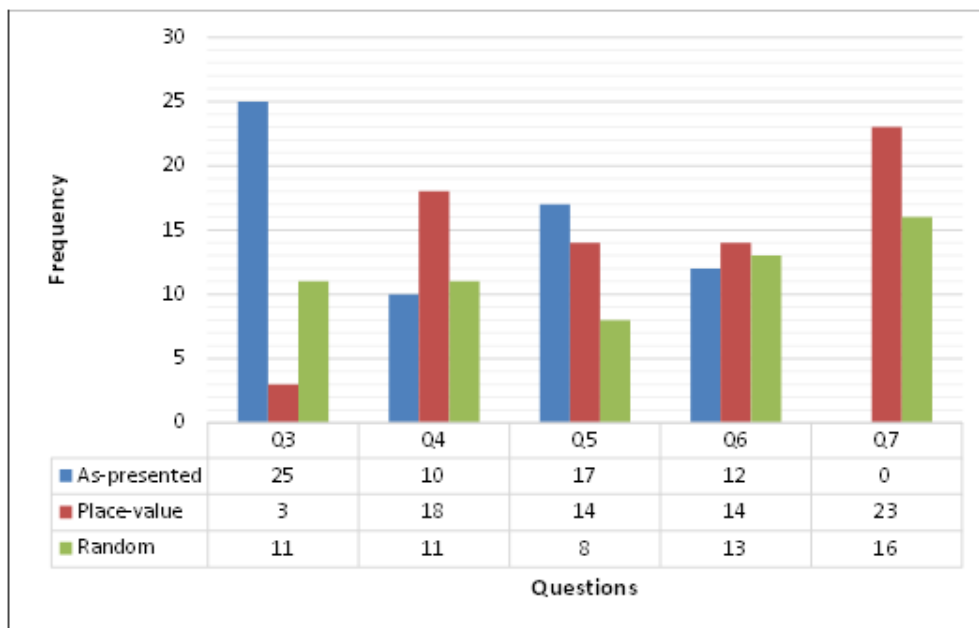
Fig. 7.5 Strategies output across questions

In order to investigate whether the participants were working strategically, the place-value strategy was subjected to a goodness-of-fit test. Had the students been selecting and pairing the numbers randomly the expected frequency for each of the three strategies on a given question would be 13 (i.e. 39÷3). The observed place-value strategy frequencies across the five questions were 3, 18, 14, 14 and 23 respectively. A one-sample chi-square test was significant but marginal, X^2 (4, N=39) =9.58, p=.048, suggesting that some but not all of the students were working strategically. This outcome is consistent with the hypothesis that some participants would use the place-value strategy, and others would add the numbers in the order presented or arbitrarily in each question.

A closer examination of how individual students consistently used the place-value strategy is shown in Table 7.4.

Table 7.4 Usage of the place-value strategy

No. of times place-value was used	No. of students
0	6
1	14
2	8
3	2
4	9

The table showed that about half of the students (20) used place-value once or not at all; Of the 14 who used place-value once, 8 used it on the control question; and of the 20 who used place-value once or not at all, 9 used mostly the as-presented strategy. This suggests the MuTAT application may have detected strategic groupings amongst the students between those who selected the numbers as presented and those who were more strategic.

II Strategy efficiency

Strategy efficiency examines how quickly and accurately a strategy leads to the solution. Figure 7.6 presents data from the MuTAT on the strategies and the average times it took children using the strategies. The data shows that the students spent less time using the place-value strategies in three out of the five questions.

	Q3	Q4	Q5	Q6	Q7
As-presented	1.10	0.90	1.14	0.92	0.58
Place-value	1.23	0.75	0.75	0.58	0.88
Random	0.91	0.98	0.97	1.02	1.05
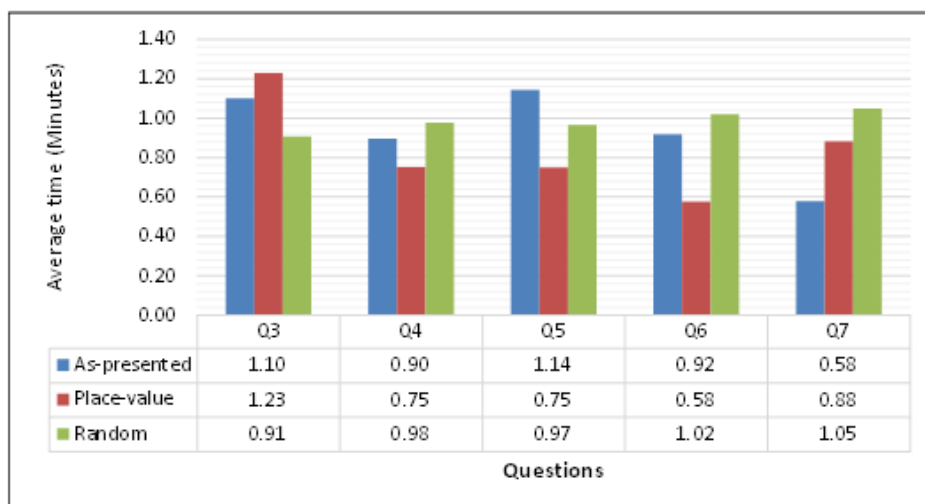
Fig. 7.6 Strategies and the average times

To compare the completion times of the three strategies, a one-way ANOVA F-test was used. The result was significant ($F = 4.9$, $p = 0.01$) indicating that there are differences between at least two of the average time values. To check for where the differences lie, a TukeyHSD pairwise comparison was carried out at 95% confidence interval. Figure 7.7 shows a plot of the differences in the mean times.
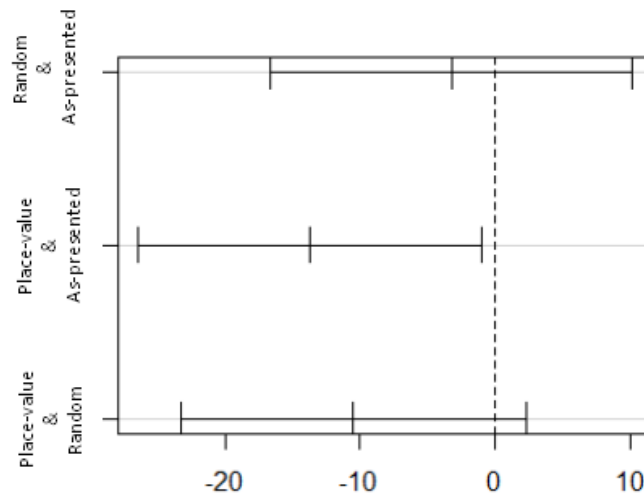
Fig. 7.7 TukeyHSD of difference in average times of strategies

The test revealed significant differences between the mean time of the place-value strategy (p=0.01) and the as-presented (p=0.13) and random (p=0.58) strategies. This agrees with Imbo and Vandierendonck (2008) who argued that computational strategies are more efficient than procedural strategies. Therefore, the MuTAT application can aid the determination of the relative efficiency of strategies thereby enabling the provision of suitable and appropriate feedback.

7.2.7   Qualitative analysis

To determine if the strategies used on paper-and-pencil corresponded with that used on the MuTAT application, the responses of Lucas, Julia, and Mark (see Figure 7.3) were closely analysed. Lucas used the as-presented strategy in all four questions of the paper-and-pencil test and he used the same strategy in all the five questions on the MuTAT. Julia used the place-value strategy in all the paper-and-pencil test questions and four of the five questions on the MuTAT. Mark, on the other hand, used a mixture of strategies: as-presented on 2 questions,

place-value on 2 questions and random strategy of the five questions on the MuTAT. As mentioned in Section 7.2.6, Mark's strategy on paper cannot be accurately determined since he did not show any workings. Analysis of the strategy output from the MuTAT suggests that he may require more support in learning to recognise relationships between numbers and working more strategically. This result suggests that the strategies used on paper may not be different from the one used on the application. However this finding cannot be extrapolated to all students as most of them did not completely present their workings on paper.

Furthermore, observations made during the study suggest that the school students genuinely enjoyed using the MuTAT and were impressed with the functionalities. Students demonstrated positive attitudes while using the tool. For example, when some of the students were working on the tool they responded with glowing comments such as "I like this", "cool", "it is fun". Generally, the students liked the look and feel of the tool and found navigation "nice and easy" and appeared to be comfortable solving problems on it.

7.3   Study 3: Teachers' Study

The previous section described the study conducted with students. This section presents the study that eplored the level of acceptance and opinions of teachers regarding the MuTAT prototype. Teachers' confidence and attitudes toward computer technology integration, and willingness to undertake a change incorporating technology use for student learning is necessary (Levin and Wadmany, 2008) . A study of their attitudes and opinions will provide in-depth insights on the usefulness or otherwise of the tool.

The goals of the study are to:

1. find out whether the teachers consider the tool to be useful

2. ascertain the attitudes of the teachers after working with the tool

3. obtain feedback that can guide future development efforts

7.3.1   Participants

The target population for the teacher study was primary school teachers. The participants should have taught or are currently teaching primary school classes. They will be able to provide an objective evaluation based on their experiences. Getting access to teachers in the schools was difficult because of their busy schedule. However, the annual BETT show which

brings together teachers from different schools provided an opportunity to access and interact with teachers from various schools (bettshow.com, 2015). The study was conducted during the 2015 edition of the event. A stand was set up and the teachers were invited to work with the software. The profile of participants is presented in Table 7.5.

Table 7.5 Participants

Teacher	Number
Early Years Foundation Stage (EYFS) - age range from birth to 5 years old	1
Key Stage 1 (KS1) also known as Year 1 and Year 2. Students pupils are aged between 5 and 7	1
Key Stage 2 (KS2) normally known as Year 3, Year 4, Year 5 and Year 6. Students are aged between 7 and 11.	2
Key Stage 3 (KS3) also known as Year 7, Year 8 and Year 9 . Students are aged between 11 and 14.	3
Key Stage 4 (KS4) also known Year 10 and 11 . Students are aged between 14 and 16	2
Others	1

From the table, it can be seen that a total of 10 participants participated in the study. Nine of them were active teachers the tenth participant was an administrator who was involved with teaching.

7.3.2   Task design

The task design is similar to that in Section 7.2. The teachers were invited to work with the MuTAT to solve the problems. After completing this, they were asked to fill a questionnaire soliciting their opinions and thoughts. The questionnaire can be found in Appendix D.

7.3.3   Results and Discussions

The questions are grouped into three categories. The first investigates the perceived usefulness of the MuTAT. The second examines the attitudes of the teachers, and the third invites the teachers to express their thoughts and comments qualitatively.

I **Perceived usefulness**

- *Question* – *How useful do you think MuTAT would be in different subject areas?*
  This question was asked in order to find out if the teachers thought that the principles applied in the design are applicable to different subjects. The responses are shown in a stacked bar chart in Figure 7.8.
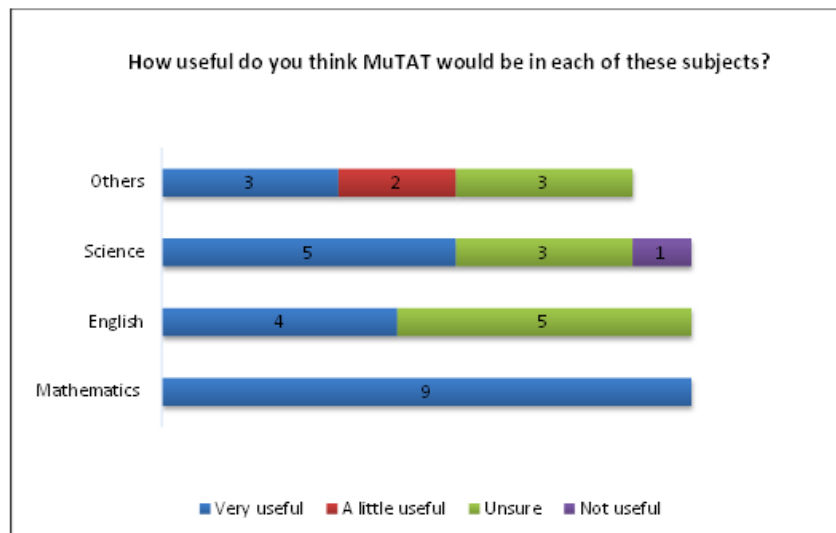
Fig. 7.8 Usefulness of the MuTAT across subjects

  Overall, the stacked bar chart indicates that the teachers thought the tool would be very useful for the major subjects. Mathematics was ranked highest (90%) followed by science (50%) and English (40%) respectively.

- *Question* - *How useful do you think the tool would be in each key stage?*
  This question sought to determine if there was any class or age group the teachers consider the tool to be most beneficial to. Figure 7.9 illustrates the breakdown of the responses given.
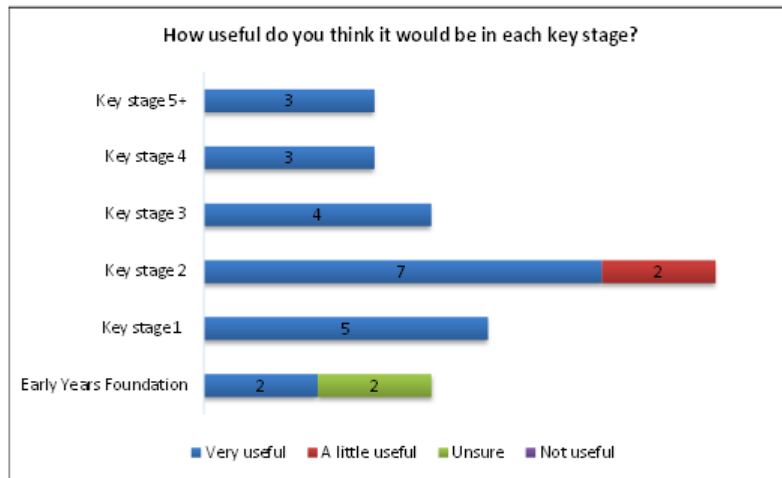
Fig. 7.9 Usefulness at various levels

Key stage2 (ages 7-11) had the highest rating with 9/10(90%) of the teachers agreeing that the application will be very useful or at least a little helpful. Generally, the respondents appear not to consider the application as useful during the early years and the later years.

- **Question** - *What would you use the tool as?*
  This question was to determine if the respondents viewed the tool as a formative or summative assessment tool. From Figure 7.10 it can be seen all those that responded (9/10) considered the MuTAT to be useful as a diagnostic and final assessment tool.
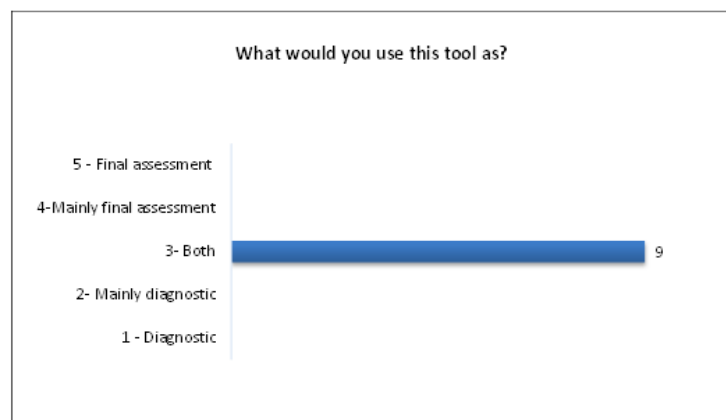
Fig. 7.10 How the MuTAT can be used

The results presented above in this part suggests that the teachers consider the tool to be useful, especially for the Keystage2 class and the major subject areas. Furthermore it can serve both as a diagnostic and final assessment tool.

II Participants' attitudes and level of excitement

- *Question – Have you seen or used a tool like this before?*
  This question was ask to find out if the participants had used similar tools and are familiar with the operations and design. Only two participants responded, one to the affirmative and the other negatively. The low response rate may be because the MuTAT is new and respondents could not easily place it alongside previous tools they had used.

- *Question – How excited are you about the tool?*
  The responses are illustrated in Figure 7.11.
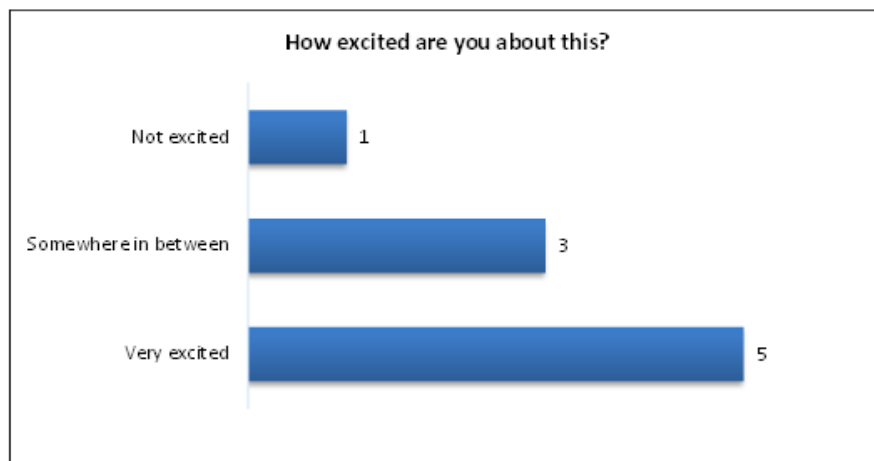
Fig. 7.11 Excitement about MuTAT

From the figure, it can be seen that only one of the respondents was not excited about it. Half (5/10) of the teachers were very excited about the application, However a third (3/10) were undecided.

- *Question – Would you like to have the app in your school?*
  This question examined the willingness of the respondents to use the tool in their classrooms. Figure 7.12 presents the responses.
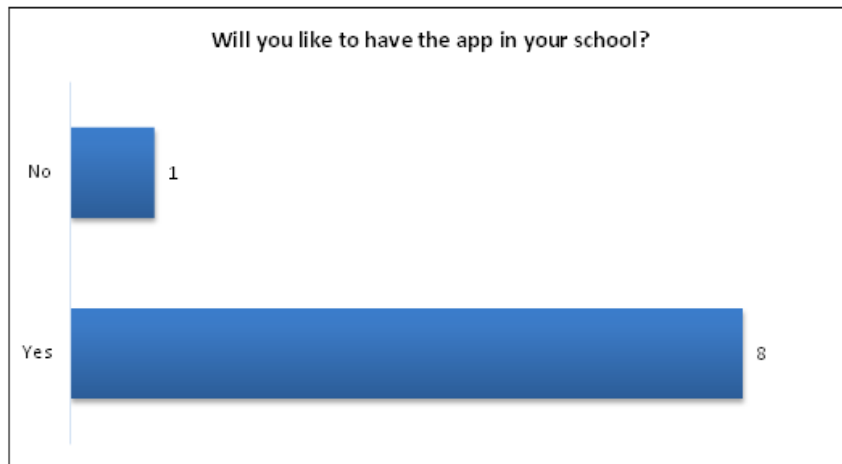
Fig. 7.12 Use of app in schools

The chart shows that the majority 8/10 (80%) of the respondents would like to use it in their classrooms. This positive response suggests that they were comfortable with adopting the tool for regular use in their classes.

- *Question* - *How would you like to see the app delivered?*
  This question was to determine if there was a preferred technology platform for the delivery of the contents of the application. Figure 7.13 shows the chart of the responses.
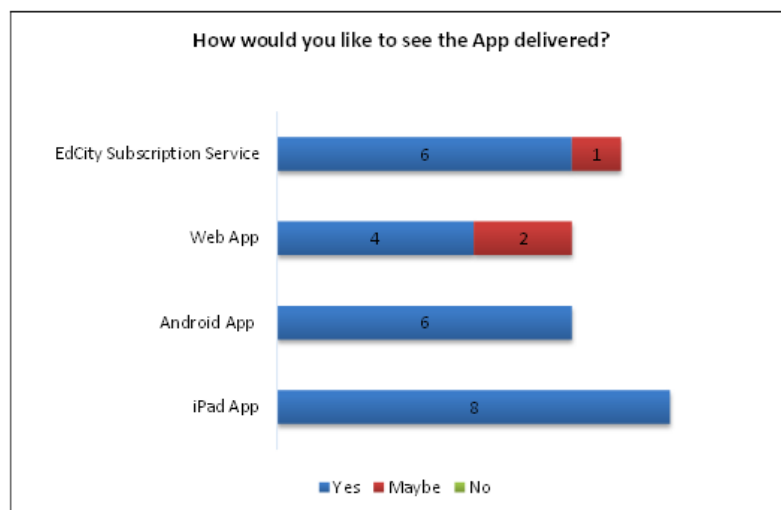
Fig. 7.13 Platform for content delivery

Most of the respondents favoured the iPad as the preferred platform for content delivery. This was closely followed by Android and the Education City proprietary solution respectively.

Overall, the results indicate that the teachers consider the tool to be usable and were willing to try it in their classes. The general attitude was positive and level of excitement modest.

III **Participants' comments and opinions**

In this part, the participants were asked to freely comment about tool. The responses are presented in Table 7.6

Table 7.6 Comments by teachers

Participant	Comments
1	The idea seems very good as solving word problems is something even ks1 children have to do. Just needs to look better and be easier to add the numbers as most times I needed to use two hands to get the operations box up, and try a couple of times. Unsure about cost but would imagine for most schools the cost would need to be as little as possible. Most Ks1 children would probably not benefit from this as they are encouraged to use physical equipment.
2	It's very handy I would want it to scribble over the top with an interactive pen to annotate, mainly as a teacher
3	It looks great from the outset and would quickly win teachers over. Appears it would save lots of time spent marking and analysing cohorts.
4	This feedback was from a secondary and a primary teacher. They loved the action of pulling numbers out. And also touching 2 numbers and choosing a operation. Would like to be able to touch an answer in its box and a floating number to then add, take away etc. that seemed a little more natural/intuitive Would like to not have a constant internet connection to use it.
5	Flow diagram confusing - better in lines eg 34 +4 Like visual pictures for younger year eg pulling down pictures of 6 sweets with the number underneath. Sentence building. Like it
6	Nil
7	Fab- practical and very good for visual learners to be able to manipulate the numbers and functions themselves in different ways and see the links between multi step problems as they work and see how one step in a problem can be incorrect ant then affect the overall outcome
8	Nil
9	Comment bank for feedback which is personalised
10	When you've clicked 2 numbers and entered an operation, you might like to swap the numbers around. Would be useful if you could mark them as they went along (as a teacher). Auto marking including 'you have used the wrong operation' and 'you have added these up incorrectly', here is the right answer. Pulling down numbers is great. would be really good to demo on a whiteboard. Not enough content to warrant a price cost. Can write questions yourself on a board and pull it down yourself using the whiteboard software. Need a lot more content already loaded. Needs to all link into a system for marking and feedback for whole class use at the same time.

As can be seen from the comments, many teachers reported positive attitudes regarding the potentials of the MuTAT for teaching and learning in their classroom. They found that the tree visualisation of the solution process to be useful and exciting. They were confident it could help in providing valuable feedback. However, some have expressed their concerns about the scope of questions the tool can handle and the initial problems encountered when trying to use the long-press multi-touch gesture.

7.4 Discussion

In this work an approach for capturing detailed problem-solving steps in arithmetic word problems was presented. Overall the MuTAT application successfully achieved the research aims. It is a novel study that demonstrated that solution steps can be captured from traces of interactive problem-solving actions in an assessment environment by using multi-touch gestures on integrated workspaces.

7.4.1 Summary of findings

The Multi-Touch Arithmetic Tool showed how solution steps could be captured from trace information obtained from interactive actions in problem-solving. It demonstrated that the products of an assessment task and the dependencies between the different parts of a solution effort as well as links to problem specification can be obtained and represented in a way that makes the thinking behind a solution visible. This is likely to provide valuable information that may be useful for detailed assessment of solution steps and the provision of more accurate feedback.

The empirical study which investigated the practicality of using the tool in classroom situations suggests that students aged 8 and 9 years are able to use the approach to solve problems. The performance scores and timings from the tool were comparable with those obtained from paper-and-pencil tests. A close examination of three student's MuTAT and paper-and-pencil solutions indicated that strategies were consistent across the two media. This suggests that the tool did not impede students' accuracy and efficiency when solving mathematics word problems. Also, the tool did not influence the students' strategy. While performances and strategies output were similar in both media, the process data output from the MuTAT application provided deeper insights into students' strategies than the paper responses. This can facilitate small-scale and large-scale assessments focused not only on answers but also

on strategies. Moreover, the approach provides students with increased opportunities to get relevant feedback to revise and improve their understanding and reasoning. Additionally, the approach provided opportunities for more detail scoring, intermediate steps such as the use of correct operations and intermediate answers can be scored independently of final answers. With this, assessors can assign proper credit for the evidence collected.

7.4.2 Implications

The study may contribute to understanding how gesture-based user interfaces are used to capture the thinking process and reveal strategies in elementary mathematics problem solving. While well-researched word problems were used in the study to allow strategies to be inferred MuTAT can be used where such strategic problems do not exist.

Computer-Aided Assessment and feedback generation can be extended beyond just providing comments and scores to the final product of elementary maths assessment tasks. Because consideration of the whole steps and actions during the problem-solving process provides useful information, teachers, parents and practitioners can use this approach to improve their feedback processes.

Some scholars have suggested that word problems may restrict children in the use of their conceptual knowledge (Gilmore and Bryant, 2006). Teachers and researchers can use the tool to examine how students understand the wording of a question.

7.4.3 Limitations

Three limitations may be related to data collection and the interpretation of results. A first limitation might be the constraint on the tool to present the stepwise information horizontally. Some students may feel better doing the sums vertically, especially when it involves a carry-over digit.

Another potential shortcoming in the study is the bias which may be introduced by the non-randomization of the MuTAT and paper-and-pencil test treatments on the participants. All the students started with the MuTAT application before crossing over to the paper-and-pencil paper test. This however, should not have significant effect on the overall outcome of the study.

A third potential limitation is related to design challenges that may result from the small size of the tablet screen (9.7-inches) that was used in the experiment. Interacting with objects

on a touch screen requires reasonable screen size. This constraint may affect problems that require three or more arithmetic operations.

7.5 Chapter summary

This chapter presented the evaluation studies carried out to examine the functionalities of the MuTAT in line with design requirements within the semi-automatic CAA framework. The most important result from this evaluative study is that it provided a proof of concept for the research approach implemented on the MuTAT. It showed that approach to be usable and comparable to conventional test formats. It also demonstrated the extended capability of the tool for process-level feedback. The students' and teachers' reactions to the system were positive as evidenced in the pilot and main studies. The technology is not perfect, but it provides opportunities for wider testing and future developmental efforts.

Chapter 8

Marking Assistant Evaluation

The design and development of the Marking Assistant tool implementing the case-based approach to marking and feedback in the semi-automatic assessment framework were presented in Chapter 6. This chapter presents the fourth study, which explores the feasibility of using the tool for marking and the study of its time-saving capabilities. The chapter presents the study design, study results and discussions of the results of the study.

8.1   Study 4: Marking Assistant evaluation

The goals of the this study are to:

1. Find out whether assessors will be able to use the Marking Assistant to score and provide feedback comments to students answers

2. Compare the times it takes to complete the marking with manual marking on paper

3. Observe the attitudes and behaviours of the assessors while marking

8.2   Study design

The main objective of the evaluation was to compare marking time on the application with that on paper. Undergraduate students (N=8) in Loughborough University were the participants in the study. The evaluation by the participants was performed on a voluntary basis (see Section 3.5.2). Each of the participants used the tool and paper independently after being briefed and

guided on how to use the MA. The details of this study and results of the observational study are described.

8.2.1 Participants

Eight participants were used in the study. There were four males and four females. They were all required to mark the responses of 20 students' on paper and the Marking Assistant tool. The students' solutions explicitly showed all the steps to the answer. The student solutions used were those obtained from the MuTAT study described in Section 7.2.

8.2.2 Questions

The study used four arithmetic word problems, solving the problems require two steps using two arithmetic operators. As with the students' solutions, the questions used were randomly selected from those used in the MuTAT study.The students'solutions were those from the MuTAT study described in Section 7.2. Participants were asked to mark four questions – two on each media. The first question to be marked on both media (tool and paper) will require the use of addition operators in both steps. The second question will require using both addition and subtraction operators. The lists of both question types are shown in Table 8.1.

Table 8.1 Questions used in study

Questions requiring addition only	Questions requiring both addition & subtraction
Question 1: Jason owned a factory that employs 53 workers. He hired another 7 workers. He then hired another 16 workers. How many workers are there at the factory altogether?[a]	**Question 2:** Sam has 21 books, he got 9 more and gave 13 to Owen. At the end, how many books did Sam have? [a]
Question 3: Sara has 8 sugar donuts. She also has 15 plain donuts and 32 jam donuts. How many donuts does Sara have altogether?[b]	**Question 4**: There are 24 books on a shelf, 11 more were added and then 6 taken away. At the end, how many books were there? [b]

(Adapted from (Carpenter and Moser, 1984) [a] and (Gilmore and Bryant, 2006)[b])

A comparative crossover experimental design was used. To minimise ordering effects different combinations of question type and marking media were created as shown in Table 8.2. The participants were randomly assigned to the four different combination categories.

Table 8.2 Order in which participants marked on paper and the Marking Assistant

Combination	Marking order	Tablet Question	Paper Question
1	Tablet, Paper, Paper,Tablet	1, 2	3, 4
2	Tablet, Paper, Paper, Tablet	3, 4	1, 2
3	Paper, Tablet, Tablet, Paper	1, 2	3, 4
4	Paper, Tablet, Tablet, Paper	3, 4	1, 2

Figure 8.1 shows how the students' solution were presented on paper for marking by the participants. This format was used so as not to burden them with too may items to deal with.

Fig. 8.1 Paper sample questions

Full details of the questions are presented in Appendix E.

8.2.3   Measures

The main measurement made was the time it took the participants to complete the marking of the questions on the two media used. On the Marking Assistant, the proportions of items marked manually and automatically were recorded.

8.3   Results

All the participants successfully completed scoring on paper and the Marking Assistant tool. The times it took the participants to complete the marking on both media are shown in Figure 8.2.
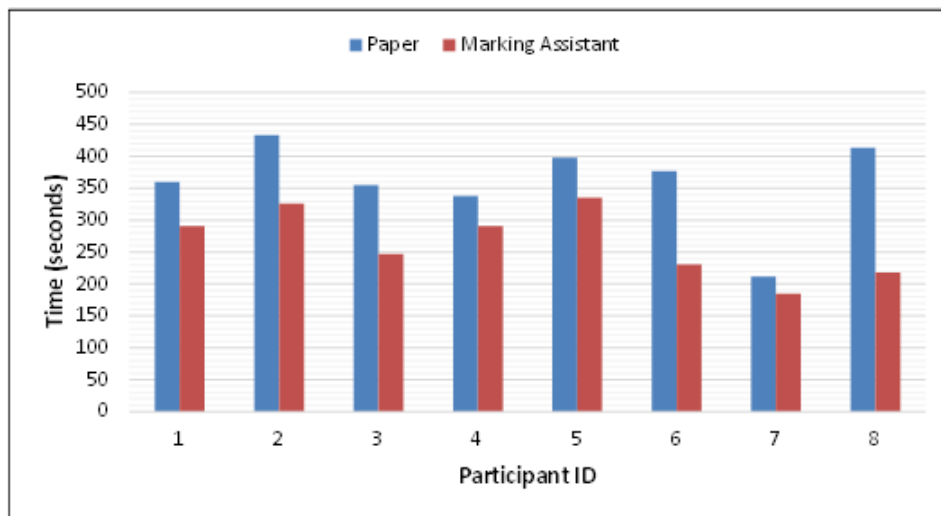
Fig. 8.2 Marking time on paper and the Marking Assistant

As can be seen from Figure 8.2, all the participants spent less time on the tool compared to paper. The mean time spent on the questions on the Marking Assistant was 265.5 seconds while 360.9 seconds was spent marking on paper. Using the t-test, this difference was significant $t(7) = 4.84$, $p = 0.00188$. The total marking time across both media for all the participants is shown in Figure 8.3. On the average about 26% of the marking time was saved by using the Marking Assistant.
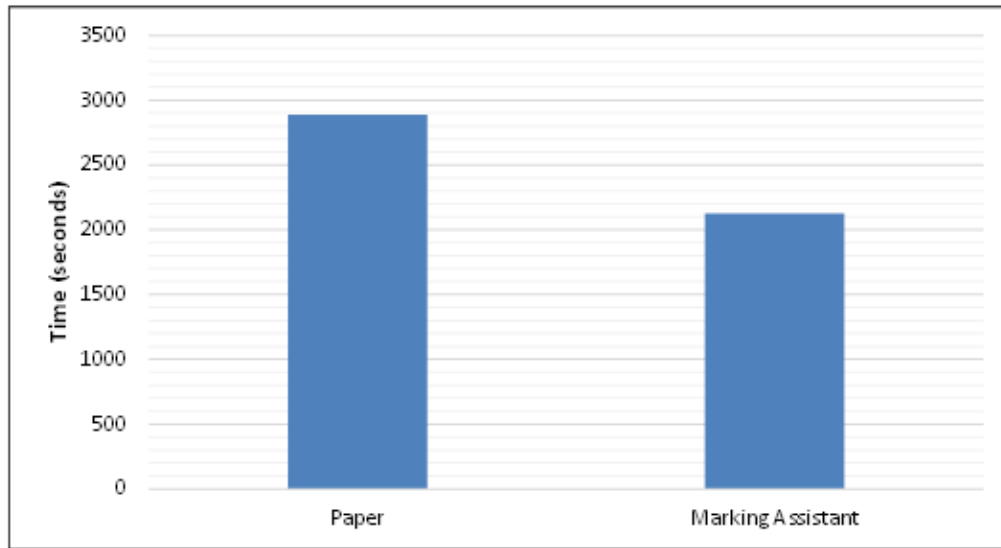
Fig. 8.3 Total marking time on paper and the Marking Assistant

The relative proportion of items manually marked and those marked automatically by the computer re-using the manually marked results is shown in Figure 8.4.
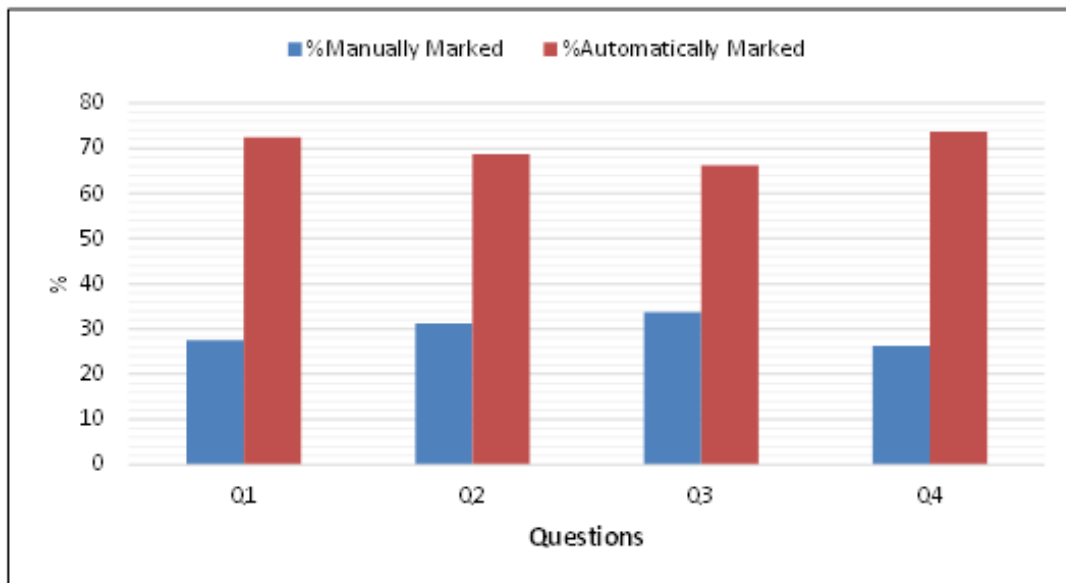
Fig. 8.4 Proportion of marking done manually and automatically

From the figure above it can be seen that in all the questions more items were marked automatically (70%) than manually (30%). This evidently enabled time to be saved on the Marking Assistant.

8.4   Discussion

This study suggests a semi-automatic assessment approach for CAA using the case-based reasoning methodology. It explored how marking efficiency and consistency may be improved while considering detailed problem-solving steps.

The first question in this study sought to determine was how a semi-automatic CAA may be designed to help assess and score more than final answers of a problem-solving effort. The approach adopted combined three techniques; capture of problem-solving steps, application of analytical rubric and case-based reasoning. The implementation of the assessment rubric and case-based-reasoning techniques were successfully carried out on the Marking Assistant tool described in Chapter 4. A two-step elementary mathematics problem was broken down into four assessable components to increase the depth of what is assessed. Because the tool automatically re-uses judgment features of the marker, consistency in the application of the judgements to problems with similar attributes is ensured.

As regards the potential efficiency of the system, the observational study provided useful insights. The results show that the Marking Assistant enabled significant time savings compared to marking done entirely manually. For the four components of the marking examined, up to 70% of the required assessment was done automatically. This suggests that the efficiency of an assessor may be improved by using as cases manually-assessed components. However, the observational study has some limitations. Although classroom test questions and solutions were used, undergraduate students were the participants in the marking exercises, rather than actual primary school teachers. Since the study objective was to obtain marking time from two mediums by the same individual, results from students with fewer marking experiences than actual classroom teachers helps demonstrate the potentials of the approach. The findings suggest several implications. The improved accuracy and consistency in marking ensure that correct and fair marks are giving to all students possibly resulting in more satisfied teaching staff, students, and administrators. The system assists students to monitor and reflect on their problem-solving processes and also in understanding how they were graded on a piece of work. The clear rubrics may also be used to communicate to parents the strengths and successes that students have demonstrated.

All assessment procedures have strengths and limitations. It is important to note that much work has to be done beforehand in the design of questions and software. Some authors have cautioned that the less writing a student does in an assessment task, the more work the assessor has do in creating the task (Brown et al., 2014; Bull and Mckenna, 2000). Sangwin (2013), also pointed out that a potential limitation in emerging semi-automatic CAA systems

is the loss of immediate feedback. This is because marking that requires the attention of a human assessor comes with an inevitable delay. Instant feedback is a big advantage of fully automatic systems. However, the gains of detailed and personalised feedback and the potential marking-time savings gained from the re-use of assessed items may allow assessors to respond quicker than with manual scoring and feedback methods.

8.5   Wider implications

In this study, assessment has been described as the process of gathering, recording, interpreting, using, and reporting information about students' progress and achievement in developing knowledge, skills and attitudes (see Section 2.1). Advances in computer technology create opportunities for improving the quality and efficiency of assessment practices for teachers and students beyond that obtained from traditional methods. There is the need to provide CAA environments that are engaging for students as they solve problems while providing them with opportunities to apply higher-order thinking skills. The detailed steps used in the process of problem solving should be captured. For teachers, there is the need to present them with opportunities to provide detailed and personalised marking and feedback comments. The preceding discussions have presented the concept, design, development and evaluation of tools based on the semi-automatic framework developed in the study. This section discusses how it can be used in other contexts and tasks.

To increase the breadth of skills assessed using CAA environments, a unique intermediate constraint question type was implemented. To actively engage students, the direct multi-touch technology was considered and applied. To support teachers in marking and feedback work, mechanisms for storing assessed work as cases and assessment objects were designed. The design was focused on reducing repetition by facilitating reuse based on similarities between assessed components. These principles can be employed in a variety of subjects and question types. Figure 8.5 summarises the operating principles.
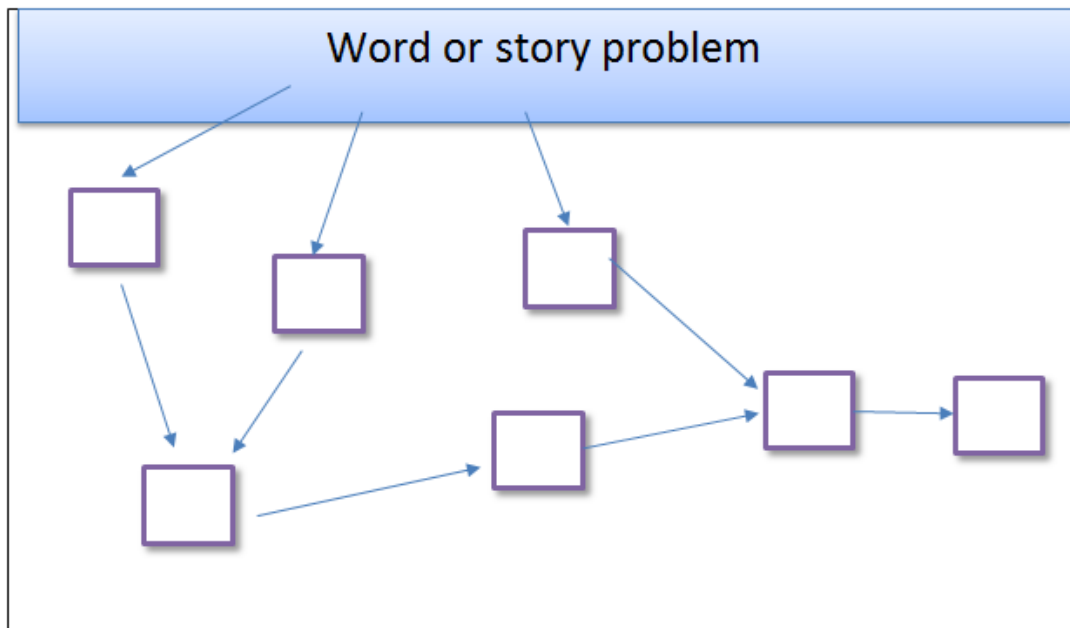
Fig. 8.5 Wider implications

The figure shows two main sections. The first is where the scenario, word or story problem is displayed. Some of the words/numbers/images in this section are designed to be selectable so they can be used to solve the problem. These intractable objects may be represented in circles or boxes, which may be linked by lines words or phrases that show the connection and special relationships between the objects. The second section is where the actual problem solving takes place using the objects selected from the word or story problem section. Here different gestures and interactions are used to create new objects, activate menus and enter answers. The diagram/solution so produced provides a visual illustration of the relationships between different components, concepts or ideas. It provides the story of the problem-solving process. This type of representation helps the teacher to see inside the student's thoughts. This information can give rich insights into what and how the students are learning and the connections they are making between concept and ideas. It is also very beneficial to the students. The process engages them in more meaningful expression of their learning.

This way for representing solution steps also enables teachers to mark and provide feedback on the work piece by piece rather than as a single element. From captured data. It is possible to analyse in great detail the processes and products of students' interactions with the tasks. The marks and comments provided can be reused as required to improve efficiency and consistency.

8.6 Chapter summary

This chapter evaluated the Marking Assistant. It presented the study designed to quantify the time savings that can result from the using the Marking Assistant compared to marking done entirely manually. Results from the study showed that 26% of the marking time was saved by using the tool. Furthermore the Marking Assistant provided the opportunities for consistent scoring and detailed feedback.

Chapter 9

Summary, Conclusions and Further Work

The aim of this research is to develop and evaluate a semi-automatic computer-aided assessment framework that can potentially extend and improve on current assessment practices by providing opportunities for detailed feedback. There are two main areas of focus: (1) assisting students to input process responses and product responses in a problem-solving task and (2) supporting teachers in assessment and feedback activities so as to be more consistent and efficient. The area of focus was on the use of the semi-automatic framework for arithmetic word problems. This thesis has reported on the design, development, and evaluation of a new semi-automatic framework with an emphasis on detailed feedback and minimising marking and feedback workloads. This chapter presents a summary of the main points of the thesis followed by conclusions, limitations, detailed list of contributions and suggestions for future work.

9.1   Thesis review and summary

There is substantial research to show that assessment design and feedback processes shape students' behaviour, learning and experience. Also, learning involves the active engagement of students in their own work. Over the years as the digital landscape advances, computers are increasingly used in teaching, learning and assessments. Many scholars have argued that computer technology has the potential to improve assessment and feedback practices beyond that possible with conventional paper-and-pencil formats. However, most CAA

practices have mainly focused on the evaluation of the product of an assessment exercises. Few studies have examined how the processes which lead to the product may be evaluated for detailed marking and feedback. This study explores how detailed marking and feedback may be provided for word problem questions in elementary mathematics. Discussions on the background of assessments, theories and arguments in literature were presented in Chapter 2. The research questions that guided the study are as follows:

1. How can a CAA assessment environment be designed to capture student's problem-solving processes?

2. Will the data captured lead to the detection of strategies and opportunities for detailed feedback in elementary mathematics word problems?

3. How can assessors score and provide feedback on students' responses consistently and efficiently?

This research has proposed a CAA semi-automatic framework that can assist students and teachers in assessments done with computers. The components of the framework and their interrelationships were described in Chapter 4. The basic principle of the semi-automatic assessment framework is drawn from the ideas - "use computers for what they are best for –i.e. computations, rich interactions, data capture and data representation", and let humans do what they do best, namely problem-solving and making informed judgements. These thoughts have been captured in the framework. The framework provided a platform to enable the integration of a number of the technologies for the assessment of students' work. Essentially, the semi-automatic framework is made up of three main components. The first is interactive problem-solving by students on a computer interface. This was discussed in Section 4.3 . The Multi-touch environment was chosen because it provided opportunities for direct and rich interactions. Also discussed in the section was a novel way for problem-solving that allowed inter-communication between the question elements and solution responses. Section 4.4 and 4.5 discussed the second component of the framework. This involves the use analytic rubrics to interpret, mark and provide feedback comments on the responses captured from students. Here the teacher or assessor uses clear and detailed rubrics to assess the various aspects of the work. A key idea from this part of the framework is the concept of Assessment Objects. These are segments of students' responses that may be assigned scores and feedback comments. These Assessment Objects are then reused and may be aggregated into large pieces to provide full and detailed representations of students' work. The third major framework component centres on the reuse of Assessment Objects. This was discussed in Section 4.6. The artificial intelligence method known as case-based reasoning was used to provide marking consistency and efficiency. The Assessment Objects complete with marks

and feedback comments are stored as cases in the case library and reused in the marking of similar students' responses. The benefit of this approach is that it delivers more efficient marking as the marker can browse sorted responses, apply one judgment which is then reused in multiple, identical responses. This way specific and detailed feedback provided in an Assessment Object need not be repeated in the marking exercise. This provides savings in marking time. The second benefit the approach provides is consistent marking. Errors and mistakes are reduced because all similar responses are assigned the same marks and feedback.

The functional requirements of the semi-automatic framework were designed and developed on two prototype tools. The Multi-Touch Arithmetic Tool provided the user-interface for the capture the students' solution steps. The design requirements and architecture of the user-interface and code implementations were discussed in Sections 5.2, 5.3 and 5.4. The Marking Assistant which supports marking and feedback activities was described in Section 6.2. Details of the marking process and the new way cases may be represented were outlined in Section 6.4. The Marking Assistant user-interface and architecture together with the implementation were discussed in Section 6.5.

The evaluation of the functions of the research prototypes was carried out in different studies. The first study was a pilot study conducted among University students to determine if the MuTAT prototype was useable for arithmetic work by undergraduate students. Details of this study were outlined in Section 7.1. The second study was the main evaluation study. This was carried out in two primary schools among Year 8 and 9 students. The full details of the study were presented in Section 7.2 . The results from the studies showed the MuTAT was usable for mathematics work. Particularly, it demonstrated the rich feedback capabilities of the approaches. Details of the results were outlined in Sections 7.3 and 7.4. In addition to studies conducted with students, a small survey was carried out among teachers after they had interacted with the MuTAT. Results show that the teachers recognised the potentials of the tool. The details are outlined in Section 7.5. The evaluation of the Marking Assistant prototype focused on the potential time gains that may be achieved when it is used to mark and provide feedback comments. Details of this study and the results were presented in Section 8.1 and 8.2.

9.2   Conclusions

A new semi-automatic framework for computer-aided assessment was developed and described. The implementation of the prototypes shows that the approaches described in the thesis to be feasible. The examples demonstrated how the approach aided the capture of detailed data from interactions on a computer user-interface. Problem-solving sequences can be monitored, allowing for the examination of how a student arrives at his or her answer. This provides information to educators on where mistakes or misunderstandings occur, potentially leading to the ability to make stronger instructional decisions. furthermore, it can allow for the extraction of patterns that correlate with varying levels of performances.

After the development of the proof-of-concept level prototypes, evaluation studies were carried out. The study which addressed questions relating to the comparability of computer and paper mathematics test scores, detection of arithmetic strategies and potential savings in marking time. The results showed the students were able to use the approach for problem solving, distinct arithmetic strategies were detected. It was possible to provide detailed feedback on the process of problem solving and about 26% of marking time may be saved.

On the whole, the students and teachers were enthusiastic about the performance of the tools. This feedback and the successful implementation of the semi-automatic framework provided initial "proof-of-concept" – level validation. We intend to use the resulting applications to plan continued developmental efforts.

9.3   Implications

This research provides tools for the exploration of semi-automatic CAA. In recent years, there has been an evolution of massive open online courses (MOOCs). MOOCs are delivered via the web and aims at having unlimited student participation. Some courses have over 10,000 students enrolled. Assessing these students beyond the typical selected response type questions has been challenging. The semi-automatic assessment framework discussed in this study provides a new way for thinking about assessing large numbers of students with detailed marking and feedback capabilities.

9.4   Limitations

Although the research has reached its aims, there were some unavoidable limitations. First, because of the time limit, this research was conducted mainly using word problems which involved addition and subtraction operators. Questions involving more operators and concepts could further expose feedback possibilities of the framework. Second, the study did not set out to investigate the best way to implement rich touch interactions for the capture of students' solution steps. The use of more intuitive multi-touch gestures may enhance the user experience and better enable the capture of students' responses. Third, the evaluation studies were conducted with only a small size population. Therefore to generalise results for larger groups, the study should have involved more participants at different levels.

Every assessment process contains errors and random chance components. As Alder (2002) noted even the metre is not a perfect measure. Some thoughts on limits the framework are outlined below.

- Perhaps only a limited range of the student's thinking is shown through physical interactions with the computer.

- While touch interactions help to capture solution processes, it is difficult to capture accompanying non-text responses.

- For complex performances, the computer-captured data can be difficult to interpret.

- Some scholars argue that interactivity can spoil some tasks because it encourages some students to engage in trial-and-error working rather than think through an analysis.

9.5   Outline of contributions

This section reiterates and provides full details of the contributions of the thesis summarised in Chapter 1. Also, it provides internal references to the related chapter for each contribution.

The contributions of this thesis are:

- Through the application of computer-aided assessment in elementary mathematics domain, it has contributed to an enhanced understanding of semi-automatic assessment. The discussion on the semi-automatic approach can be found in Chapter 4.

- The new assessment framework proposed provided an integration of a variety of methods which may be used to increase automation of assessment in mathematics. The complete system view of the framework can be found in Section 4.2.

- The interactive problem-solving part of the framework on a multi-touch environment provided a new way of delivering assessment for intermediate constrained question types (discussed in Section 2.5). This part of the framework achieves Objective 1 identified in Chapter 1. The unique design implemented on the user-interface aided the detailed capture of problem-solving steps.

- The empirical studies contribute to providing support on the feasibility of capturing problem-solving process in normal school classrooms. It showed that students' performances were not affected by the new approach. Students' performance results when the new approach was used compared favourably with traditional paper-and-pencil test approaches. The empirical studies can be found in Chapter 7. The studies achieve objective 2 of the research.

- A new way for defining cases for as assessment objects. These are self-contained segments of students' responses that may be assigned scores and feedback comments. The assessment objects may be reused at varying levels of granularity. The application of artificial intelligence methods on the case definitions for different student submissions contributes to ensuring consistency and efficiency in marking. The new case definition was explained in Section 6.4. The evaluation of savings in marking time was outlined in Chapter 8. This achieves the third research objective.

- The novel partial marking style using analytic rubrics and detailed feedback comments provides a way for providing personalised and detailed feedback to students.

- A new way to obtaining distracters when setting multiple-choice questions. The data captured for the study describe in Section 7.2 was used in a study for determining how to obtain good distracters for MCQ types. This was reported in masters thesis Automatic Multiple Choice Questions Generator (Mazin, 2015).

9.6 Further Work

This work has revealed a number of areas where further research can yield potential contributions to knowledge and practise.

- What is now needed is research into using the framework for other types of questions and domains of study.

- Further investigations on the best and most intuitive gestures for arithmetic work may be carried out.

- More research is required to determine how teachers and researchers may use the MuTAT to examine how students understand the wordings of a question by their interactions with the different parts. It may be useful in mathematics education in studies looking improving to improve the writing of questions.

- The research within the thesis has been specific to CAA in elementary mathematics. However, recommendations can be made to assessment areas in other domains and question tasks.

- There are more extensions to the CBR search algorithm that may be explored. Two additional ideas include support for distance-weighted contribution for the k-most similar instances to the prediction and more advanced data tree-based structures for searching for similar instances.

- It would be interesting to assess how the multi-touch environment enables collaborative problem-solving. Studies on how students solve problems as a group may be carried out.

9.7   Closure

This research has been an enlightening experience for the author, and it is hoped that it has provided some insights into the salient characteristics of semi-automatic assessment and the role that modern technology could play in facilitating the whole assessment exercise. The approach described is just the beginning of what is hoped will be an even greater journey towards an extensive system that will meet the expectations of students, teachers and administrators for rich feedback and efficient marking.

References

Aamodt, A. and Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, 7(1):39–59.

Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for english language learners. *Applied Measurement in Education*, 27(4):261–272.

AERA (1999). *Standards for educational and psychological testing*. American Psychological Association, and National Council on Measurement in Education, Washington, DC.

Ahoniemi, T. and Reinikainen, T. (2006). Aloha-a grading tool for semi-automatic assessment of mass programming courses. In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006*, pages 139–140. ACM.

Ainsworth, L. and Viegut, D. (2006). *Common Formative Assessments: How to Connect Standards-Based Instruction and Assessment*. SAGE publications.

Airasian, P. W. (1997). *Classroom Assessment. Third Edition.* McGraw Hill,.

Al-Smadi, M. and Gütl, C. (2008). Past, Present and Future of e-Assessment: Towards a Flexible e-Assessment System. In *ICL2008, Villach , Austria*.

Alder, K. (2002). *The Measure of All Things: The Seven-year Odyssey and Hidden Error that Transformed the World*. Simon and Schuster.

Allan, A. (2013). *Learning iOS Programming: From Xcode to App Store*. O'Reilly Media, Inc.

Allen, S., Graupera, V., and Lundrigan, L. (2010). *Pro smartphone cross-platform development: iPhone, blackberry, windows mobile and android development and distribution*. Apress.

Anderson, L. W. and Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Longman, New York.

Anderson, W. L., Mitchell, S. M., and Osgood, M. P. (2008). Gauging the Gaps in Student Problem-Solving Skills: Assessment of Individual and Group Use of Problem-Solving Strategies Using Online Discussions. *CBE Life Sciences Education*, 7(2):254–262.

Apple Inc. (2015). ios human interface guidelines. Retrieved August 2, 2016, from https://developer.apple.com/ios/human-interface-guidelines/.

Arnett, T. (2013). Will computers replace teachers. *Clayton Christensen Institute for Disruptive Innovation*.

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., and Yoshida, C. (2009). Cognitive developmental robotics: a survey. *IEEE Transactions on Autonomous Mental Development*, 1(1):12–34.

Ashton, H. S., Beevers, C. E., Korabinski, A. A., and Youngson, M. A. (2006). Incorporating partial credit in computer-aided assessment of Mathematics in secondary education. *British Journal of Educational Technology*, 37(1):93–119.

Azevedo, R. and Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition–implications for the design of computer-based scaffolds. *Instructional Science*, 33(5):367–379.

Bailey, R. and Garner, M. (2010). Is the feedback in higher education assessment worth the paper it is written on? Teachers' reflections on their practices. *Teaching in Higher Education*, 15(2):187–198.

Bailly, G., Müller, J., and Lecolinet, E. (2012). Design and evaluation of finger-count interaction: Combining multitouch gestures and menus. *International Journal of Human-Computer Studies*, 70(10):673–689.

Baker, E. and Mayer, R. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior*, 15(3–4):269–282.

Baker, E. L., O'Neil, H. F., and Linn, R. L. (1994). Introduction: Policy and validity prospects for performance-based assessment. *Journal for the Education of the Gifted*, 17(4):331–353.

Ballera, M., Lukandu, I. A., and Radwan, A. (2015). Improving learning throughput in e-learning using interactive-cognitive based assessment. *International Journal of E-learning and Educational Technologies in Digital media (IJEETDM)*, 1(1):32–49.

Bank, C. (2015). Gestures and animations: The pillars of mobile design. Retrieved August 12, 2016, from https://uxmag.com/articles/gestures-animations-the-pillars-of-mobile-design.

Bara, F., Gentaz, E., Colé, P., and Sprenger-Charolles, L. (2004). The visuo-haptic and haptic exploration of letters increases the kindergarten-children's understanding of the alphabetic principle. *Cognitive Development*, 19(3):433–449.

Barnes, K., Marateo, R. C., and Ferris, S. P. (2007). Teaching and learning with the net generation. *Innovate: Journal of Online Education*, 3(4):1.

Baroody, A. J. and Gannon, K. E. (1984). The Development of the Commutativity Principle and Economical Addition Strategies. *Cognition and Instruction*, 1(3):321–339.

Baroody, A. J. and Ginsburg, H. (1986). The relationship between initial meaningful and mechanical knowledge of arithmetic. In *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., and Ruppert, J. A. (2003). Social Embodiment. In *Psychology of Learning and Motivation*, volume Volume 43, pages 43–92. Academic Press.

Batmaz, F., Stone, R., and Hinde, C. J. (2009). Personal Feedback with Semi-Automated Assessment Tool for Conceptual Database Model. In *10th Annual Conference of HEA-ICS*, pages 115–119, University of Kent, Canterbury UK.

Beevers, C. E. and Paterson, J. S. (2003). Automatic Assessment of Problem-Solving Skills in Mathematics. *Active Learning in Higher Education*, 4(2):127–144.

Bennett, R. E. (1991). On the meanings of constructed response. *ETS Research Report Series*, 1991(2):i–46.

Bennett, R. E. (2011). Automated Scoring of Constructed-Response Literacy and Mathematics Items. Retrieved August 2, 2016, from http://www.ets.org/s/k12/pdf/k12_commonassess_automated_scoring_math.pdf.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., and Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning and Assessment*, 6(9).

Bennett, R. E., Persky, H., Weiss, A. R., and Jenkins, F. (2007). Problem Solving in Technology-Rich Environments. A Report from the NAEP Technology-Based Assessment Project, Research and Development Series. NCES 2007-466. *National Center for Education Statistics*.

Berger, D. and Dreher, H. (2011). *Supporting tool for moderation in the grading process of summative assessments: Design and prototype of a software tool for moderation and assessment with variable rubrics*. Graz University of Technology, Graz, Austria.

Bergmann, R., Kolodner, J., and Plaza, E. (2005). Representation in case-based reasoning. *The Knowledge Engineering Review*, 20(03):209–213.

Bescherer, C., Herding, D., Kortenkamp, U., Muller, W., and Zimmermann, M. (2011). *E-Learning Tools with Intelligent Assessment and Feedback for Mathematics Study*. Idea Group Inc (IGI).

bettshow.com (2015). Bettshow. Retrieved August 2, 2016, from http://www.bettshow.com/.

Bijou, S. W., Birnbrauer, J. S., Kidder, J. D., and Tague, C. (1966). Programmed instruction as an approach to teaching of reading, writing, and arithmetic to retarded children. *The Psychological Record*, 16(4):505.

Black, P. and Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, 21(1):5–31.

Black, P. and William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals* . Longman Group.

Bloom, B. S. (1984). *Taxonomy of Educational Objectives Book 1: Cognitive Domain.* Addison Wesley Publishing Company, 2nd edition edition.

Borthwick, A. and Harcourt-Heath, M. (2010). Calculating: What can year 5 children do? *Society for Research into Learning Mathematics*, 30(3):13.

Broadfoot, P. (2007). *An introduction to assessment.* Continuum Intl Pub Group.

Brown, G., Bull, J., and Pendlebury, M. (1997). *Assessing student learning in higher education.* Routledge.

Brown, G., Irving, E., and Keegan, P. (2014). *An Introduction to Educational Assessment, Measurement and Evaluation.* Pearson Education New Zealand Limited.

Brown, G. T. L., Glasswell, K., and Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, 9(2):105–121.

Bruner, J. S. (1966). *The Process of Education.* Harvard University Press, Cambridge.

Bryman, A. (2012). *Social Research Methods.* Oxford University Press Oxford.

Bryman, A. and Bell, E. (2012). *Business Research Methods.* Oxford University Press Oxford.

Bull, J. and Mckenna, C. (2000). Computer-assisted Assessment Centre (TLTP3) Update. In *Computer-assisted assessment centre (TLTP3) update, in: M. Danson (Ed.)*, Loughborough University.

Bull, J. and McKenna, C. (2004). *Blueprint for computer-assisted assessment.* Routledge-Falmer, London.

Bunderson, C. V., Inouye, D. K., and Olsen, J. B. (1988). The Four Generations of Computerized Educational Measurement. *ETS Research Report Series*, 1988(1):i–148.

Burrows, S. and Shortis, M. (2011). An Evaluation of Semi-Automated, Collaborative Marking and Feedback Systems: Academic Staff Perspectives. *Australasian Journal of Educational Technology*, 27(7):1135–1154.

Burstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing.

Butcher, P. G. and Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, 55(2):489–499.

Byrnes, J. P. and Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology*, 27(5):777–786.

Campbell, A. (2005). Application of ict and rubrics to the assessment process where professional judgement is involved: the features of an e-marking tool. *Assessment & Evaluation in Higher Education*, 30(5):529–537.

Carpenter, T. P. (1981). *Results from the Second Mathematics Assessment of the National Assessment of Educational Progress*. ERIC.

Carpenter, T. P. and Moser, J. M. (1984). The Acquisition of Addition and Subtraction Concepts in Grades One through Three. *Journal for Research in Mathematics Education*, 15(3):179–202.

Carroll, J. M. (2010). Conceptualizing a possible discipline of human–computer interaction. *Interacting with Computers*, 22(1):3–12.

Ceobanu, C. and Boncu, Ş. (2014). The challenges of the mobile technology in the young adult education. *Procedia-Social and Behavioral Sciences*, 142:647–652.

Chan, M. S. and Black, J. B. (2006). Direct-manipulation animation: Incorporating the haptic channel in the learning process to support middle school students in science learning and mental model acquisition. In *Proceedings of the 7th international conference on Learning sciences*, pages 64–70. International Society of the Learning Sciences.

Chandler, P. and Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, 8(4):293–332.

Charland, A. and Leroux, B. (2011). Mobile application development: web vs. native. *Commun. ACM*, 54(5):49–53.

Charles, R., Lester, Frank, and O'Daffer, Phares (1987). *How To Evaluate Progress in Problem Solving*. National Council of Teachers of Mathematics, 1906 Association Drive, Reston, VA 22901.

Chien, C.-F., Lin, K.-Y., and Yu, A. P.-I. (2014). User-experience of tablet operating system: An experimental investigation of windows 8, ios 6, and android 4.2. *Computers & Industrial Engineering*, 73:75–84.

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior*, 28(5):1580–1586.

Chung, G. K. W. K. and Baker, E. L. (2003). An Exploratory Study to Examine the Feasibility of Measuring Problem-Solving Processes Using a Click-Through Interface. *The Journal of Technology, Learning and Assessment*, 2(2).

Clarke, B., Svanaes, S., and Zimmermann, S. (2013). One-to-one tablets in secondary schools: an evaluation study. *Tablets for Schools*.

Clements, D. H. (2000). From exercises and tasks to problems and projects: Unique contributions of computers to innovative mathematics education. *The Journal of Mathematical Behavior*, 19(1):9–47.

Cocks, A. J. (2006). The ethical maze finding an inclusive path towards gaining children's agreement to research participation. *Childhood*, 13(2):247–266.

Conole, G. and Dyke, M. (2004). What are the affordances of information and communication technologies? *Association for Learning Technology Journal*, 12(2):113–124.

Conole, G. and Warburton, B. (2005). A review of computer-assisted assessment. *ALT-J*, 13(1):17–31.

Cooper, P. A. (1993). Paradigm shifts in designed instruction: From behaviorism to cognitivism to constructivism. *Educational technology*, 33(5):12–19.

Cowan, R. and Renton, M. (1996). Do They Know What They Are Doing? Children's Use of Economical Addition Strategies and Knowledge of Commutativity. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 16(4):407–20.

Creswell, J. W. (2013). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. SAGE Publications.

Dahlman, C. (2007). Technology, globalization, and international competitiveness: Challenges for developing countries. *asdf*, page 29.

Darling-Hammond, L. and Sykes, G. (2003). Wanted: A national manpower policy for education. *Denver, CO: Education Commission of the States*.

Davies, S. (2010). Effective Assessment in a Digital Age: A guide to technology-enhanced assessment and feedback. *JISC 2009*, pages 1–64.

Debuse, J. C. and Lawley, M. (2016). Benefits and drawbacks of computer-based assessment and feedback systems: Student and educator perspectives. *British Journal of Educational Technology*, 47(2):294–301.

Delen, E. (2015). Enhancing a computer-based testing environment with optimum item response time. *Eurasia Journal of Mathematics, Science & Technology Education*, 11(6):1457–1472.

Denton, P. (2003). Enhancing student learning using electronic feedback. *Exchange*, 4:23–24.

Denzin, N. K. and Lincoln, Y. S. (2005). *The SAGE Handbook of Qualitative Research*. SAGE Publications.

Devore, J. L. (2015). *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

Dewey, J. (2007). *Experience and education*. Simon and Schuster.

Dienes, Z. P. (1971). *Approach to modern mathematics*. Herder and Herder, New York.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Dolan, D., Williamson, J., and Muri, M. (2012). *Mathematics Activities for Elementary School Teachers, Problem Solving Approach to Mathematics*. Pearson, 11 edition.

Don, L. and Smith, S. P. (2010). Applying bimanual interaction principles to text input on multi-touch surfaces and tabletops. In *ACM International Conference on Interactive Tabletops and Surfaces*, pages 253–254. ACM.

Downing, S. M. and Haladyna, T. M. (2006). *Handbook of test development.* Lawrence Erlbaum Associates Publishers.

Dubé, L. and Paré, G. (2003). Rigor in information systems positivist case research: current practices, trends, and recommendations. *MIS quarterly*, pages 597–636.

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice*, volume 373. Springer Science & Business Media.

Elliott, B. (2008). E-pedagogy and e-assessment. In *12th International Computer Assisted Assessment Conference : Proceedings of the Conference*. © Loughborough University.

Ertmer, P. A. and Newby, T. J. (1993). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance improvement quarterly*, 6(4):50–72.

Feldman, M. S., Bell, J., and Berger, M. T. (2004). *Gaining access: A practical and theoretical guide for qualitative researchers*. Rowman Altamira.

Firestone, W. A., Fitz, J., and Broadfoot, P. (1999). Power, Learning, and Legitimation: Assessment Implementation Across Levels in the United States and the United Kingdom. *American Educational Research Journal*, 36(4):759–793.

Fitts, P. M. and Peterson, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology*, 67(2):103–112.

Flanagan, D. (2006). *JavaScript: the definitive guide*. " O'Reilly Media, Inc.".

Flewitt*, R. (2005). Conducting research with young children: Some ethical considerations. *Early child development and care*, 175(6):553–565.

Flick, U. (2015). *Introducing research methodology: A beginner's guide to doing a research project*. Sage.

Framework (2016). Businessdictionary.com. Retrieved August 12, 2016, from http://www.businessdictionary.com/definition/framework.html.

Fyfe, E. R., Rittle Johnson, B., and Decaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: prior knowledge matters. *Journal of Educational Psychology*, 104(4):1094.

Gagne, R. (1980). *The conditions of learning*. Rinehart, & Winston., Holt,New York.

Galliers, R. (1992). *Information systems research: Issues, methods and practical guidelines*. Blackwell Scientific.

Gallup (2013). Americans grade math as the most valuable school subject. Retrieved August 2, 2016, from http://www.gallup.com/poll/164249/americans-grade-math-valuable-school-subject.aspx.

García-Peñalvo, F. J. (2008). *Advances in E-Learning: Experiences and Methodologies: Experiences and Methodologies*. IGI Global.

Garfield, J. . B. (1994). Beyond testing and grading: Using assessment to improve student learning. *Journal of Statistics Education*, 2(1):1–11.

gartner.com (2014). Gartner says worldwide tablet sales grew 68 percent in 2013, with android capturing 62 percent of the market. Retrieved August 2, 2016, from http://www.gartner.com/newsroom/id/2674215.

Gibbs, G. and Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and teaching in higher education*, 1(1):3–31.

Gilmore, C. K. and Bryant, P. (2006). Individual differences in children's understanding of inversion and arithmetical skill. *The British journal of educational psychology*, 76(Pt 2):309–331.

Glasersfeld, E. v. (1989). Constructivism in education. *The international encyclopedia of education*, 1:162–163.

Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., and Kaschak, M. P. (2004). Activity and Imagined Activity Can Enhance Young Children's Reading Comprehension. *Journal of Educational Psychology*, 96(3):424–436.

Goldkuhl, G. (2012). Pragmatism vs interpretivism in qualitative information systems research. *European Journal of Information Systems*, 21(2):135–146.

Goldman, S. R. (1989). Strategy Instruction in Mathematics. *Learning Disability Quarterly*, 12(1):43–55.

Goundar, S. (2011). What is the potential impact of using mobile devices in education. In *Proceedings of SIG GlobDev Fourth Annual Workshop*.

Gravemeijer, K. (1997). Solving word problems: A case of modelling? *Learning and Instruction*, 7(4):389–397.

Gray, E. M. and Tall, D. O. (1994). Duality, Ambiguity, and Flexibility: A "Proceptual" View of Simple Arithmetic. *Journal for Research in Mathematics Education*, 25(2):116–140.

Gray, L., Thomas, N., and Lewis, L. (2010). Teachers' use of educational technology in us public schools: 2009(nces 2010-040). *National Center for Education Statistics*.

Green, L. (1999). Bridging the digital divide in our schools-achieving technology equity7 for all students. *Education Market Research. Education Week's*.

Gregory, R. and Muntermann, J. (2011). Theorizing in design science research: inductive versus deductive approaches. In *Thirty second international conference on information systems, Shanghai, China*. Shanghai.

Grønli, T.-M., Hansen, J., Ghinea, G., and Younas, M. (2014). Mobile application platform heterogeneity: Android vs windows phone vs ios vs firefox os. In *2014 IEEE 28th International Conference on Advanced Information Networking and Applications*, pages 635–641. IEEE.

Hallman, G., Paley, I., Han, I., and Black, J. (2009). Possibilities of haptic feedback simulation for physics learning. In *Proceedings of world conference on educational multimedia, hypermedia and telecommunications*, pages 3597–3602.

Hanna, G. S. and Dettmer, P. A. (2004). *Assessment for effective teaching using context-adaptive planning.* Pearson, New York.

Hansen, T. E., Hourcade, J. P., Virbel, M., Patali, S., and Serra, T. (2009). PyMT: a post-WIMP multi-touch user interface toolkit. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, ITS '09, pages 17–24, New York, NY, USA. ACM.

Harrington, P. (2012). *Machine learning in action.* Manning.

Harry, T. and John, P. (1998). *Investigating Formative Assessment: Teaching, Learning and Assessment in the Classroom.* McGraw-Hill Education (UK).

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* Routledge.

Hattie, J. and Brown, G. (2012). Assessment and evaluation. In *C.Rubie-Davies (Ed.) Educational pyschology: Concepts, research and challenges*, pages 102–117.

Hattie, J. and Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

Hayes, B. G. (1999). Where's the Data: Is Multimedia Instruction Effective in Training Counselors? *Journal of Technology in Counseling*, 1(1):n1.

Hegarty, M., Mayer, R. E., and Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, 87(1):18–32.

Heinrich, P. (2011). The iPad as a Tool for Education: A study of the introduction of iPads at Longfield Academy, Kent. *Nottingham: NAACE: The ICT Association.*

Herding, D. and Schroeder, U. (2012). Using Capture & Replay for Semi-automatic Assessment. *International Journal of e-Assessment*, 2:1.

Herl, H. E., Baker, E. L., and Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of us history knowledge. *The Journal of Educational Research*, 89(4):206–218.

Hevner, A. and Chatterjee, S. (2010). *Design science research in information systems.* Springer.

Hiebert, J. (1984). Children's Mathematics Learning: The Struggle to Link Form and Understanding. *The Elementary School Journal*, 84(5):496.

Hogan, J. M., Carlson, J. G., and Dua, J. (2002). Stressors and Stress Reactions Among University Personnel. *International Journal of Stress Management*, 9(4):289–310.

Hollingsworth, J. (1960). Automatic graders for programming classes. *Commun. ACM*, 3(10):528–529.

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, 48(1):71–74.

Hosseini, M., Abidin, M. J. Z., and Baghdarnia, M. (2014). Comparability of test results of computer based tests (cbt) and paper and pencil tests (ppt) among english language learners in iran. *Procedia-Social and Behavioral Sciences*, 98:659–667.

Hüllermeier, E. and Minor, M. (2015). *Case-Based Reasoning Research and Development: 23rd International Conference, ICCBR 2015, Frankfurt Am Main, Germany, September 28-30, 2015. Proceedings*, volume 9343. Springer.

Humanreaders.org (2013). Professionals against machine scoring of student essays in high-stakes assessment. Retrieved August 2, 2016, from http://humanreaders.org/petition/.

Humanreaders.org (2016). Professionals against machine scoring of student essays in high-stakes assessment. Retrieved August 3, 2013, from http://humanreaders.org/petition/research_findings.htm.

Hunt, T. (2012). Computer-marked assessment in Moodle: Past, present and future. In *CAA 2012 International Conference, Southampton*.

Hutton, D. (2013). The quest for artificial intelligence: A history of ideas and achievements. *Kybernetes*.

Iivari, J. and Venable, J. (2009). Action research and design science research–seemingly similar but decisively dissimilar. In *17th European Conference on Information Systems*, pages 1–13.

Imbo, I. and Vandierendonck, A. (2008). Effects of problem size, operation, and working-memory span on simple-arithmetic strategies: differences between children and adults? *Psychological research*, 72(3):331–346.

Jacko, J. A. (2012). *Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications*. CRC press.

Jackson, P. (1998). *Introduction to Expert Systems*. Addison-Wesley Longman Publishing Co., Inc.

Jacucci, G., Morrison, A., Richard, G. T., Kleimola, J., Peltonen, P., Parisi, L., and Laitinen, T. (2010). Worlds of information: designing for engagement at a public multi-touch display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2267–2276. ACM.

Jang, S.-J. (2010). Integrating the interactive whiteboard and peer coaching to develop the tpack of secondary science teachers. *Computers & Education*, 55(4):1744–1751.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, 33(4):410–422.

Jiao, X., Deng, H., and Wang, F. (2010). An Investigation of Two-Handed Manipulation and Related Techniques in Multi-touch Interaction. In *2010 International Conference on Machine Vision and Human-Machine Interface (MVHI)*, pages 565 –568.

Johannesen, T. and Eide, E. M. (2000). The role of the teacher in the age of technology: Will the role change with use of information and communication technology in education? *European Journal of Open, Distance and E-learning*, 3(2).

Jonassen, D. H. (1999). *Designing constructivist learning environments. In C. M. Reigeluth (Ed.)*, volume II of *Instructional design theories and models: A new paradigm of instructional theory*. Lawrence Erlbaum Associates, Mahwah, NJ.

Jonassen, D. H. (2000). Toward a Design Theory of Problem Solving. *Educational Technology Research and Development*, 48(4):63–85.

Jones, M. (1985). Applications of artificial intelligence within education. *Computers & mathematics with applications*, 11(5):517–526.

Jonsson, A. and Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2):130–144.

Jordan, S. (2013). E-assessment: Past, present and future. *New Directions*, 9(1):87–106.

Jordan, S. and Mitchell, T. (2009). e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, 40(2):371–385.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kakkonen, T. and Sutinen, E. (2009). Semi-automatic Assessment Model of Student Texts - Pedagogical Foundations. In *Proceedings of the 12th International Conference on Interactive Computer-aided Learning*, Villach, Austria.

Kammer, D., Wojdziak, J., Keck, M., Groh, R., and Taranko, S. (2010). Towards a formalization of multi-touch gestures. In *ACM International Conference on Interactive Tabletops and Surfaces*, pages 49–58. ACM.

Kaplan, B., Maxwell, J., Anderson, J., Aydin, C., and Jay, S. (1994). Evaluating health care information systems: Methods and applications. *Qualitative Research Methods for Evaluating Computer Information Systems*.

Kolodner, J. (1993). *Case-based reasoning*. Morgan Kaufmann.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lawson, D. (2012). Computer-aided assessment in mathematics: Panacea or propaganda? *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, 9(1).

Leake, D. B. (1996). *Case-Based Reasoning: Experiences, Lessons and Future Directions*. MIT Press, Cambridge, MA, USA, 1st edition.

Lembke, E. S., Hampton, D., and Beyers, S. J. (2012). Response to intervention in mathematics: Critical elements. *Psychology in the Schools*, 49(3):257–272.

Lesh, R. and Behr, M. (1987). Representations and translations among representations in mathematics learning and problem solving. In *Problems of representation in the teaching and learning of mathematics*, pages 33–40.

Levin, T. and Wadmany, R. (2008). Teachers' views on factors affecting effective integration of information technology in the classroom: Developmental scenery. *Journal of Technology and Teacher Education*, 16(2):233.

Lewis, J. (2006). Usability testing. In *Handbook of Human Factors and Ergonomics*, pages 1275–1316. John Wiley, New York.

Li, K. C., Wong, T.-L., Cheung, S. K., Lam, J., and Ng, K. K. (2015). *Technology in Education. Transforming Educational Practices with Technology: International Conference, ICTE 2014, Hong Kong, China, July 2-4, 2014. Revised Selected Papers*, volume 494. Springer.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. *Educational Testing Service R & D Connections*, 11:1–8.

Maple, T. (2016). Automated online testing & assessment. Retrieved August 2, 2016, from http://www.maplesoft.com/products/mapleta.

Mason, R. L., Gunst, R. F., and Hess, J. L. (2003). *Statistical design and analysis of experiments: with applications to engineering and science*, volume 474. John Wiley & Sons.

Mayer, R. and Wittrock, M. C. (1996). Problem-solving transfer. In *Handbook of educational psychology*, pages 47–62.

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science*, 26(1-2):49–63.

Mazin, A.-S. (2015). *Automatic Multiple Choice Questions Generator*. Masters Thesis, Loughborough University.

Mazza, R. and Milani, C. (2005). Exploring usage analysis in learning systems: Gaining insights from visualisations. In *Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education*, pages 65–72.

McAlpine, M. (2002). *Principles of assessment*. Robert Clark Center for Technological Education University of Glasgow.

McCloskey, M. (2007). Quantitative literacy and developmental dyscalculias. *Why is math so hard for some children*, pages 415–429.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, 39(3):299–312.

McNeil, N. M., Rittle-Johnson, B., Hattikudur, S., and Petersen, L. A. (2010). Continuity in Representation Between Children and Adults: Arithmetic Knowledge Hinders Undergraduates' Algebraic Problem Solving. *Journal of Cognition and Development*, 11(4):437–457.

Mercier, E. M. and Higgins, S. E. (2013). Collaborative learning with multi-touch technology: Developing adaptive expertise. *Learning and Instruction*, 25:13–23.

Messick, S. (1987). Validity. *ETS Research Report Series*, 1987(2):i–208.

Mills, C. N., Potenza, M. T., Fremer, J. J., and Ward, W. C. (2002). *Computer-Based Testing: Building the Foundation for Future Assessments*. Psychology Press.

Mingers, J. (2003). The paucity of multimethod research: a review of the information systems literature. *Information Systems Journal*, 13(3):233–249.

Mitra, S. (2003). Minimally invasive education: a progress report on the "hole-in-the-wall" experiments. *British journal of educational technology*, 34(3):367–371.

Mittal, S. and Mittal, A. (2011). Versatile question answering systems: seeing in synthesis. *International Journal of Intelligent Information and Database Systems*, 5(2):119–142.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, 13(1):29–39.

Morrow, V. (2008). Ethical dilemmas in research with children and young people about their social environments. *Children's geographies*, 6(1):49–61.

Mory, E. H. (2004). Feedback research revisited. *Handbook of research on educational communications and technology*, 2:745–783.

Moscovich, T. (2007). *Principles and Applications of Multi-touch Interaction*. PhD thesis, Brown University, Providence, RI, USA.

Moyer, P. S. (2001). Are We Having Fun Yet? How Teachers Use Manipulatives to Teach Mathematics. *Educational Studies in Mathematics*, 47(2):175–197.

MSDN (2016). Chapter 21: Designing web applications. Retrieved August 2, 2016, from https://msdn.microsoft.com/en-us/library/ee658099.aspx.

Mullet, K. and Sano, D. (1994). *Designing visual interfaces: Communication oriented techniques*. Prentice Hall.

Myers, B. A. (1998). A brief history of human-computer interaction technology. *interactions*, 5(2):44–54.

Myers, M. D. and Avison, D. (2002). *Qualitative Research in Information Systems*. SAGE Publications.

Myhrvold, N. (2013). Moore's Law Corollary: Pixel Power. Retrieved August 2, 2016, from http://www.nytimes.com/2006/06/07/technology/circuits/07essay.html.

Nahavandipoor, V. (2012). *iOS 6 programming cookbook*. O'Reilly Media, Inc.

NCCA (2006). Mathematic curriculum. Retrieved August 2, 2016, from http://www.ncca.ie/uploadedfiles/Curriculum/Maths_Curr.pdf.

NCCA (2016). Primary School Curriculum. Retrieved August 12, 2016, from http://www.ncca.ie/en/Curriculum_and_Assessment/Early_Childhood_and_Primary_Education/Primary-Education/Primary_School_Curriculum/.

NCTM (2000). *Principles and Standards for School Mathematics*. National Council of Teachers in Mathematics (NCTM), Reston, VA.

Neuburg, M. (2012). *Programming iOS 5: Fundamentals of iPhone, iPad, and iPod touch Development*. O'Reilly Media, Inc.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, 14(2):149–170.

Nicol, D. and Milligan, C. (2006a). Rethinking technology-supported assessment practices in relation to the seven principles of good feedback practice. *Innovative assessment in higher education*, pages 64–77.

Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218.

Nicol, D. J. and Milligan, C. (2006b). Rethinking technology-supported assessment in terms of the seven principles of good feedback practice. In *Innovative Assessment in Higher Education*. Routledge, Taylor and Francis Group, London.

Nielsen, J. (2010). Why you only need to test with 5 users. Retrieved August 2, 2016, from https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/.

Nielsen, J. (2012). How many test users in a usability study? Retrieved August 2, 2016, from https://www.nngroup.com/articles/how-many-test-users/.

Nitko, A. J. (2001). *Educational assessment of students*. ERIC.

Norouzi, M., Fleet, D. J., and Salakhutdinov, R. R. (2012). Hamming distance metric learning. In *Advances in neural information processing systems*, pages 1061–1069.

Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge University Press.

Nunamaker Jr, J. F., Chen, M., and Purdin, T. D. (1990). Systems development in information systems research. *Journal of management information systems*, 7(3):89–106.

O'bannon, B. W. and Thomas, K. (2014). Teacher perceptions of using mobile phones in the classroom: Age matters! *Computers & Education*, 74:15–25.

Oblinger, D., Oblinger, J. L., and Lippincott, J. K. (2005). *Educating the net generation*. Boulder, Colorado.

OECD (2001). *Understanding the digital divide*. OECD Publications, Paris.

Oliver-Hoyo, M. and Allen, D. (2006). The use of triangulation methods in qualitative educational research. *Journal of College Science Teaching*, 35(4):42.

Pachler, N. (2005). Theories of learning and ict. *Learning to teach using ICT in the secondary school: A companion to school experience*, page 189.

Pal, S. K., Dillon, T. S., and Yeung, D. S. (2012). *Soft computing in case based reasoning*. Springer Science & Business Media.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3):45–77.

Pellegrino, J. W. and Quellmalz, E. S. (2010). Perspectives on the Integration of Technology and Assessment. *Journal of Research on Technology in Education*, 43(2):119–134.

Pestalozzi, J. H. (1803). *A.B.C. der Anschauung oder Anschauungs-Lehre der Maßverhältnisse*. Gessner.

Piaget, J. (1971). *The Psychology of Intelligence*. Routledge and Kegan, Boston.

Podsakoff, P. M., MacKenzie, S. B., and Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual review of psychology*, 63:539–569.

Polsani, P. R. (2006). Use and abuse of reusable learning objects. *Journal of Digital information*, 3(4).

Polya, G. (1957). *How to Solve It: A New Aspect of Mathematical Method*. Doubleday, Garden City.

Poole, D. L., Mackworth, A. K., and Goebel, R. (1998). *Computational intelligence: a logical approach*, volume 1. Oxford University Press New York.

Poslad, S. (2009). *Ubiquitous computing smart devices, smart environments and smart interaction*. Wiley.

Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the horizon*, 9(5):1–6.

Puentedura, R (2012). Samr and tpck: Intro to advanced practice. Retrieved August 12, 2016, from http://hippasus.com/resources/sweden2010/SAMR_TPCK_IntroToAdvancedPractice.pdf.

Puentedura, R (2014). Samr model explained through examples. Retrieved August 12, 2016, from http://www.educatorstechnology.com/2013/08/samr-model-explained-through-examples.html.

Rasila, A., Harjula, M., and Zenger, K. (2007). Automatic assessment of mathematics exercises: Experiences and future prospects. *ReflekTori 2007 Symposium of Engineering Education*, 1:70–80.

Rebholz, S. and Zimmermann, M. (2013). Applying computer-aided intelligent assessment in the context of mathematical induction. In *Transactions on Edutainment X*, pages 191–201. Springer.

Recker, J. (2012). *Scientific research in information systems: a beginner's guide*. Springer Science & Business Media.

Redecker, C. et al. (2013). The use of ict for the assessment of key competences. *Joint Research Centre Institute for Prospective Technological Studies. Scientific and Technical Research series*.

Reigeluth, C. M. and Nelson, L. M. (1997). A New Paradigm of ISD? *Educational Media and Technology Yearbook*, 22:24–35.

Resnick, M. (1998). Technologies for lifelong kindergarten. *Educational Technology Research and Development*, 46(4):43–55.

Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K., and Silverman, B. (1998). Digital manipulatives: new toys to think with. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–287. ACM Press/Addison-Wesley Publishing Co.

Richardson, M., Baird, J.-A., Ridgway, J., Ripley, M., Shorrocks-Taylor, D., and Swan, M. (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computers in Human Behavior*, 18(6):633–649.

Richter, M. M. and Rosina, O. (2013). *Case-based reasoning: a textbook*. Springer, London.

Riesbeck, C. K. and Schank, R. C. (2013). *Inside case-based reasoning*. Psychology Press.

Ripley, M. (2009). Transformational computer-based testing. *The transition to computer-based assessment*, page 92.

Rittle-Johnson, B. and Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, 91(1):175–189.

Rodríguez-Gómez, G. and Ibarra-Sáiz, M. S. (2015). Assessment as learning and empowerment: towards sustainable learning in higher education. In *Sustainable Learning in Higher Education*, pages 1–20. Springer.

Rogers, P. R., Miller, A., and Judge, W. Q. (1999). Using information-processing theory to understand planning/performance relationships in the context of strategy. *Strategic Management Journal*, 20(6):567–577.

Ruiz-Primo, M. A., Schultz, S. E., and Shavelson, R. J. (1997). *Concept map-based assessment in science: Two exploratory studies*. Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., and Edwards, D. D. (2009). *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River.

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144.

Salem, A.-B. M. (2008). Data mining technology in e-learning. In *Proceedings of the 6th International Conference on Emerging e-Learning Technologies and Applications, ICETA*.

Salem, A.-B. M. (2011). Intellectual e-learning systems. In *Proc. Of the Annual International Conference on " Virtual and Augmented Reality in Education"(VARE 2011)), Valmiera, Latvia*, pages 16–23.

Salem, A.-B. M. (2015). Towards of intelligence education and learning. In *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 196–202. IEEE.

Sangwin, C. (2013). *Computer Aided Assessment of Mathematics*. Oxford University Press.

Sangwin, C. J. (2007). Assessing elementary algebra with stack. *International Journal of Mathematical Education in Science and Technology*, 38(8):987–1002.

Sargent, J, Wood, M.M, and Aderson, S.M (2004). A human-computer collaborative approach to the marking of free text answers. In *8th International Computer Assisted Assessment Conference,*, Loughborough University.

Saunders, M., Lewis, P., and Thornhill, A. (2012). *Research Methods for Business Students*. Pearson.

Savery, J. R. and Duffy, T. M. (1996). *Problem Based Learning: An instructional model and its constructivist framework*. Educational Technology Publications Englewood Cliffs, NJ.

Scalise, K. and Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *The Journal of Technology, Learning, and Assessment*, 4(6).

Schacter, J., Herl, H., Chung, G., Dennis, R., and O'Neil, H. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior*, 15(3):403–418.

Schank, R. C. (1999). *Dynamic memory revisited*. Cambridge University Press.

Schau, C., Mattern, N., Zeilik, M., Teague, K. W., and Weber, R. J. (2001). Select-and-fill-in concept map scores as a measure of students' connected understanding of science. *Educational and Psychological Measurement*, 61(1):136–158.

Scheuermann, F. and Björnsson, J. (2009). The transition to computer-based assessment. *Luxembourg: Office for Official Publications of the European Communities*.

Schoenfeld, A. H. (1989). Explorations of Students' Mathematical Beliefs and Behavior. *Journal for Research in Mathematics Education*, 20(4):338–355.

Segal, A. (2011). *Do Gestural Interfaces Promote Thinking? Embodied Interaction: Congruent Gestures and Direct Touch Promote Performance in Math.* ERIC.

Shiu, S. C. and Pal, S. K. (2004). Case-based reasoning: concepts, features and soft computing. *Applied Intelligence*, 21(3):233–238.

Shneiderman, B. (2010). *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.

Shohamy, E. (2014). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. Routledge.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1):153–189.

Shuttleworth, M. (2009). What is the scientific method? Retrieved August 2, 2016, from https://explorable.com/scientific-method.

Siegle, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford University Press.

Silver, E. A. (1986). Using conceptual and procedural knowledge: A focus on relationships. In *Conceptual and procedural knowledge: The case of mathematics*, pages 181–198. Lawrence Erlbaum Associates, Inc.

Silverman, D. (2013). *Doing qualitative research: A practical handbook*. SAGE Publications Limited.

Skinner, E. A. and Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, 85(4):571–581.

Snieder, R. and Larner, K. (2009). *The art of being a scientist: A guide for graduate students and their mentors*. Cambridge University Press.

Sourceforge.net (2013). Reduce:a portable general-purpose computer algebra system. Retrieved August 2, 2016, from http://reduce-algebra.sourceforge.net/.

Stair, R. and Reynolds, G. (2012). *Fundamentals of information systems*. Cengage Learning.

Stein, M. K. and Bovalino, J. W. (2001). Manipulatives: One piece of the puzzle. *Mathematics Teaching in the Middle School*, 6(6):356.

Stevens, D. D. and Levi, A. J. (2011). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus Publishing, LLC.

Stone, R. G., Batmaz, F., and Hinde, C. J. (2009). Drawing and marking graph diagrams. *ITALICS*, 8(2).

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Szetela, W. (1992). Evaluating Problem Solving in Mathematics. *Educational Leadership*, 49(8):42–45.

Thomas, P., Smith, N., and Waugh, K. (2008). Automatically assessing graph based diagrams. *Learning, Media and Technology*, 33(3):249–267.

Thomson, A. J. and Schmoldt, D. L. (2001). Ethics in computer software design and development. *Computers and Electronics in Agriculture*, 30(1–3):85–102.

Tomorrow, P. (2014). The new digital learning playbook: Understanding the spectrum of students' activities and aspirations. Retrieved August 2, 2016, from http://www.tomorrow. org/speakup/SU13DigitalLearningPlaybook_StudentReport.html.

Torbeyns, J., Obersteiner, A., and Verschaffel, L. (2015). Number sense in early and elementary mathematics education. *Yearbook of the Department of Early Childhood Studies*, 5:60–75.

Tselonis, C., Sargeant, J., and McGee Wood, M. (2005). Diagram matching for human-computer collaborative assessment. In *Proceedings of the 9th International Computer Assisted Assessment Conference (CAA'05)*. © Loughborough University.

UK, Essays (2015). Explanation of the concept of research onion psychology essay. Retrieved August 2, 2016, from https://www.ukessays.com/essays/psychology/ explanation-of-the-concept-of-research-onion-psychology-essay.php.

UNESCO (2014). Global teacher shortage. Retrieved August 2, 2016, from http://www.uis. unesco.org/Education/Pages/global-teacher-shortage.aspx.

UNESCO (2016). Bridging the digital divide. Retrieved August 2, 2016, from http://portal.unesco.org/en/ev.php-URL_ID=15738&URL_DO=DO_TOPIC&URL_ SECTION=201.html.

UPI.com (2013). Tablets giving PCs a run for the money. Retrieved August 18, 2013, from http://www.upi.com/Business_News/2012/03/05/ Tablets-giving-PCs-a-run-for-the-money/UPI-37341330981057/.

Vaishnavi, V. K. and Kuechler, W. (2015). *Design science research methods and patterns: innovating information and communication technology*. Crc Press.

van Dam, A. (1997). Post-WIMP user interfaces. *Commun. ACM*, 40(2):63–67.

Van den Heuvel-Panhuizen, M., Kolovou, A., and Peltenburg, M. (2011). Using ict to improve assessment. *Assessment in the mathematics classroom: Yearbook*, pages 165–185.

Van Lehn, K. (1986). Arithmetic procedures are induced from examples. In *Conceptual and procedural knowledge: The case of mathematics*. Lawrence Erlbaum Associates, Inc.

Vantage Learning (2016). IntelliMetric®: How it Works. Retrieved August 2, 2016, from http://www.vantagelearning.com/products/intellimetric/intellimetric-how-it-works/.

Venable, J. R., Aitken, A., Chang, V., Dreher, H., Issa, T., von Konsky, B., and Wood, L. (2012). Developing a research design for comparative evaluation of marking and feedback support systems. In *Proceedings of the 21st Annual Teaching Learning Forum, Murdoch University, Perth.* Retrieved August 2, 2016, from https://ctl.curtin.edu.au/events/conferences/tlf/tlf2012/refereed/venable.html.

Vendlinksi, T. and Stevens, R. (2002). Assessing Student Problem-Solving Skills With Complex Computer-Based Tasks. *The Journal of Technology, Learning and Assessment*, 1(3).

Verschaffel, L., Greer, B., and De Corte, E. (2000). *Making sense of word problems*. Swets & Zeitlinger.

Villar, N., Izadi, S., Rosenfeld, D., Benko, H., Helmes, J., Westhues, J., Hodges, S., Ofek, E., Butler, A., Cao, X., et al. (2009). Mouse 2.0: multi-touch meets the mouse. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 33–42. ACM.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Walvoord, B. E. (2010). *Assessment clear and simple: A practical guide for institutions, departments, and general education*. John Wiley & Sons.

Ward, W. C. and Bennett, R. E. (2012). *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Routledge.

Warschauer, M. (2011). Tablet computers in education. eventually tablets will facilitate more personalized and interactive learning. *Educational Technology Debate–Exploring Learning in Developing Countries*.

Wastiau, P., Blamire, R., Kearney, C., Quittre, V., Van de Gaer, E., and Monseur, C. (2013). The use of ict in education: a survey of schools in europe. *European Journal of Education*, 48(1):11–27.

Watanabe, H., Arai, M., and Takei, S. (2001). Case-based evaluation of novice programs. *Proc. of AI-ED2001*, 1:55–64.

Watson, I. (1998). *Applying case-based reasoning: techniques for enterprise systems*. Morgan Kaufmann Publishers Inc.

Watson, I. (1999). Case-based reasoning is a methodology not a technology. *Knowledge-based systems*, 12(5):303–308.

Weber, S. (2004). *The success of open source*, volume 368. Cambridge Univ Press.

Weinerth, K., Koenig, V., Brunner, M., and Martin, R. (2014). Concept maps: A useful and usable tool for computer-based knowledge assessment? a literature review with a focus on usability. *Computers & Education*, 78:201–209.

Welling, L. and Thomson, L. (2003). *PHP and MySQL Web development*. Sams Publishing.

Wigdor, D. and Wixon, D. (2011). *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture.* Elsevier.

Wiley, D. A. (2001). *Instructional use of learning objects.* Agency for instructional technology. Retrieved August 2, 2016, from http://www.reusability.org/read.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., and DeMark, S. F. (2004). Design Rationale for a Complex Performance Assessment. *International Journal of Testing*, 4(4):303–332.

Wirth, J. (2008). Computer-based tests: Alternatives for test and item design. *Assessment of competencies in educational contexts*, pages 235–252.

World Health Organization (2016). Ethical standards and procedures for research with human beings. Retrieved August 2, 2016, from http://www.who.int/ethics/research/en/.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., and Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, 15(4):391–412.

Yeap, B. and Kaur, B. (2001). Semantic characteristics that make arithmetic word problems difficult. In *Proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia Incorporated (MERGA) on "Numeracy and Beyond.* Australia.

Zimmerman, B. J., Boekarts, M., Pintrich, P., and Zeidner, M. (2000). A social cognitive perspective. *Handbook of self-regulation*, 13.

Zoanetti, N. (2010). Interactive computer based assessment tasks: How problem-solving process data can inform instruction. *Australasian Journal of Educational Technology*, 26(5):585–606.

Zuckerman, O., Arida, S., and Resnick, M. (2005). Extending tangible interfaces for education: digital montessori-inspired manipulatives. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 859–868. New York, ACM.

Zuppo, C. M. (2012). Defining ict in a boundaryless world: The development of a working hierarchy. *International Journal of Managing Information Technology*, 4(3):13.

Appendix A

Ethics documents

This appendix provides the ethics documents used for the study in line with descriptions given in Section 3.10. The documents includes the ethics approval letter, Head Teacher request letter,Head Teacher consent form, participants recruitment letter, participant assent form, parent notification letter, parental opt-out form.

A.1   Ethics sub-committee approval

Ref No: R13-P104

LOUGHBOROUGHUNIVERSITY
ETHICS APPROVALS (HUMAN PARTICIPANTS)SUB-COMMITTEE

RESEARCH PROPOSAL
INVOLVING HUMAN PARTICIPANTS

Title: Using Multi-Touch Gestures for Arithmetic Work

Applicant: Dr R Stone, A Adesina, Dr F Batmaz, Dr I Jones

Department: Computer Science/MEC

Date of clearance: 10 May 2013

Comments of the Sub-Committee:

The Sub-Committee agreed to issue conditional approval, subject to the following conditions:

- That the start date for data collection **had not** already passed. It appeared that data collection had already started and the Sub-Committee emphasized that Ethical Clearance **cannot be granted** after the data collection had started.
- That written confirmation was provided that the Headteacher of the school was happy for the study to proceed.
- That the Participant Information Sheet was amended to:
  - Include full contact details of all investigators.
  - Include all the relevant questions from the Participant Information Sheet Template.

A.2   Head Teacher letter

Loughborough University

Computer Science Department
LoughboroughUniversity
Loughborough
LE11 3TU

Email: a.o.adesina@lboro.ac.uk
Tel: 01509 225905

Dear Head teacher,

Using Multi-Touch Gestures for Arithmetic Work

I am writing following my supervisor IanJones's recent email regarding the possibility of undertaking some research work at your school. I am a PhD studnt at Loughborough University investigating children's arithmetic strategies. I am writing to request written consent to undertake a study that seeks to improve the quality of assessment and feedback processes in primary school arithmetic work.

To do this we would like five groups of four or five Y5 pupils to undertake arithmetic word problems on iPads provided by us. Pupils with average or above average literacy levels will be well suited for the study. The participants will be required to solve series of word problems using touch gestures. The study will check if the pupils are able to successfully solve the problems using the interactive touch gestures. It will also examine if solution steps captured can be used to detect arithmetic strategies. Pupils will also complete a short written test which will be used to investigate the suitability of implemented gestures. A quiet room is required for the activities with each group session expected to last around 30 minutes. All the researchers are former teachers and are DBS (CRB) cleared.

We will not collect video or audio records of the children activities. The information we collect will be kept strictly confidential. Test results and solutions will be identified by a code number only, and kept in a locked filing cabinet at the University. We plan to publish the findings in a scientific journal, although this can take two to three years from the end of the study. We will send to your school the results from the study. The proposed period for the study is from 3rd June 2013 to the end of the term. Your school may require more than a day to complete the study on the five groups.

In order to undertake the study we must provide written consent from you to the university's ethical committee. If you are happy for pupils and mathematics teachers to participate in the study please be kind enough to complete the attached, and return by email to **a.o.adesina@lboro.ac.uk.**

Kind regards,

AdewaleAdesina

A.3   Head Teacher consent form

Using Multi-Touch Gestures for Arithmetic Work

HEADTEACHER WRITTEN CONSENT FORM

Primary School
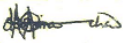
I consent for the study to take place at our primary school

Brenda Davies

Head teacher's signature: _____

Date: _____

A.4   Participant recruitment letter

**Using Multi-Touch Gestures for Arithmetic
Work**

**Loughborough
University**

Computer Science Department
LoughboroughUniversity
Loughborough
LE11 3TU

Email: a.o.adesina@lboro.ac.uk
Tel: 01509 225905

Dear Student,

Your school is taking part in a study about using multi-touch gestures for arithmetic
work. We would like to you to be part of this study. Please read the following infor-
mation and if anything is unclear ask your teacher for further information.

Researchers from Loughborough University are investigating how to help pupils can
use touch gestures to solve arithmetic problems. To do this study we want groups of
students in your year to assess and solve a sequence of word problems.

If you choose to take part you will have a researcher from the university will give a
short demonstration on how to complete the tasks on the iPad. After which you will
be given three arithmetic word problems to work on.

The information we collect is kept strictly confidential. We will not collect names or
personal details, ID numbers will be used to identify completed tasks.

If you choose not to take part when the researcher comes to take the lesson then the
teacher will instead provide some usual class work for you.

Please complete the form on the next page and return it to your teacher.

Adewale Adesina

A.5   Participant assent form

Loughborough University

Using Multi-Touch Gestures for Arithmetic Work

STUDENT ASSENT FORM

Your school has agreed to take part in a study run by researchers at Loughborough University looking at helping pupils use touch gestures for arithmetic work. Please tick ONE of the following options.

☐   I choose to take part.

☐   I choose not to take part.

- — -

Name: _____   _____
        forename   surname

Date of birth:_____   Class: _____

Signature:_____

Date:_____

A.6   Parent letter

Using Multi-Touch Gestures for Arithmetic Work

Loughborough University

Computer Science Department
LoughboroughUniversity
Loughborough
LE11 3TU

Email: a.o.adesina@lboro.ac.uk
Tel: 01509 225905

Dear Parent,

Your child's school is taking part in a study on how multi-touch gestures can be used to assist students in arithmetic work and simultaneously capture solution steps. We would like to your child to be part of this study. Please read the following information and if anything is unclear ask your teacher for further information.

Researchers from Loughborough University are investigating how feedback and assessment of arithmetic work can be improved by the use of touch gestures on a tablet device. To do this study we want pupils to attempt solving arithmetic word problems involving addition and subtraction.

Findings from literature suggests that touch interactions can aid cognition and learning. Our previous work indicates that useful feedback on solution steps can be obtained from the process. The activity will be beneficial for your child's understanding of the important arithmetic concepts. A researcher from Loughborough University will guide your child on completing the tasks. The tests will be anonymised and then assessed by the research team.

What happens to the results of the research study?

The information we collect is kept strictly confidential. We would not give it to anyone else without written consent from you. Test results and predicted grades will be identified by a code number only, and kept in a locked filing cabinet at the University. We plan to publish the findings in a scientific journal, although this can take two to three years from the end of the study. We will also send the results of the study to your child's school.

Who is conducting this study?

The principal investigator is Adewale Adesina a PhD student at Loughborough University. If you have any more questions please contact Adewale using the details given overleaf, or ask your child's teacher for further information.

What should I do next?

Please fill in the enclosed form and return it to your child's class teacher if you would prefer your child not to take part in this study.

Adewale Adesina

A.7   Parental opt-out form

Loughborough University

Computer Science Department,
Loughborough University, UK

Using Multi-Touch Gestures for Arithmetic Work

PARENTAL OPT-OUT FORM

Your child's school has agreed to take part in a study run by researchers at Loughborough

University looking at using multi-touch gestures for arithmetic work.

- **If you do not want your child to take part, please fill in the form below and return it to your child's class teacher as soon as possible.**

- To find out more about the project, please read the attached information sheet. You can also contact Adewale (email: a.o.adesina@lboro.ac.uk; Tel: 01509225905) if you have any questions.

- -

Name of child: _____ _____

                    forename   surname

Date of birth:_____ Class: _____

I do not wantmy child to take part in the above study.

Name of parent/guardian:_____

Signature:_____

Date:_____

Appendix B

Pilot Study

This appendix present the details and results of the Pilot Study carried out at the early stages of the research to study how the approach to study was received among students. Section 7.2 provided a summary on the study. This section includes the questions, questionnaire and results obtained in the study.

B.1 Participants

Seventeen Loughborough University students were recruited as participants for the study. It was assumed that they will not have difficulty with arithmetic tasks but are likely to be unfamiliar with the novel multi-touch approach. Although the participants are university students, we believe the findings may be relevant to younger learners as well. An introduction session was given to each participant on sample question to intimate them on how to use the tool to solve problems. After this, they were asked to solve three word problems using the techniques demonstrated

B.2 Questions used for study

The numbers in the questions were chosen to support the use of the place-value strategy by students such that in each problem there is a large (two-digit) number, and a corresponding small (single-digit) number that sum to a multiple of 10. In each problem the two-digit number is presented in a different position: 2nd in question 1; 3rd in question 2; 1st in

question 3. The particular values were selected so that adding the single digit numbers was not too easy, i.e. every single digit addition requires a carry over. The large numbers were selected such that each question is most easily answered by starting with the large number, and then one of the smaller numbers (i.e. the place-value strategy). Question 3 presents the numbers in strategic order. This is a control question to help us work out if any participants consistently either (i) just chose numbers from left to right or (ii) just choose numbers arbitrarily.

The questions uses for the study are presented below.

Table B.1 Pilot study questions

Problem	Strategy
Q1.William had 7 bottles of wine. His father gave him 41 more bottles of wine. His mother gave him 9 more bottles of wine. How many bottles of wine did William have altogether?	start with 41 then add 9 then add 7
Q2. Sara has 8 sugar donuts. She also has 5 plain donuts and 32 jam donuts. How many donuts does Sara have altogether?	start with 32 then add 8 then add 5
Q3. Jason owned a factory that employs 53 workers. He hired another 7 workers. He then hired another 6 workers. How many workers are there at the factory altogether?	start with 53 then add 7 then add 6

B.3   Questionnaire

Loughborough University

Pilot Study

Investigating the Potentials of Multi-Touch Arithmetic Tool in Solving Arithmetic Problems

The following questionnaire is designed to investigate the potential of the iPad application in carrying out simple arithmetic operations and also identify usability issues.

Thank you in advance for your time taken to complete this questionnaire, your views are extremely valuable and much appreciated.

All responses are anonymous and all results are strictly confidential.

Please complete the following questions as best as you can.

1. What is your department?

_____

 2. Are you able to use the tool to solve the prob-   ☐ ☐

                                                             Yes No

3. How easy was it to solve the problem with the tool?

Definitely easy   ☐

Moderately easy   ☐

Sort of easy   ☐

Sort of hard   ☐

Moderately hard   ☐

Definitely hard   ☐

 4. What do you like about the tool?

☐

5. What don't you like about the tool?

☐

Fig. B.1 Questionnaire

Appendix C

MuTAT Evaluation Study details and results

This appendix provides additional details and results on the main study carried out in the schools. It presents the participant briefing document, full details of the study questions,and complementary results to those already provided in Sections 7.3.

C.1   Participant briefing

To ensure consistency in the study all participant were given the same instructions before they commencement of the experimental tasks. A sample of the script used is shown below.

Raw script

My name is Adewale Adesina , I am here with my supervisor Dr. Ian Jones and I will be running this session with you today.

First of all, thank you for agreeing to participate in study today. I really appreciate it. We should be here together for about 30 minutes doing some simple math on the iPad after which you can all go back to your classes.

We are researches working on educational software to improve teaching and learning and today we have software on the iPad. We want to find out if it works for real users people like you. This is not a test of you or your mathematics abilities; instead you will be helping us to test the software in the iPad if it has got problems. So, please feel free as you perform the tasks.

Before we start, I have a little bit of paperwork. Let me have your names, age, gender and the color of the sticker on the iPad you have in front of you.

Thank you, now we will like to see how easy it is for you to go about using hand gestures to do math on the iPad. I will demonstrate how to do the problems. You should all watch what I am doing on the laptop screen in front of us here. After this, you call have a go on the remaining problems. There are eleven of them. If something unexpected happens, don't worry, you didn't break anything. I will be here to let you know what to do in those situations. After you have finished the question on the iPad, I will give you a paper with addition problems for you to solve as well. Do you have any questions before we begin?

OK, let's start with the demonstration

[After the demonstration] Thank you. Now let's move on to the questions on the paper. Remember to write your names at the top of the paper. *[After all tasks]* Thank you. You have been very helpful. How was it working on the problems with the iPad? Do you have any questions for me about what you just worked on?

[After all] OK, we have finished. As a way of thanking you, I'd like to give you some sweets. Can you please select any two you like? *[Hand participant sweet packet]*

Thanks once again. If you're ready, you can now go back to your classes.

[End]

Fig. C.1 participants briefing script

C.2   Study questions

The list of questions used in the study for the MuTAT and paper-and-pencil tasks are listed below.

Table C.1 MuTAT study questions

SNo.	Question	Category
1	Raj ordered some crayons, a bottle of juice and a glass of water. The crayons cost £4 , the bottle of juice cost £7 and the glass of water cost £1 . How much did Raj pay altogether?	Training question
2	A phone costs £80 . A TV cost £110 . How much money did Ken need to buy the phone and the TV ?	Training question
3	Joe had 7 toy cars. His father gave him 26 toy cars. His mother gave him 4 more toy cars. How many toy cars did Joe have altogether?	Test I
4	Kate has 6 red sweets. She also has 8 green sweets and 14 brown sweets. How many sweets does Sara have altogether ?	Test I
5	William had 37 chocolate bars. His father gave him 41 chocolate bars. His mother gave him 9 more chocolate bars. How many chocolate bars did William have altogether?	Test I
6	Sara has 8 sugar donuts. She also has 15 plain donuts and 32 jam donuts. How many donuts does Sara have altogether?	Test I
7	Jason owned a factory that employs 53 workers. He hired another 7 workers. He then hired another 16 workers. How many workers are there at the factory altogether?	Control for Test I
8	There are some people in a queue, 12 more arrive and then 8 leave. At the end there are 16 people. How many were there to start with?	Test II
9	There are 24 books on a shelf, some more are added and then 6 taken away. At the end there are 29 books. How many books were added?	Test II
10	Emily had 18 stickers, she bought some more and then gave 8 to Sam. At the end Emily had 23. How many did Emily buy?	Test II
11	There are 22 sheep in the field, 13 more are put there and some moved away. At the end there are 28 sheep in the field. How many were moved away?	Test II
12	Sam had 21 books, he got 9 more and then gave 13 to Owen. At the end, how many books did Sam have?	Test II

Table C.2 Structure of questions in paper-and-pencil test

SNo.	Question
1	34 + 6 + 18
2	29 + 17 +1
3	21 + 9 + 10
4	18 + 25 + 2

C.3 Additional results

C.3.1 Performance of male vs female
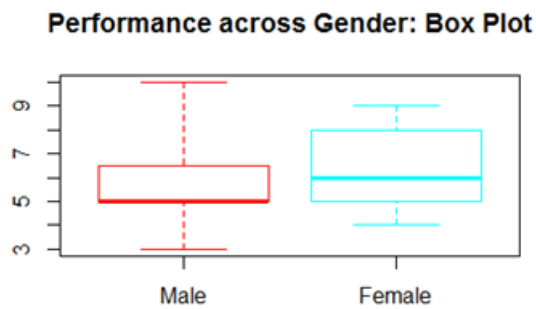
Fig. C.2 Box plot of performance across gender

C.3.2 Students relative performance across schools
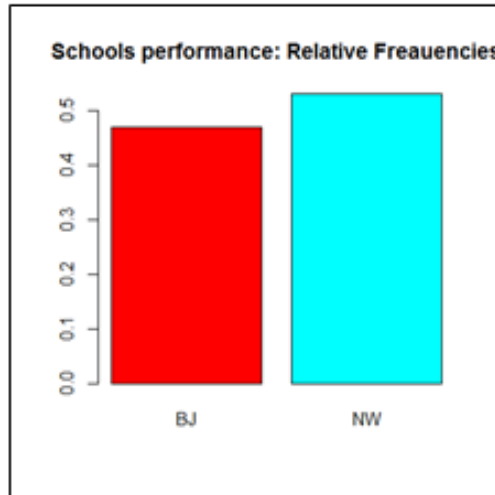
Fig. C.3 Students'relative performances

C.3.3 Students'final score across questions
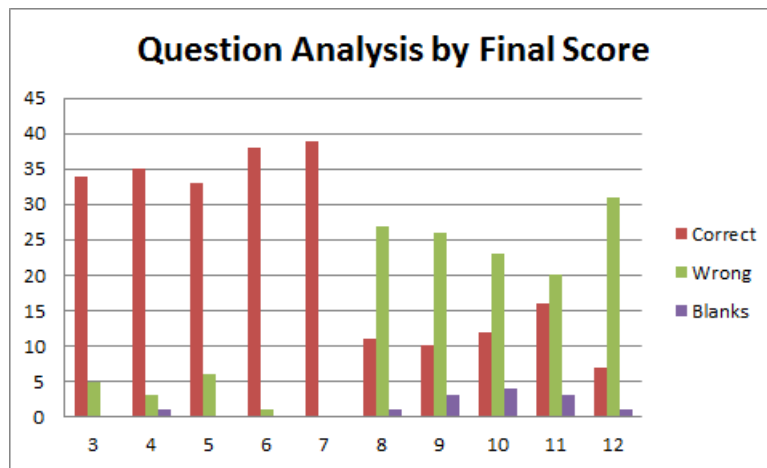
Fig. C.4 Students performances by final score in all questions

C.3.4 Scattered plot of average task times
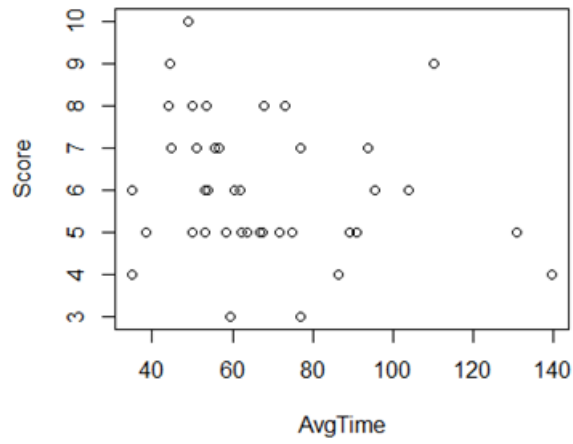
Fig. C.5 Histogram of task times

C.3.5 Strategies
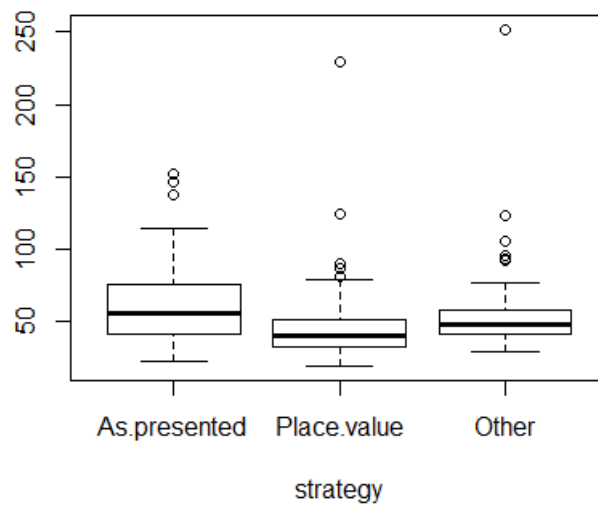
Fig. C.6 Box plot of stragegies

Appendix D

Questionnaires and Data from the Study with Teachers

This appendix presents the data obtained from teachers who attended the 2015 BETT show. The BETT show is an annual event brings together educators and exhibitors for the advancement of teaching and learning. EducationCity actively participated in the 2015 event [1]. EducationCity is a leading teaching, and assessment resource center. which focuses on children aged between 3-12 years old and children with Special Educational Needs.

D.1 Questionnaire

The questions administered and completed by the teachers is presented below:

EducationCity **Loughborough University**

Educational Tool

Teacher Survey 2015

This survey is being carried out to find out how you feel about the Multi-Touch Arithmetic tool supporting teachers like you in your work. Please answer the questions freely. You cannot be identified from the information you provide, and no information about individuals will be disclosed.

All information you provide will be treated in the strictest confidence. Your decision to participate in this research if entirely voluntary. If you do not wish to take part, just do not return the questionnaire. If you decide to take part, the questionnaire should take you about five minutes to complete after you have interacted with the tool. Please answer the questions in the spaces provided.

Thank you for your help

Adewale Adesina

Ellen Hartshorn

Please score the following according t o the ratings shown: Please tick the appropriate box.

1. Have you seen an educational tool like this before? ☐ ☐

Yes No

2. How useful do you think it would be in each of these subjects?

	Very useful	A little useful	Not sure	Not Useful
Maths	☐	☐	☐	☐
English	☐	☐	☐	☐
Science	☐	☐	☐	☐
Other	☐	☐	☐	☐

3. How useful do you think it would be in each key stage?

	Very useful	A little useful	Not sure	Not Useful
Early Years Foundation Stage (EYFS)	☐	☐	☐	☐
Key Stage 1 (KS1)	☐	☐	☐	☐
Key Stage 2 (KS2)	☐	☐	☐	☐
Key Stage 3 (KS3)	☐	☐	☐	☐
Key Stage 4 (KS4)	☐	☐	☐	☐
Key Stage 5+ (KS5+)	☐	☐	☐	☐

4. What would you rate this tool as ?

	1	2	3	5	5	
Diagnostic tool	☐	☐	☐	☐	☐	Final assessment tool

5. What would you expect data to look like at the end?

6. How would you most like to see this?

	Yes	Maybe	No
App (iPad)	☐	☐	☐
App (android)	☐	☐	☐
Online access	☐	☐	☐
Part of EduCity Subscription	☐	☐	☐

7. How much would you expect it to cost per year?

8. How excited are you about this?

Very	In between	Not at all
☐	☐	☐

9. If money was no object would you have it in your school?

Very	No
☐	☐

10. In 5 years do you think that the majority of schools will be close to having 1-1 devices for each child?

	Yes	No	Not Sure
Nursery/reception	☐	☐	☐
KS1	☐	☐	☐
KS2	☐	☐	☐
KS3/4	☐	☐	☐
Ks5	☐	☐	☐

11. Anything else you would like to say about this tool, or see it do?

D.2   Additional results

Question How much would you expect it to cost per year:

We pay per app, therefore £1.99- £2.99 × amount of ipads free or £0.69 each free or £0.69 each Per child Per class (around max 35) Per school £475 max total

Question In 5 years do you think that the majority of schools will be close to having 1-1 devices for each child?

Table D.1 Future projections

	Yes	No	Not sure	Total responses
Nursery/reception	1	3	3	7
KS1	3	2	2	7
KS2	5	1		6
KS3/4	4	1		5
KS5	2	1	2	5

Appendix E

Dataset for the Marking Experiment

This appendix presents the dataset used for the marking experiments [1] outlined in Chapter 8.

S/No	Student No.	Question	Num1	Num2	First Op	Intermediate Answer	Num3	Second Op	Final Answer
1	1	4	8	32	1	40	15	1	55
2	1	3	21	9	1	30	13	2	17
3	1	1	24	11	1	35	6	2	29
4	1	2	53	7	1	60	16	1	76
5	2	2	53	7	1	60	16	1	76
6	2	4	8	15	1	23	32	1	55
7	2	1	11	24	1	35	6	2	29
8	2	3	9	21	1	30	13	2	17
9	3	3	9	21	1	30	13	2	17
10	3	1	11	6	2	5	24	1	29
11	3	4	8	15	1	22	32	1	54
12	3	2	7	53	1	60	16	1	76
13	4	3	21	9	1	30	13	2	17
14	4	1	24	11	1	35	6	2	29
15	4	4	15	32	1	47	8	1	52
16	4	2	53	7	1	59	16	1	75
17	5	2	16	7	1	23	53	1	76
18	5	1	24	11	1	34	6	2	28
19	5	3	21	13	2	8	9	1	17
20	5	4	32	8	1	40	15	1	55
21	6	2	53	7	1	60	16	1	76
22	6	4	8	32	1	40	15	1	55
23	6	3	9	13	2	-4	21	1	17
24	6	1	24	11	1	35	6	2	29
25	7	1	24	6	2	18	11	1	29
26	7	2	16	7	1	23	53	1	76
27	7	3	21	9	1	30	13	2	17
28	7	4	8	32	1	40	15	1	55
29	8	3	21	13	2	7	9	1	16
30	8	1	11	6	2	5	24	1	29
31	8	2	16	53	1	69	7	1	77
32	8	4	8	15	1	23	32	1	55
33	9	3	21	9	2	13	13	1	26
34	9	2	7	16	1	33	53	1	86
35	9	4	8	32	1	40	15	1	55
36	9	1	24	11	1	35	6	2	29
37	10	4	8	15	1	25	32	1	57
38	10	3	21	13	2	8	9	1	17
39	10	2	53	7	1	60	16	1	76
40	10	1	24	11	1	25	6	2	19
41	11	1	24	11	1	35	6	2	29
42	11	4	32	8	1	40	15	1	55
43	11	3	21	9	2	12	13	1	25
44	11	2	53	7	1	60	16	1	76

S/No	Student No.	Question	Num1	Num2	First Op	Intermediate Answer	Num3	Second Op	Final Answer
45	12	3	21	9	1	30	13	2	17
46	12	2	53	7	1	60	16	1	76
47	12	4	8	32	1	40	15	1	55
48	12	1	11	6	2	5	24	1	29
49	13	4	8	32	1	40	15	1	55
50	13	2	7	53	1	60	16	1	76
51	13	3	9	21	1	30	13	2	17
52	13	1	24	11	1	35	6	2	29
53	14	2	53	16	1	59	7	1	66
54	14	1	24	11	1	35	6	2	29
55	14	3	21	9	1	30	13	2	17
56	14	4	8	15	1	23	32	1	55
57	15	2	7	53	1	60	16	1	76
58	15	3	21	13	2	8	9	1	17
59	15	4	8	15	1	23	32	1	55
60	15	1	24	11	1	35	6	2	29
61	16	3	21	9	1	30	13	2	17
62	16	1	11	6	2	5	24	1	29
63	16	4	15	32	1	47	8	1	55
64	16	2	16	53	1	69	7	1	76
65	17	3	21	9	1	30	13	2	17
66	17	1	11	24	1	35	6	2	29
67	17	2	53	7	1	60	16	1	76
68	17	4	32	8	1	40	15	1	55
69	18	1	24	11	1	35	6	2	27
70	18	4	8	32	1	40	15	1	55
71	18	3	21	13	2	8	9	1	17

Appendix F

Code Snippets

This appendix provides some code snippets from the implementation of the prototypes.

F.1 Longpress gesture

The following code shows how the longpress gesture recognizer was implemented on the MuTAT.

```
- (void)longPressDetected:(UILongPressGestureRecognizer*)
    paramSender{
  if (calController) {
      return;
  }

  if (self.touchedPhraseOne != nil && self.
      touchedPhraseOne.layer.borderWidth == 0) {
      self.touchedPhraseOne = nil;
  }

  MWComponent *currentSender = (MWComponent*)paramSender.
      view;
  if (currentSender.tag == Entity || currentSender.tag ==
      Result) {
```

```
        if (paramSender.state ==
            UIGestureRecognizerStateBegan) {
          currentSender.layer.borderWidth = 6.0;
          currentSender.layer.borderColor = [UIColor
              yellowColor].CGColor;

          if (currentSender == self.touchedPhraseOne) {
            return;
          }

          if (self.touchedPhraseOne == nil) {
            self.touchedPhraseOne = currentSender;
          } else if (self.touchedPhraseTwo == nil) {
            self.touchedPhraseTwo = currentSender;
            [self calculatePhrases];
          }
        }

    } else if (paramSender.state ==
        UIGestureRecognizerStateEnded){
      currentSender.layer.borderWidth= 0.0;
    }
}
```

F.2   Arithmetic operators

The following shows the code that presents the arithmetic operators to students using the MVC design pattern.

```
- (void)addNumberPad
{
    self.numberPadView.center = CGPointMake(self.view.center
        .x, self.view.frame.size.height / 2 + 80);
    self.numberPadView.autoresizingMask =
        UIViewAutoresizingFlexibleLeftMargin |
        UIViewAutoresizingFlexibleRightMargin |
```

```
                UIViewAutoresizingFlexibleTopMargin  |
                UIViewAutoresizingFlexibleBottomMargin ;


        [ self . view  addSubview : self . numberPadView ] ;
}


- ( IBAction ) addAction : ( id ) sender
{
    UIView  *buttonView  =  ( UIView  *) sender ;
    buttonView . frame  =  [ buttonView  convertRect : buttonView .
            frame  toView : self . view ] ;
    operation  =  Addition ;
    [ self  addOperatorImage :@"+"  center : buttonView . center ] ;
}


- ( IBAction ) minusAction : ( id ) sender
{
    UIView  *buttonView  =  ( UIView  *) sender ;
    buttonView . frame  =  [ buttonView  convertRect : buttonView .
            frame  toView : self . view ] ;
    operation  =  Minus ;
    [ self  addOperatorImage :@"−"  center : buttonView . center ] ;
}


- ( IBAction ) multiplyAction : ( id ) sender
{
    UIView  *buttonView  =  ( UIView  *) sender ;
    buttonView . frame  =  [ buttonView  convertRect : buttonView .
            frame  toView : self . view ] ;
    operation  =  Multiply ;
    [ self  addOperatorImage :@"x"  center : buttonView . center ] ;
}


- ( IBAction ) divideAction : ( id ) sender
{
    UIView  *buttonView  =  ( UIView  *) sender ;
```

```
buttonView.frame = [buttonView convertRect:buttonView.
        frame toView:self.view];
operation = Division;
[self addOperatorImage:@"d" center:buttonView.center];
}
```

F.3   Presenting the visual representation of the solution steps

The following code makes the diagram of the solution steps.

```
- (void)viewDidLoad
{
    [super viewDidLoad];
    // Do any additional setup after loading the view from
            its nib.
    self.view.backgroundColor = [UIColor colorWithWhite:0
            alpha:0.8];
    self.view.autoresizingMask =
            UIViewAutoresizingFlexibleLeftMargin |
            UIViewAutoresizingFlexibleRightMargin |
            UIViewAutoresizingFlexibleTopMargin |
            UIViewAutoresizingFlexibleBottomMargin |
            UIViewAutoresizingFlexibleWidth |
            UIViewAutoresizingFlexibleHeight;

    UIButton *backButton = [UIButton buttonWithType:
            UIButtonTypeCustom];
    backButton.frame = self.view.frame;
    backButton.autoresizingMask =
            UIViewAutoresizingFlexibleLeftMargin |
            UIViewAutoresizingFlexibleRightMargin |
            UIViewAutoresizingFlexibleTopMargin |
            UIViewAutoresizingFlexibleBottomMargin |
            UIViewAutoresizingFlexibleWidth |
            UIViewAutoresizingFlexibleHeight;
```

```objc
    [backButton addTarget:self action:@selector(backAction:)
            forControlEvents:UIControlEventTouchDown];
    [self.view addSubview:backButton];


    componentItems = [[NSMutableArray alloc] init];
    relationShipLines = [[NSMutableArray alloc] init];
}


- (void)drawRelationShipLines
{
    if (relationShipLines) {
        for (UIView *v in relationShipLines) {
            [v removeFromSuperview];
        }
        [relationShipLines removeAllObjects];
    }

    [self drawLineLeftView:self.participateOne rightView:
            self.participateTwo toView:self.participateFour];
    if (self.expressionTwo != nil) {
        if (self.expressionTwo.participantOne.tag == Entity)
                {
            [self drawLineLeftView:self.participateThree
                    rightView:self.participateFour toView:self.
                    participateFive];
        } else {
            [self drawLineLeftView:self.participateFour
                    rightView:self.participateThree toView:self
                    .participateFive];
        }
    }
}
```